

2023-08

Machine learning model for prediction of malaria in low and high endemic areas of Tanzania

Mariki, Martina

NM-AIST

<https://dspace.nm-aist.ac.tz/handle/20.500.12479/2581>

Provided with love from The Nelson Mandela African Institution of Science and Technology

**MACHINE LEARNING MODEL FOR PREDICTION OF MALARIA IN
LOW AND HIGH ENDEMIC AREAS OF TANZANIA**

Martina Wilfred Mariki

**A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy in Information and Communication Science and Engineering of
the Nelson Mandela African Institution of Science and Technology**

Arusha, Tanzania

August 2023

ABSTRACT

Presumptive treatment and self-medication with anti-malaria drugs is a common practice in most limited resource settings that hinders proper management of malaria. However, these approaches have been considered unreliable due to the unnecessary use of malaria medication and untreated diseases that relate to malaria. This study aimed to develop a machine-learning model for malaria diagnosis using patients' symptoms and non-symptomatic features in high and low endemic areas of Tanzania. The malaria diagnosis dataset with 2556 patient's records and 36 features was collected in two regions of Tanzania: Morogoro and Kilimanjaro from 2015 -2019. Machine learning classifiers with the k-fold cross-validation methods were used to train and validate the model. To improve the performance of the diagnostic model, important features for malaria diagnosis were selected, and it was observed that the ranking of features differs among regions and when combined dataset. Significant features selected are residence area, fever, age, general body malaise, visit date, and headache. Random Forest and Decision Tree algorithms were the best performing classifiers in modelling malaria diagnosis datasets and attained 96%, 99% and 98% prediction accuracy for Kilimanjaro, Combined and Morogoro dataset respectively. These best-performing classifiers were evaluated using the unseen malaria diagnosis dataset and performed well in classifying malaria patients from sick patients. The final developed model showed that only a specific combination of features can predict malaria accurately. The results of this study revealed that malaria diagnosis using patients' symptoms and demographic features is possible. Also, the study results offer additional knowledge and shed light on the state diagnosis of malaria in the country. The developed machine learning model enables prediction of patient's malaria state using symptoms observed and non-symptomatic features before prescription of anti-malaria drugs. Apart from that the output of this study will be a necessary step in designing a malaria diagnosis decision support system through the developed model. Furthermore, towards reducing drug resistance, the results of this study can be used by the policymakers and the Ministry of Health for better management of malaria disease in health facilities and drug dispensing outlets to avoid self-medication and presumptive treatment.

DECLARATION

I, Martina Wilfred Mariki, do hereby declare to the Senate of the Nelson Mandela African Institution of Science and Technology Arusha that this dissertation is my original work and that it has neither been submitted nor is concurrently submitted for degree award in any other institution.

Martina Wilfred Mariki

Name of Candidate

Signature

Date

The above declaration is confirmed by:

Dr. Elizabeth Mkoba

Name of Supervisor 1

Signature

Date

Dr. Neema Mduma

Name of Supervisor 2

Signature

Date

COPYRIGHT

This dissertation is copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, on that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the office of Deputy Vice-Chancellor for Academic, Research and Innovation on behalf of both the author and the Nelson Mandela African Institution of Science and Technology.

.

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology (NM-AIST) a dissertation titled Machine learning model for Prediction of Malaria in Low and High endemic areas of Tanzania, in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology (NM-AIST).

Dr. Elizabeth Mkoba

Name of Supervisor 1

Signature

Date

Dr. Neema Mduma

Name of Supervisor 2

Signature

Date

ACKNOWLEDGEMENTS

First, I am grateful to Almighty God for protection and guidance throughout my studies.

I acknowledge and appreciate my Supervisors, Dr. Elizabeth Mkoba and Dr. Neema Mduma, for accepting and supervising my research with tireless efforts and valuable guidance towards completing my PhD studies. They have been inspirational, wonderful advisors, excellent supervisors, and good mentors in my PhD study.

I am grateful to the NM-AIST community and the United Republic of Tanzania government for offering me a scholarship and a supportive environment to accomplish this study.

Also, I would like to express my sincere gratitude to all the health facilities we have visited, the patients who took their time to talk to us, and the government offices where our research permission was provided. This wouldn't have been possible without you.

Lastly, I would like to thank my family for all the love and encouragement. Thanks to my fantastic husband, Dr. Kennedy Michael Ngowi, for endless encouragement and support throughout my study. To my beautiful children, Jayden, Jayleen, Ethan, and Ian, thank you for your patience throughout my study program.

DEDICATION

To my Husband, Dr Kennedy Michael Ngowi, and my Children, Jayden, Jayleen and Ethan. I could not and most certainly would not have done it without you guys by my side.

TABLE OF CONTENTS

ABSTRACT.....	i
DECLARATION	ii
COPYRIGHT.....	iii
CERTIFICATION	iv
ACKNOWLEDGEMENTS	v
DEDICATION.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	xii
ABBREVIATIONS AND SYMBOLS.....	xvi
CHAPTER ONE.....	1
INTRODUCTION	1
1.1 Background of the Problem	1
1.1.1 Global Malaria Burden	1
1.1.2 Malaria in Tanzania	2
1.1.3 Malaria Cases Management.....	5
1.2 Statement of the Problem.....	7
1.3 Rationale of the Study.....	8
1.4 Research Objectives.....	9
1.4.1 General Objective	9
1.4.2 Specific Objectives	9
1.5 Research Questions.....	9
1.6 Significance of the Study	9
1.7 Delineation of the Study	10
CHAPTER TWO	11
LITERATURE REVIEW	11

2.1	Introduction.....	11
2.2	Conceptual Definitions	11
2.2.1	Malaria Disease.....	11
2.2.2	Resource-Poor Country, Area, Settings	11
2.2.3	Disease Diagnosis	12
2.2.4	Clinical Diagnosis.....	12
2.2.5	Self-Medication/ Self- Treatment	12
2.2.6	Presumptive Treatment	13
2.2.7	Machine Learning	14
2.3	Theoretical Literature Review	14
2.3.1	Problem Definition.....	15
2.3.2	Data Identification and Collection.....	15
2.3.3	Data Cleaning and Pre-Processing.....	16
2.3.4	Machine Learning Model Development	16
2.3.5	Model Evaluation.....	18
2.3.6	Model Validation	19
2.4	Empirical Literature Review.....	20
2.4.1	Malaria Diagnosis and Presumptive Treatment of Malaria	20
2.4.2	Self-Medication.....	21
2.4.3	Machine Learning for Disease Diagnosis	21
2.4.4	Machine Learning for Malaria Diagnosis	22
2.5	Research Gap	25
2.6	Conceptual Framework of the Study	25
	CHAPTER THREE	27
	CHARACTERISATION OF MALARIA DIAGNOSIS DATA IN HIGH AND LOW ENDEMIC AREAS OF TANZANIA	27

3.1	Abstract.....	27
3.2	Introduction.....	28
3.3	Material and Methods	29
3.3.1	Study Design.....	29
3.3.2	Ethical Clearance	30
3.3.3	Study Area and Scope of the Study	30
3.3.4	Study Population.....	31
3.3.5	Data Collection	32
3.3.6	Data Analysis	32
3.4	Results.....	33
3.4.1	Document Review results	33
3.4.2	Malaria Patients Survey	39
3.5	Discussion	42
3.6	Conclusion	44
	CHAPTER FOUR.....	45
	DEVELOPMENT OF A MACHINE LEARNING MODEL FOR MALARIA PREDICTION	
	45
4.1	Abstract.....	45
4.2	Introduction.....	45
4.2.1	Related Works.....	47
4.2.2	Theoretical Background.....	48
4.3	Materials and Methods.....	49
4.3.1	Dataset Collection and Description	51
4.3.2	Dataset Descriptions and Pre-processing.....	51
4.3.3	Feature Selection.....	53
4.3.4	Prediction Classifiers	55

4.3.5	Machine Learning Classifiers Validation	55
4.3.6	Machine Learning Model Performance Evaluation.....	56
4.3.7	Development of Regional-Specific Malaria Diagnosis Models	56
4.4	Result	56
4.4.1	Feature Selection Results	56
4.4.2	Important Features Validation: Healthcare Worker's Perspective	60
4.4.3	Machine Learning Classifiers Performance with Important Features	63
4.4.4	Development of Final, Regional-Specific Malaria Diagnosis Models	69
4.5	Discussion	76
4.6	Conclusion	79
CHAPTER FIVE		80
MALARIA PREDICTION MODEL VALIDATION		80
5.1	Introduction.....	80
5.2	Validation Process with unseen malaria diagnosis dataset	80
5.3	Validation Results.....	81
5.3.1	Description of Unseen Malaria Diagnosis Dataset	81
5.3.2	The validated important features	82
5.3.3	Model performance on validation dataset.....	84
CHAPTER SIX.....		87
CONCLUSION AND RECOMMENDATIONS		87
6.1	Introduction.....	87
6.2	Summary of the Study Findings	87
6.2.1	Characterisation of malaria diagnosis records	87
6.2.2	Development of machine learning model for malaria diagnosis	88
6.2.3	Validation of the developed machine leaning model for malaria diagnosis	89

6.3	Contributions of the Study	89
6.3.1	Scientific Contributions	89
6.3.2	New knowledge added from the study.....	90
6.4	Recommendations	90
6.4.1	To the government and policymakers.....	90
6.4.2	To the practitioners	90
6.5	Limitation of the Study	91
6.6	Future Research	91
	REFERENCES	93
	APPENDICES	120

LIST OF TABLES

Table 1:	Confusion matrix	18
Table 2:	Summary of machine learning models for malaria diagnosis	24
Table 3:	Reviewed malaria patients records preliminary information.....	34
Table 4:	Malaria symptoms observed with malaria positivity in document review	38
Table 5:	Multivariate analysis of the significant factors to malaria positivity results (a)....	39
Table 6:	Multivariate analysis of the significant factors to malaria positivity results (b) ...	39
Table 7:	Survey respondents' demographics information	40
Table 8:	Malaria symptoms identified by the survey respondents	41
Table 9:	Malaria diagnosis and treatment history	42
Table 10:	Malaria diagnosis dataset features description	53
Table 11:	Regional based Important Features in Malaria Diagnosis	59
Table 12:	Medical doctors' perspective on malaria diagnosis symptoms	61
Table 13:	Medical doctors' perspective on other factors to be considered for malaria diagnosis	62
Table 14:	10-fold CV classification performance evaluation of different classifiers on malaria diagnosis dataset on full features.	63
Table 15:	10-fold CV classification performance evaluation of different classifiers on malaria diagnosis dataset ten important features	64
Table 16:	Summary of classifiers performance on Morogoro dataset (%).....	68
Table 17:	Summary of classifiers performance on Kilimanjaro dataset (%).....	69
Table 18:	Excellent performance metrics results and best classifiers.....	69
Table 19:	Models' performance for predicting malaria. The results are accuracies obtained by models developed (Decision tree (DT) and Random Forest (RF)) using different sets of important features selected (%)	72

LIST OF FIGURES

Figure 1:	World map showing distribution of malaria cases in 2021	2
Figure 2:	Malaria cases and death distribution in 2021	4
Figure 3:	Malaria distribution in Tanzania (SMO, 2020)	5
Figure 4:	Decision tree description	18
Figure 5:	Machine learning model validation process (Datatron, 2022)	20
Figure 6:	Conceptual framework for the study	26
Figure 7:	Study area	31
Figure 8:	Number of records per health facility	34
Figure 9:	Number of patients per month visiting the health facility	35
Figure 10:	Frequency of malaria symptoms observed	35
Figure 11:	Machine learning framework employed in features selection, model development and validation	50
Figure 12:	Important Features with Random Forest in High Endemic Area (Morogoro)	57
Figure 13:	Important features with random forest in low endemic area (Kilimanjaro)	57
Figure 14:	Important features with random forest in combined dataset	58
Figure 15:	AUC, Sensitivity and Specificity performance of different classifiers on full features dataset	64
Figure 16:	AUC, Sensitivity and Specificity performance of different ML classifiers on important features of the whole Malaria diagnosis dataset	65
Figure 17:	AUC, Sensitivity and Specificity performance of classifiers on ten important features on Kilimanjaro dataset	66
Figure 18:	F1 score comparison on the two regions' datasets	67
Figure 19:	Classification Accuracy comparison in two regions dataset	67
Figure 20:	Classifier's Sensitivity comparison on the two regions' datasets	68
Figure 21:	ROC plot for Random Forest performance evaluation	68
Figure 22:	Decision tree for low endemic area (Kilimanjaro)	73

Figure 23: Decision tree for high endemic area (Morogoro).....	74
Figure 24: Decision tree for combined malaria diagnosis dataset	75
Figure 25: Frequency of residence area and age of the patient.....	81
Figure 26: Frequency of patients based on their visit month.....	82
Figure 27: Patients distribution based on their Sex	82
Figure 28: Important features in the two regions.....	83
Figure 29: Important features in low endemic area	83
Figure 30: Important features in high endemic areas.....	84
Figure 31: Prediction Accuracy and F1 score for the validation dataset	85
Figure 32: Precision and recall scores for the validation dataset.....	85
Figure 33: Confusion matrix of DT and RF on a validation dataset.....	86

LIST OF APPENDICES

Appendix 1:	Study ethical clearance from NIMR	120
Appendix 2:	Malaria Patients Records Collection Form	121
Appendix 3:	Questionnaire Used for Patient’s Survey	122
Appendix 4:	Questionnaire Used for Medical Doctors Survey	125
Appendix 5:	Python code that were employed for features selection.....	128
Appendix 6:	Python code that was employed for regional model development	130
Appendix 7:	Research Outputs - Poster and Published Papers.....	131

ABBREVIATIONS AND SYMBOLS

AUROC	Area Under the Receiver Operating Characteristic Curve
BS	Blood Slide
CDC	Centre for Disease Control
DT	Decision Tree
FPR	False Positive Rate
GDP	Gross Domestic Product
IPT	Intermittent Presumptive Treatment
IRS	Indoor Residual Spraying
ITN	Insecticide Treated Net
KNN	K-nearest Neighbors Algorithm
ML	Machine Learning
MRDT	Malaria Rapid Diagnosis Test
NMCP	National Malaria Control Protocol
PPT	Periodic Presumptive Treatment
RF	Random Forest
SVM	Support Vector Machine
TPR	True Positive Rate
WHO	World Health Organization
LR	Logistic Regression
AUC	Area under the Curve
CV	Cross Validation
AUC	Area under the ROC Curve
CSV	Comma-Separated Values
RDT	Rapid Diagnostic Test
NIMR	National Institute for Medical Research in Tanzania
OTC	Over The Counter

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Malaria is a disease caused by the plasmodium parasite transmitted by the bite of an infected female anopheles mosquito. Malaria remains a substantial public health issue in sub-Saharan Africa, accounting for around 1 million fatalities and more than 400 million cases annually (SMO, 2020). According to reports, there will be 229 million instances of malaria and 409 000 deaths from it in 2019, with 94% of these deaths occurring in Africa (WHO, 2020). Tanzania accounted for 5% of the world's malaria deaths in 2019, concentrated in 31 nations. (WHO, 2020; WHO AFRICA, 2018). Although there was a tremendous decrease in malaria death from 435 000 in 2017 to 409 000 in 2019 globally, malaria cases have increased (WHO, 2019). Tanzania reported more than six million (6 000 000) malaria confirmed and presumed cases and more than 21 000 deaths which are 5.18% of total death in 2019 (SMO, 2020). Since malaria is regarded as treatable and preventable, reducing the burden of illness and fatalities while maintaining the long-term goal of eradicating malaria is a global concern. According to a National Malaria Control Program, Tanzania has achieved significant strides in the previous few decades in ensuring access to malaria control measures, but the nation is stated to be still far from its intended goal of eliminating malaria at a large percentage (Group *et al.*, 2017; Ngasala & Bushukatale, 2019; SMO, 2020; USAID, 2018). The number of people affected by malaria continues to rise, with about two million deaths anticipated yearly despite international efforts to combat the disease. Treatment adherence, effectiveness, and clinical care of severe malaria cases continue to be severely hampered by the absence of proper diagnosis (Andrade *et al.*, 2010). Families living in poverty in rural areas are disproportionately disadvantaged when accessing modern health care.

1.1.1 Global Malaria Burden

Malaria is one of the most severe issues affecting public health worldwide. In 87 countries and territories, people live in locations with a danger of contracting malaria, affecting about half of the world's population, as depicted in Fig. 1. According to the World Malaria Report published in 2020 by the World Health Organization (WHO), malaria is the top cause of mortality and disease in several underdeveloped nations. Thirty-one (31) countries accounted for around 95% of all malaria deaths. About 51% of all deaths from malaria in 2019 occurred in just ten

countries: Nigeria (23%), the Democratic Republic of the Congo (11%), the United Republic of Tanzania (5%), Mozambique (4%), Niger (4%), and Burkina Faso (4%) The African Region of WHO is responsible for 82% of all cases and 94% of all deaths worldwide. In 2019, there are 46 nations where malaria is a problem, up from 26 in 2000 (WHO, 2019, 2020; WHO AFRICA, 2018).

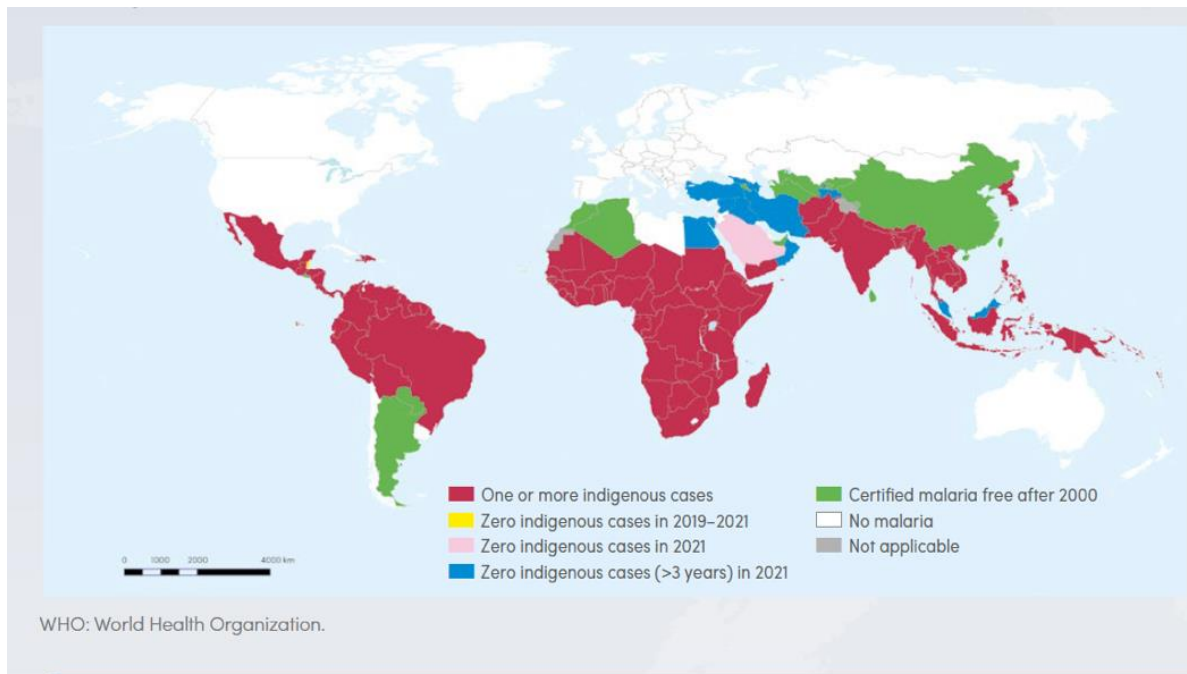


Figure 1: World map showing distribution of malaria cases in 2021

Malaria imposes high costs on both individuals and governments in the endemic areas. Drugs for treating malaria at home, transportation to and from dispensaries and clinics, missed work days or school, preventative measures, and burial costs in the event of death all add up to a hefty price tag for individuals and their families in malaria management (Belachew Gutema *et al.*, 2011; Kajeguka *et al.*, 2017; Frøkjær *et al.*, 2012; Marealle & Kirutu, 2018; Mwita *et al.*, 2019). Public health measures against malaria, such as insecticide spraying or distribution of insecticide-treated bed nets, have a cost, as do lost days of work and the income they would have generated, as well as the opportunity cost of lost tourists and collaborative economic ventures. The annual loss in GDP growth due to direct expenses (such as disease, treatment, and premature mortality) has been estimated to be at least 1.3%, or US\$ 12 billion (CDC, 2021).

1.1.2 Malaria in Tanzania

Tanzania is a country in Eastern Africa, bordered by the great lakes of Victoria to the north, Tanganyika to the west and Malawi to the south. It comprises a mainland and the Zanzibar

archipelago. More than ninety-three per cent of Tanzania's mainland population is in malaria-endemic regions. Tanzania ranks seventh among the top ten countries with the highest malaria infection and mortality rates (3% of global cases, 13.4% of patients in East and Southern Africa, and 4% of global deaths) (WHO, 2019), as shown in Fig. 2. Despite stagnation in case incidence between 2015 and 2018, with rates hovering at 122–124 per 1000 of the population at risk, deaths decreased by around 4 per cent, from 0.4 to 0.38 per 1000 of those at risk. Recent years have shown a rise in the rate of new cases and deaths from them. Over the past few years, Zanzibar has seen a steady decline in the number of cases of malaria, and in 2018, the positive rate among patients seeking treatment was only 1.3%. Nonetheless, confirmed malaria cases rose from 4171 in 2017 to 5146 in 2018, with five deaths attributed to the disease (USAID, 2018). According to the Tanzania Commission on AIDS, the risk of malaria transmission and prevalence in Tanzania varies significantly between 1% and 33%, with an average of around 10%, as seen in Fig. 3 (SMO, 2020; Thawer *et al.*, 2020).

The prevalence variation between places and times of the year is affected by climatic and non-climatic factors (Chirombo *et al.*, 2020). Climatic factors, including temperature, rainfall and relative humidity, greatly influence the pattern and levels of malaria (Hagenlocher, 2015; Snow, 2005; Rumisha *et al.*, 2019). Non-climatic factors influencing malaria risk include vectors, parasite species, host immunity, insecticide and drug resistance, environmental development and urbanisation, population movements, and other socio-economic factors, including livelihoods. Ninety-three per cent (93%) of mainland Tanzania's population resides in malaria-endemic areas. In 2015, there were estimated to be 7.3 million clinical and confirmed cases of malaria reported in the country (Michael & Mkunde, 2017).

Tanzania ranks seventh among the top ten countries with the highest malaria infection and mortality rates (3% of global cases, 13.4% of patients in East and Southern Africa, and 4% of global deaths) (WHO, 2019) as shown in Fig. 2.

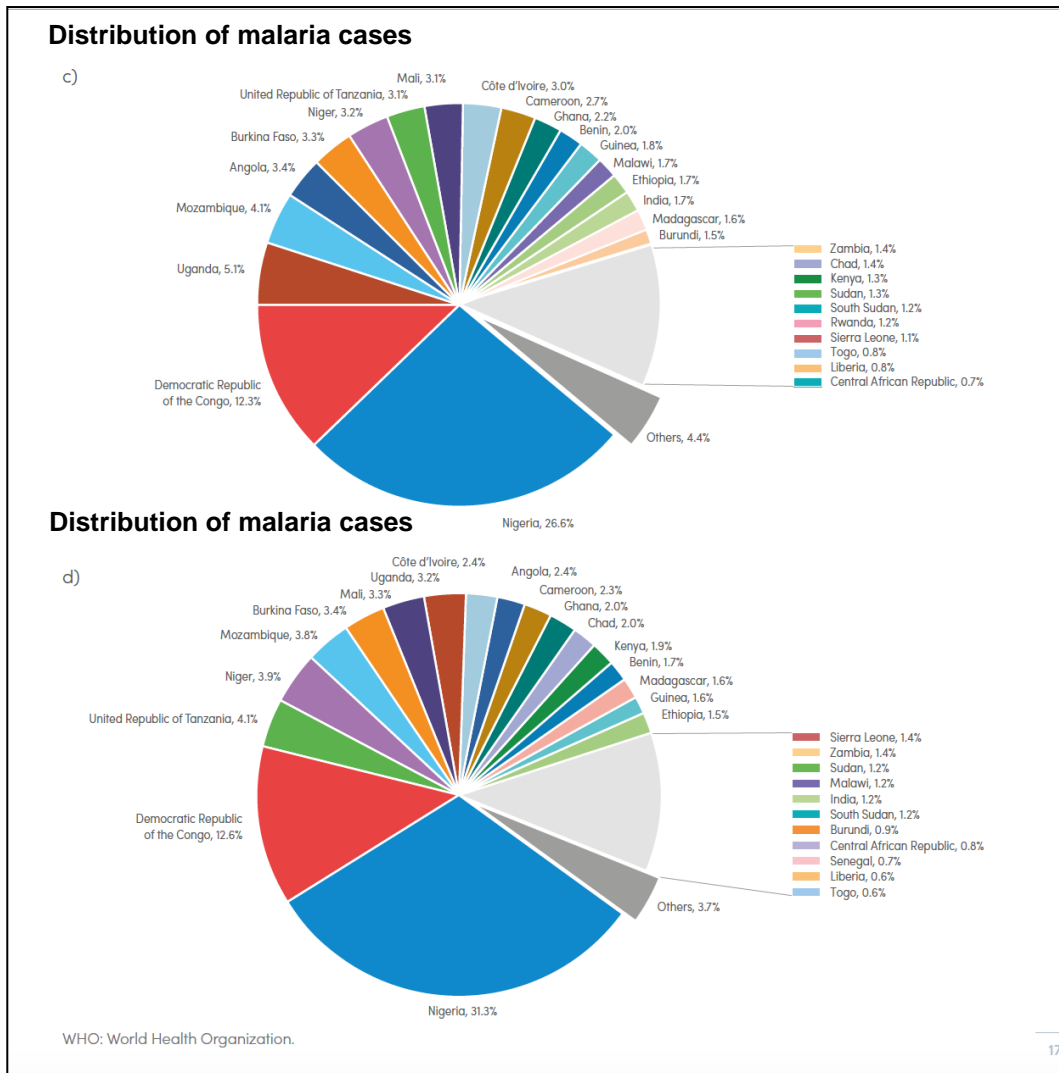


Figure 2: Malaria cases and death distribution in 2021

Three types of malaria transmission seasons exist in the country: stable perennial transmission, which accounts for 60% of the country, stable malaria transmission (with seasonal variation), which accounts for 20% of the country, and unstable seasonal transmission, which accounts for about 20% of the country (SMO, 2020). The poorest 46% of the population have the highest risk of contracting malaria and progressing to severe cases because of their living conditions and lack of access to effective treatment and malaria control measures. Even while the overall malaria burden is still relatively high, the recent rise in this number should serve as a wake-up call and prompt the creation of innovative methods for controlling the disease. Preventing and treating malaria is possible wherever in the world. While the ultimate goal of malaria eradication is essential, the immediate objective is to lessen the burden of illness and mortality. The frequency of malaria varies throughout regions in Tanzania, as shown in Fig. 3.

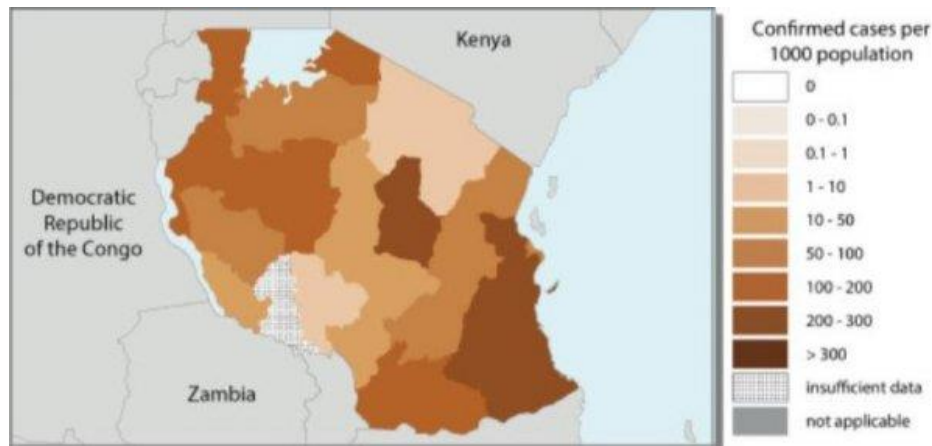


Figure 3: Malaria distribution in Tanzania (SMO, 2020)

According to reports from the National Malaria Control Programme, Tanzania has made significant progress over the previous several decades through various programmes and efforts to guarantee access to malaria control therapies (National Malaria Control Programme *et al.*, 2013). Insecticide-treated nets (ITN), indoor residual spraying (IRS), intermittent preventative treatment (IPT), and case management, which includes diagnostic tests for patients of all ages, are all examples of measures that are being implemented to combat malaria. However, the country reportedly has a long way to go before it reaches the planned National Malaria Control Programme (NMCP) aim of eliminating malaria by a significant percentage. This is because the NMCP relies on accurate health statistics, which are currently lacking (WHO, 2018). Therefore, a more robust healthcare infrastructure is essential for precise malaria diagnosis, effective treatment, and vigilant monitoring.

1.1.3 Malaria Cases Management

Malaria case management, the process of identifying and successfully treating malaria cases, remains crucial to malaria control and elimination efforts. The WHO has issued a guideline on the diagnosis and treatment of malaria that has to be followed by health professionals in managing malaria cases (WHO, 2019; WHO-Guidelines, 2015). The 2014 Tanzania Mainland’s National Guidelines for the Diagnosis and Treatment of Malaria aligns with this guideline (Group *et al.*, 2017). In malaria-endemic areas, the World Health Organization (WHO) recommends that patients with a history of fever or temperature of 37.5 °C who have no other apparent explanation should be tested for malaria. The disease should also be suspected where it is routinely transmitted (or during the high-transmission period of seasonal malaria). Comprehensive parasitological testing of all fever cases may not be cost-effective

when malaria prevalence is low. Healthcare providers in these locations should be trained to recognise patients who may have been exposed to malaria (e.g., recent unprotected travel to a malaria-endemic area) and who present with a fever or a history of fever without an apparent reason before ordering a parasitological test. The parasitological diagnosis should be readily available when a patient is brought in (within 2 hours). Without a parasitological diagnosis, the decision to treat with antimalarials must be based on the probability that the disease is malaria (WHO Guidelines, 2015).

The absence of a definitive diagnosis is a serious obstacle to treatment compliance, efficacy, and clinical care of severe malaria cases. The current Malaria Reduction Strategic Plan recommends using parasite-based diagnosis with Microscopy Blood Slides (BS) and Malaria Rapid Diagnostic Test (MRDT) as part of malaria case management across all health facility levels for all age groups and targeted groups (children under five) at the community level (2014–2025) (Wang *et al.*, 2019; WHO, 2015). Even though the current diagnostic and treatment guidelines have been implemented in public health facilities in most countries, policy compliance is still far from ideal, with some malaria diagnoses still based on clinical symptoms and inconsistent therapeutic intervention based on test results. In many nations with a high incidence of malaria, issues including overdiagnosis, overprescribing of malaria treatments, and a lack of malaria medicine stockpiles are well-known issues. For healthcare workers who work in rural areas with limited resources, the long-standing practice of treating all fevers as malaria has significantly changed with the development of MRDTs. For rapid diagnostic testing to be effective, healthcare providers must trust the accuracy of the tests and take action based on the findings. By adopting parasite-based testing, healthcare professionals can feel more confident diagnosing individuals with malaria (Altaras *et al.*, 2016). Once a diagnosis result is established, more information is needed to successfully manage a negative malaria diagnosis since parasite-based diagnosis is challenging in most health institutions for reasons other than the ineffectiveness of the diagnostic techniques.

The current policy for treating malaria is that it is better to treat numerous cases of non-malarial febrile fever with an antimalarial treatment than to miss one real case in an era when antimalarial drugs are cheap and inexhaustible (Nadjm *et al.*, 2010). However, in some regions with insufficient resources and diagnostic techniques, malaria has been improperly diagnosed, with the widespread belief that each incidence of fever must be caused by malaria (Sapkota *et al.*, 2010). Apart from that, people do not visit health centres for diagnosis but rather visit

pharmacies (Menard & Dondorp, 2017). Therefore, malaria management needs to upgrade disease diagnostic tools and procedures. As a result, a better approach enables correct, concise, and quick malaria diagnosis for patients in resource-limited areas. Fever has been the most common complaint among outpatient clinic patients, with malaria being a likely cause of such febrile diseases.

In many African countries, malaria is the primary medical diagnosis. However, the low specificity of malaria symptoms and signs restricts the accuracy of clinical diagnosis (Reyburn *et al.*, 2007). In light of current antimalarial treatment guidelines, it is advised to treat multiple episodes of non-malarial febrile illness with an antimalarial drug rather than failing to treat a single malaria case (Winskill *et al.*, 2011). The development of a machine learning-based approach will aid medical experts in determining whether or not a patient with a negative diagnosis is clear of malaria and provide information on other probable conditions the patient may be experiencing. In addition, individuals who want to self-medicate should ensure they have the condition by having malaria-related symptoms evaluated by the model. Finally, health officials require accurate and reliable predictions of the disease's occurrence to contain a malaria epidemic effectively.

Consequently, this research aims to create a machine learning model for detecting malaria based on symptoms and non-symptomatic characteristics, such as patient demographic information. Patient demographics, symptoms, and test results for malaria were gathered from the patient's healthcare records. The collected records were computed under the model's specifications to aid in correct diagnosis.

1.2 Statement of the Problem

Over 229 million cases and 409 000 fatalities worldwide in over 87 endemic countries demonstrate that malaria is still one of the world's worst infectious illnesses (Bria *et al.*, 2021). Mortality from malaria is overwhelmingly concentrated in Africa (94% of global deaths), which is also the leading cause of illness overall (Caminade *et al.*, 2014). The National Malaria Control Programme reports that Tanzania, accountable for 5% of worldwide malaria deaths in 2019, has made significant progress toward securing the complete elimination of the illness through programmes such as countrywide Malaria control programmes, insecticide-treated nets, indoor residual spraying and intermittent preventive therapy. Although there have been significant advances in both preventing and treating malaria, it remains a severe threat to public

health (Dhiman, 2019; Patouillard *et al.*, 2017). Prompt, precise, and appropriate malaria diagnosis and quick treatment are paramount for managing malaria. The World Health Organization (WHO) recommends conducting a parasitological test on anyone with malaria symptoms, regardless of where they are. Public health facilities worldwide have adopted pre-existing norms for diagnosis and treatment, but enforcing compliance has proven difficult. In most impoverished nations, especially rural areas, malaria is diagnosed and managed through presumed therapy and self-medication with antimalaria medications (Gosling *et al.*, 2008; UM, 2016). According to reports, presumptive therapy and self-medication are increasing (Ansumana *et al.*, 2013). The existence of massive amounts of patient records has the potential to support a wide variety of medical and healthcare support systems, such as, among others, clinical decision support, disease surveillance, and population health management, all of which are motivated by mandatory requirements and the possibility of improving the quality of healthcare delivery and the diagnosis of malaria while reducing costs.

Despite the tremendous rise in machine learning in tackling social issues, there is a lack of a machine learning-based malaria detection model that will leverage patients' symptoms and demographic information. Most modern methods for diagnosing malaria rely on studying blood smears' microscopic images. The challenge is that not all hospitals have access to the necessary technology or trained personnel to interpret these images correctly. It's worth noting that some people habitually treat themselves with antimalaria drugs whenever they observe malaria-related symptoms such as fever.

1.3 Rationale of the Study

This study aims to develop a machine learning-based model for malaria diagnosis. While using machine learning models to assist malaria diagnosis in previous studies, most of these studies focused on analysing microscopic images to help in fast and accurate malaria diagnosis. While these are good strategies for accurately diagnosing malaria, presumptive treatment and self-medication still need to be addressed in properly managing malaria. Development of antimalarial drugs resistance, misuse of drugs and untreated friable diseases are a few global concerns regarding presumptive treatment and self-medication. Machine learning is the field that assesses and learns from data to identify various patterns and assist in making decisions. While machine learning is currently being used for other systems, there is potential for the technology to do much more in properly diagnosing malaria. The significant benefits of the machine learning model for malaria diagnosis are accuracy in learning from datasets and fast

and easy prediction of malaria. The developed models would be expected to be used for early malaria prediction for patients who have experienced malaria-related symptoms before using any antimalarial drugs or when the proper diagnosis is unavailable to avoid unnecessary antimalarial drugs.

1.4 Research Objectives

1.4.1 General Objective

The main objective of this study is to develop a malaria machine learning model for the prediction of malaria in low and high-endemic areas of Tanzania.

1.4.2 Specific Objectives

The following specific objectives were used as a guide towards achieving the main objective:

- (i) To analyse available malaria diagnosis data for model training and validation
- (ii) To develop a malaria prediction model to improve malaria diagnosis in low and high-endemic areas of Tanzania.
- (iii) To validate the performance of the developed malaria prediction model.

1.5 Research Questions

- (i) What attributes/variables within malaria patient records can be used to train and validate the malaria diagnosis model?
- (ii) What is the suitable model for malaria prediction to improve malaria diagnosis in low and high endemic areas of Tanzania?
- (iii) How does a proposed malaria prediction model improve malaria diagnosis in low and high endemic areas of Tanzania?

1.6 Significance of the Study

This research work's findings will benefit malaria patients, health facilities, the government, policymakers, malaria management programmes and the academic world. Malaria patients can confirm their malaria status before self-medicating themselves with anti-malaria drugs

whenever they observe malaria symptoms. Health facilities with inadequate diagnostic tools can confirm their patient's status before presuming that the cases are malaria. The research work will inform malaria treatment policymakers on the regional-specific key features significant in malaria diagnosis. More importantly for the government, this work improves current knowledge of malaria management programs by demonstrating how sustainable machine learning can be accommodated in malaria diagnosis. This research work also paves the way for developing an artificially intelligent tool that will be used for malaria prediction in the absence of proper testing tools in health facilities and in events of self-medication among malaria patients. Apart from that, this research work is expected to inspire other researchers to look into and enhance the diagnosis of other diseases in the country using machine learning.

1.7 Delineation of the Study

This study focuses on using machine learning to assist in diagnosing malaria. Machine learning is a branch of artificial intelligence based on the idea that the system/computer can learn from data to identify patterns, predict future events and make decisions with minimal human intervention without being explicitly programmed. The algorithms use statistical analysis to predict output and update results as new data becomes available. The programmer does not directly provide machine learning instructions. Thus, the study aimed to construct a machine-learning model to fit the given dataset. The programs designed had to perform a repetitive feature and model selection process and modify various algorithm parameters to obtain a robust model for malaria diagnosis. Model development in this study was defined as taking data collected from patients' records, analysing them and using results to predict the future patient's state of malaria based on the symptoms and non-symptomatic factors presented.

The malaria diagnosis features are defined as the variables presented by the patient. For this study, two features identified include malaria-related symptoms presented or observed by the patient and non-symptomatic factors such as the patient's demographic information. A patient indicated in this study is a person who has observed malaria-related symptoms such as headache and fever and is ready to seek treatment.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter discusses and reviews various theories relating the malaria diagnosis and machine learning models. The chapter also describes the related studies on machine learning for malaria diagnosis, including the research gap in this study. Finally, the conceptual framework to guide the study was developed and discussed. The overall goal of this chapter is to understand the concept of malaria diagnosis, self-medication, and machine learning and create a relationship.

2.2 Conceptual Definitions

2.2.1 Malaria Disease

Malaria is a parasitic disease. Humans contract the parasite from the bites of mosquitoes carrying the disease (Barber *et al.*, 2017; Lozano *et al.*, 2020). Malaria patients frequently experience severe illness, including high fever and chills (Sanchez *et al.*, 2020). The disease is brought on by parasites called Plasmodium. Female Anopheles mosquitoes are malaria vectors responsible for transmitting these parasites to humans through their bites. Five (5) different parasite species can infect people, but the most dangerous ones are Plasmodium falciparum and Plasmodium vivax (SMO, 2020; WHO, 2019). Malaria is a severe, rapidly progressing fever. After being bitten by an infected mosquito, it typically takes between 10 and 14 days for symptoms to manifest in a person who is not immune. Malaria presents first with fever, headache, and chills, although these symptoms may be mild and hard to identify. Malaria caused by P. falciparum, if left untreated for more than 24 hours, can cause severe sickness and even death.

2.2.2 Resource-Poor Country, Area, Settings

Resource-poor countries and regions only have access to minimal equipment, supplies, and personnel when treating patients with life-threatening illnesses. Three levels of resource scarcity have been identified by researchers (Geiling *et al.*, 2014; Yapa & Bärnighausen, 2018): absence of resources, limited resources, and limited resources with potential referral to higher care capability. The majority of developing nations, particularly those in sub-Saharan Africa, are resource-poor (Barber *et al.*, 2017; Clair *et al.*, 2017; Yapa & Bärnighausen, 2018)

2.2.3 Disease Diagnosis

Diagnosis refers to determining the nature of a health problem by examining a patient's symptoms (Zimmerman & Howes, 2015). Hawkes and Kain (2014) state that diagnosis aims to determine the presence and nature of a disease, ailment, or damage by observing the patient's signs and symptoms. Diagnosis may involve the patient's medical history, a physical examination, and several testing (such as blood work, imaging studies, and biopsies). For example, patients are examined to see if they have been infected with malaria. There are three different ways malaria can be diagnosed: through a microscopic examination, a quick diagnostic test, or a clinical evaluation (Prevention, 2019). Clinical diagnosis is based on the patient's symptoms and exam results. Therefore, clinical diagnosis is based on the patient's symptoms and exam results.

2.2.4 Clinical Diagnosis

A patient's symptoms and physical examination results form the basis of a clinical diagnosis (Tangpukdee *et al.*, 2009). Fever, headache, weakness, myalgia, chills, dizziness, abdominal discomfort, diarrhoea, nausea, vomiting, anorexia, and pruritus are some of the early symptoms of malaria, which are pretty general and changeable (CDC, 2021; Tangpukdee *et al.*, 2009). Malaria parasites can be recognised microscopically by spreading a drop of the patient's blood out as a "blood smear" on a microscope slide. Parasites in a specimen are given a unique look by staining it (often with the Giemsa stain) before it is examined. This method is still the most reliable for diagnosing malaria in the lab. However, the results will vary according to the calibre of the laboratory's reagents, microscope, and technicians (Kumar *et al.*, 2021). Rapid diagnostic tests for malaria rely on detecting specific antigens made by the parasite that causes the disease. Therefore, parasite density is a crucial criterion in monitoring a patient undergoing treatment for severe malaria. However, the tests are not quantitative, and the association between antigen concentration and parasitaemia is not well known (Gillet *et al.*, 2011; Mouatcho *et al.*, 2013).

2.2.5 Self-Medication/ Self- Treatment

The World Health Organization (WHO) defines self-medication as the use of medications to treat diseases or symptoms that one has self-diagnosed or the use of prescribed medication on an as-needed basis to address chronic or recurrent sickness or symptoms (Alghanim, 2011). Taking medicines or drugs to cure a problem without consulting a doctor is known as self-medication or self-treatment (Barber *et al.*, 2017; Lozano *et al.*, 2020). People who self-

medicate use substances to lessen the adverse effects of their mental illness or its treatment (Ansari, 2018). Self-medicate is using medication for one's diagnosis and treatment without consulting a medical professional (Alefian & Halboup, 2016). Misuse of OTC medications, using many medicines at once, and using home remedies for potentially life-threatening disorders are all examples of how self-treatment can go wrong and lead to incorrect diagnoses or even the concealment of health issues (Ansumana *et al.*, 2013; Gil-Rivas & McWhorter, 2013; Sundram & Pereira, 2007). To self-medicate is to choose and utilise medication (including herbal and traditional remedies) to cure one's perceived health problems.

2.2.6 Presumptive Treatment

Periodic Presumptive Treatment (PPT) refers to the standard delivery of presumptive treatment (Population Council & WHO, 2008). Presumptive treatment treats patients with clinical suspicions before or without confirmation of laboratory findings (Graz *et al.*, 2011). Presumptive therapy is giving antimalarial medicine to someone who may have malaria before they have been tested or before blood tests can be performed (Graz *et al.*, 2011; Nadjm *et al.*, 2010; WHO, 2019a). Individuals or populations at high risk of disease may receive one-time "presumptive treatment," in which they are given medication to combat an infection based on the assumption that they have it. For example, if malaria is suspected but a definitive diagnosis cannot be made, the situation is called an assumed case.

Additionally, intermittent presumptive treatment (IPT) manages malaria among pregnant women. Pregnant women can benefit from this method, which entails dosing them with a curative dose of an efficient antimalarial medicine at regular intervals during their pregnancies. *Plasmodium falciparum* malaria is particularly harmful to pregnant women; hence IPT was initially implemented in high-transmission areas (White, 2005; Yeung & White, 2005). Antimalarial medication should be administered based on clinical suspicion in areas where a parasitological diagnosis is not feasible. Any patient presenting with a history of fever or temperature of 37.5 °C in a malaria-endemic area should be evaluated for malaria. If a child has palmar pallor or a haemoglobin level of 8 g/dL or less and lives where malaria transmission is consistent (or during the high-transmission phase of seasonal malaria), then malaria should be considered. Many areas of sub-Saharan Africa and some of Oceania are considered high-transmission settings.

2.2.7 Machine Learning

To discover patterns and generate meaningful classifications based on the correlation of each variable with the disease outcome, machine learning methods employ algorithms based on statistical assumptions and mathematical principles (Lee *et al.*, 2021; Morang'a *et al.*, 2020). It is a subfield of AI that seeks to automate as many mundane tasks as possible by teaching computers to learn from data and spot patterns independently (Dash *et al.*, 2021). In data science, the choice of the algorithm relies on the nature of the data being predicted. The methods by which a classical machine learning algorithm improves its predictive abilities are commonly used to classify the field. There are four primary machine learning methods: Data scientists employing supervised machine learning provide algorithms with labelled training data and identify the variables they want the computer to examine for correlations. Unsupervised learning is a subfield of machine learning in which algorithms are trained using unlabelled data rather than labelled data. The programme searches all of the data to find patterns. Semi-supervised learning is a machine learning strategy that combines deterministic and nondeterministic training data and presents prediction or recommendation output (Jiang *et al.*, 2017). In other cases, data scientists only use labelled training data to feed an algorithm. The model may still independently investigate the dataset and grow its knowledge of it (Jiang *et al.*, 2017); semi-supervised learning is the approach to machine learning involving a mix of the two preceding types. Data scientists may feed an algorithm mostly labelled training data. However, the model is still free to explore the data independently and develop its understanding of the data set (Davenport & Kalakota, 2019) and Reinforcement learning, which data scientists use to train computers to carry out complex tasks following previously established principles. Algorithms can be trained by data scientists who provide them with positive and negative reinforcement as the algorithms determine how best to carry out their tasks. However, the algorithm typically makes autonomous decisions about what to do next (Uddin *et al.*, 2019).

2.3 Theoretical Literature Review

Machine learning methods employ statistical assumptions and mathematical principles to classify data and predict disease outcomes (Lee *et al.*, 2021; Morang'a *et al.*, 2020). These machine-learning models enhance the quality of care provided to patients by medical professionals (Liang *et al.*, 2017; Sriporn *et al.*, 2020). Following these guidelines will help ensure that your machine-learning model produces reliable results. Methods and theories for creating machine learning models in general and healthcare-oriented models, in particular, are

discussed below. Machine learning models are typically developed using these standard procedures: problem definition, data identification, data cleaning and pre-processing; model construction; model evaluation; model improvement; model validation (Chen *et al.*, 2019).

2.3.1 Problem Definition

Defining the problem and a prediction task is the first step in developing a machine learning model. A successful machine learning model in healthcare is expected to impact patient care by providing actionable insights. This step aims to understand the requirement and the problem that needs a solution before attempting to code it. The best approach to problem definition is to understand the project's objective. Secondly, reshape the obtained knowledge to define the problem. Lastly, formulate an opening plan for attaining the goals of the project.

The primary goal of this study is to create a machine learning model that can determine if a patient has malaria or not based solely on their symptoms, non-symptomatic factors like the patient's demographics and travel history. The aspect of a dataset you want to understand more thoroughly is its target variable. A supervised machine learning method uses previous data to identify patterns and find connections between the goal and other elements of your dataset. Whatever the input variables' outcome, that is what the target is. In the case of a classification problem, it might be the specific classes to which the input variables might be mapped or the possible range of output values in the case of a regression problem. The aim is the training output values that will be taken into account if the training set is taken into account. The aim variable in creating a machine learning model for diagnosing malaria is diagnostic, which can be either positive or negative. The patients' demographic data (residence area, age, and sex), malaria-related symptoms, and diagnosis outcome are the input data or variables for the malaria diagnosis model. These data points were gathered based on the information provided by individuals when they visit a healthcare facility and request treatment for symptoms associated with malaria. In the other research, the same characteristics were likewise used to control malaria (Baltzell *et al.*, 2019; Bria *et al.*, 2021). We are addressing a classification issue in this study. A supervised learning method, a classification task, allows the computer programme to learn from the data and generate new observations or classifications (Sarkar *et al.*, 2018).

2.3.2 Data Identification and Collection

This is the first natural step toward developing a machine-learning model and collecting the data. This critical step will cascade into how good the model will be. The more and better data

collected, the better the model can perform. For example, the dataset used to build a malaria diagnosis model was created from scratch using malaria patients' records. These records were collected from the patient's files in the health facilities. The records were collected using a designed form to mirror the information about patients visiting the health facility with malaria-related symptoms.

2.3.3 Data Cleaning and Pre-Processing

Data preparation consists of cleansing, augmentation, normalisation, aggregation, transformation, and labelling. This step involves pre-processing data by eliminating, normalising, error corrections, and removing duplicity.

2.3.4 Machine Learning Model Development

This step aims to achieve close to 100% accuracy in the model's performance. The fit of the algorithm, the completeness of the feature set, and the sufficiency of training data are the three main components determining machine learning models' accuracy. The modelling process is repeated until the required accuracy is achieved or progress has stalled.

The first step in creating a model is to decide on an appropriate algorithm. The algorithm is used to develop or train the model using the training data. A good model that can become a good business tool requires the correct algorithm for a given machine-learning problem. This study uses the most common supervised machine learning classifiers to build a malaria diagnosis model (Uddin *et al.*, 2019). The popular machine learning classifiers for disease diagnosis are Logistics Regression (LR), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), which were used in the model development (Ibarra *et al.* 2021). These algorithms were adopted since they have proven to work best with healthcare datasets, as shown in other studies done by Aminu *et al.* (2016), Ghumbre and Ghatol (2012), Iyer *et al.* (2015), Laghmati *et al.* (2019), Mishra *et al.* (2019), Mohan *et al.* (2019), Priyadarshini *et al.* (2014), Ritthipravat, (2009), and Sengar *et al.* (2020).

Logistic Regression, also known as LR, is a method for supervised classification that is both reliable and widely used. It is possible to think of it as an extension of regular regression. The sole variable it can model is a dichotomous one, which typically indicates the occurrence or non-occurrence of an event. This approach aims to determine the likelihood that the newly

created instance belongs to a specific class. Since it is a probability, the result is between 0 and 1 (Swaminathan *et al.*, 2017; Ullah *et al.*, 2019).

The K-Nearest Neighbour (KNN) technique can be used. Researchers from other fields worked together to develop this method. $D = \{x_i \mid i = 1, \dots, N\}$ where $x_i \in \mathbb{R}^d$ is the i th data point as the input to the method. For the initial step of the algorithm, a cluster of k points is chosen in \mathbb{R}^d . Initial seeds can be selected using various techniques, such as random sampling, clustering, and perturbing the global mean of the data k times (Krishnani *et al.*, 2019; Patil *et al.*, 2018).

The Support Vector Machine (SVM) algorithm can classify linear and non-linear data. It usually starts by mapping each data item into an n -dimensional feature space, where n denotes the number of features. The hyperplane that separates the data items into two classes is then described, with the marginal distance for both types best realised and classification errors significantly reduced (Krishnani *et al.*, 2019; Ibarra *et al.*, 2021; Ullah *et al.*, 2019).

Decision Tree (DT) is one of the earliest and most prominent machine learning algorithms. A decision tree models the decision logic, i.e., tests and corresponds to outcomes for classifying data items into a tree-like structure. The nodes of a DT tree typically have multiple levels where the first or top-most node is called the root node, as shown in Fig. 4. All internal nodes (i.e., having at least one child) represent tests on input variables or attributes. Depending on the test outcome, the classification algorithm branches toward the appropriate child node, where the process of examination and branching repeats until it reaches the leaf node. The leaf or terminal nodes correspond to the decision outcomes. Decision Trees have been found easy to interpret and learn quickly and are a common component of many medical diagnostic protocols. When traversing the tree for sample classification, the outcomes of all tests at each node along the path will provide sufficient information to conjecture about its class (Krishnani *et al.*, 2019; Saranya & Pravin, 2020; Swaminathan *et al.*, 2017).

A Random Forest (RF) is a type of ensemble classifier made up of numerous Decision Trees (DTs), just as a forest is a collection of many individual trees. When DTs are allowed to grow exceedingly deep, it is common for the training data to overfit, which leads to a high degree of variation in the classification results for a given level of change in the input data. In addition, they are biased by their training data, which makes them susceptible to errors when applied to the test dataset (Azar *et al.*, 2014; Chen *et al.*, 2019; Iyer *et al.*, 2015).

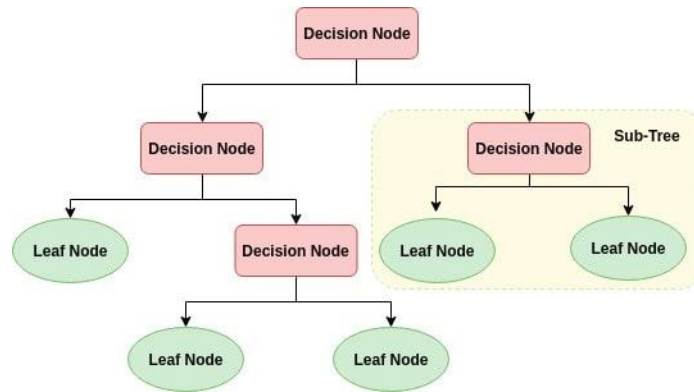


Figure 4: Decision tree description

2.3.5 Model Evaluation

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. Each model was evaluated with variable sets of features selected by the different feature selection algorithms. In addition, models were evaluated using prediction accuracy for classification problems (being malaria positive or negative). Therefore, the confusion matrix provides a more insightful picture of the performance of a predictive model and which classes are being predicted correctly and incorrectly, and what type of errors are being made. This illustrates how the four-classification metrics are calculated (True Positive (TP), False Positive (TN), False Negative FN, True Negative (TN)), and the predicted value compared to the actual value as shown in Table 1. Classification accuracy, AUROC (Area Under the Receiver Operating Characteristics), Precision, Specificity, Sensitivity/recall, and F1 score were used as evaluation matrices for this study.

Table 1: Confusion matrix

		Actual Value	
		Positive	Negative
Predictive Value	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

Note:

- True Positive – Observation is positive and is predicted to be positive
- False Positive - Observation is negative but is predicted to be positive
- True Negative - Observation is negative and is predicted to be negative
- False Negative - Observation is positive but is predicted to be negative

Classification accuracy is the percentage of correctly predicted cases relative to all examined examples. How many correct predictions were made close to the total number of input samples. For instance, the expense of misdiagnosing a rare but fatal condition far outweighs the cost of sending a healthy person for additional tests.

Area Under the Receiver Operating Characteristics (AUROC) as evaluation metrics for checking the classification model's performance tells how much the model can distinguish between classes. Higher the AUC, the better the model predicts 0s as 0s and 1s as 1s. Generally, it plots True Positive Rate (TPR) against False Positive Rate (FPR). This curve generates two essential metrics: sensitivity and specificity. The other metrics used are Sensitivity/ recall (true positive rate), which corresponds to the proportion of positive data points that are correctly considered as positive concerning all positive data points, and Specificity (false positive rate) corresponds to the ratio of negative data points that are mistakenly considered as positive, concerning all negative data points, Precision is the number of correct positive results divided by the number of positive results predicted by the classifier, and F1 score is the harmonic mean between precision and recall. It measures the accuracy of tests and directly indicates the model's performance. The range of the F1 score is between 0 to 1, with the goal being to get as close as possible to 1.

2.3.6 Model Validation

When a trained model is assessed using a testing data set, this process is referred to as model validation in machine learning (Wang & Zheng, 2013). Separate from the data used for training, the testing dataset is used to evaluate the model's performance. The main reason for utilising the testing dataset is to evaluate the trained model's generalisation capacity. As shown in Fig. 5, after training a model, it must be validated to determine which one provides the best results. Validation of a model ensures the accuracy of its results by comparing them mathematically and logically with the actual output. There are two primary methods for achieving model validation: (a) in-sample validation, in which validation is conducted on data from the same dataset used to construct the model, and (b) out-of-sample validation, in which validation is conducted on data from a new dataset that was not used to construct the model (Gill, 2022).

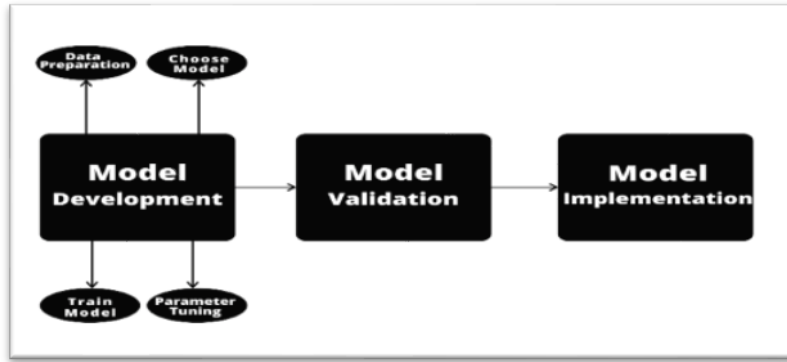


Figure 5: Machine learning model validation process (Datatron, 2022)

2.4 Empirical Literature Review

2.4.1 Malaria Diagnosis and Presumptive Treatment of Malaria

Malaria shares similar symptoms with other febrile diseases such as dengue fever, typhoid fever, common cold, respiratory tract infection, dyspepsia, and pneumonia (Abba *et al.*, 2011; Crump *et al.*, 2017; Nadjm *et al.*, 2010). Parasitological tests, like microscopic and rapid diagnostic tests (RDT), are the recommended and standard tools for diagnosing malaria (WHO, 2018, 2019, 2020). On the other hand, in regions where parasitological testing for malaria is not easily accessible, the difficulty of malaria diagnosis may lead to incorrect diagnoses, overdiagnoses, and unsuitable presumptive treatments (Gosling *et al.*, 2008; Graz *et al.*, 2011; Isiguzo *et al.*, 2014; Maro *et al.*, 2011; UM, 2016). As specified by WHO, in situations such as rural areas where there is no parasitological test available within 2 hours of presenting for treatment in medical centres, medical doctors can provide a prognosis using a clinical examination and physical examination to treat suspected patients (WHO, 2019; WHO-Guidelines, 2015). Consequently, suspected patients would be presumptively treated. Among medical professionals, a clinical diagnosis of malaria is customary. This technique is the most common because it is also the most affordable. Presumptive treatment is a clinical diagnosis based on physical examination findings and the patient's signs and symptoms. Early signs of malaria include fever, headache, bodily weakness, chills, dizziness, abdominal discomfort, diarrhoea, nausea, vomiting, anorexia, and pruritus. These early signs are also highly nonspecific. Due to inadequate awareness of significant malaria symptoms (other than shivering, fever, and sweating) and non-malaria-related variables, the clinical diagnosis of malaria is susceptible to misdiagnosis (Bria *et al.*, 2021). In addition, presumptive treatment could increase the use of unnecessary anti-malarial drugs, which have side effects and increase the spread of drug resistance (Budimu *et al.*, 2020; Gosling *et al.*, 2008; Hertz *et al.*, 2019).

2.4.2 Self-Medication

Apart from that, there is a significant tendency for self-treatment/medication with over-the-counter medication when malaria-related symptoms are observed. Based on the studies done in Tanzania, it was observed that drug-dispensing shops still sell non-prescription drugs frequently, although it is advised that the anti-malarial medications should be administered after a parasitological confirmation of the disease dispense prescription-only treatments (Michael & Mkunde, 2017; Ndomondo-Sigonda *et al.*, 2005). This could lead to disease mismanagement, drug resistance, and drug shortage (Grobusch & Schlagenhauf, 2019; Mboera *et al.*, 2007; Metta *et al.*, 2014; Mwai *et al.*, 2009; Wang *et al.*, 2019). General public awareness on the impact of self-medication and improvement of health services provision is one of the ways that self-medication can be eradicated in the society. With the emergence of technology such as mobile phones and artificial intelligence, development of tools that can assist in disease management in the manner that suits the general population can be a more feasible way to eradicate self-medication. In the efforts to eliminate these issues, the government of Tanzania has established a “not every fever is Malaria” campaign, which aims to educate people that not every fever episode experienced is a malaria case (Baltzell *et al.*, 2019) since there are other diseases such as typhoid, dengue, chikungunya, and urinary tract infections that present the same symptoms as malaria (Goodyer, 2015). The significance of these issues was a substantial drive to develop a malaria prediction model using patients’ symptoms and demographic information. In addition, machine learning techniques have been used as tools for predicting the risk of diseases such as heart disease, diabetes, brain stroke, liver, thyroids disease, and brain cancer (Dwyer *et al.*, 2018; Habib *et al.*, 2018; Kim *et al.*, 2021; Kim *et al.*, 2019; Mishra *et al.*, 2019; SirSat *et al.*, 2020; Priyadarshini *et al.*, 2014; Rao & Renuka, 2020).

2.4.3 Machine Learning for Disease Diagnosis

This section shows studies on how machine learning has been employed in different disease management. Machine learning has been used to detect whether a person is suffering from cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. The study by Garg *et al.* (2021) and Karthick *et al.* (2022) used ML classifiers using patients' symptoms and features to detect the risks of heart disease among patients. In these studies, RF and KNN attained good classification accuracy. Most of these studies focused on using machine learning to identify risk factors and prediction of early signs of the disease (Bhatt *et al.*, 2023; Chang *et al.*, 2022; Karthick *et al.*, 2022;

Nagavelli *et al.*, 2022; Nandal *et al.*, 2022). The studies by Ganiger and Rajashekharaiiah (2018), Kasturiwale *et al.* (2022) and Yuan *et al.* (2022) developed predictive models for diagnosing and forecasting chronic diseases. Apart from that, studies by Imran *et al.* (2019), Imran *et al.* (2019), Nithya *et al.* (2020) and Walse *et al.* (2021) used machine learning to classify kidney patients. The algorithms showed great performance in classifying kidney patients from healthy ones. Cancer is another disease that researchers have used machine learning models to manage the disease. In the study by Ma and Karki, (2020) through machine learning, they were able to classify skin lesions between benign and melanoma using different machine learning techniques. Radhika *et al.* (2019) used preliminary symptoms to predict lung cancer patients. Machine learning techniques have been used as tools for predicting the risk of diseases such as heart disease, diabetes, brain stroke, liver, thyroids disease, and brain cancer (Dwyer *et al.*, 2018; Habib *et al.*, 2018; Kim *et al.*, 2021; Sirsat *et al.*, 2020; Priyadarshini *et al.*, 2014; Rao & Renuka, 2020). Machine learning techniques in these studies successfully predicted and classified diseases, proving that machine learning can also be used to classify malaria patients.

2.4.4 Machine Learning for Malaria Diagnosis

Malaria, like any other disease, has harnessed the power of machine learning to manage the disease from diagnosis, risk analysis, and disease outbreak prediction. Machine learning has been utilised in malaria diagnosis from diagnostic tools to predict illness using patient symptoms and indicators. Malaria research has been conducted over the last decade in the areas of diagnostic testing, malaria Rapid Diagnostic Test (mRDT), and microscopy, specifically the automation of these techniques (Brown *et al.*, 2020; Dharap & Raimbault, 2020; Ford *et al.*, 2020; Ravalji *et al.*, 2020; Shekalaghe *et al.*, 2013). These studies elicited how machine learning can assist in reading microscopic blood smear images to diagnose malaria and automate the complete blood count. This test screens for infection in the blood. The performance of machine learning in the automation of these tools has improved, and classifier prediction accuracy has shown potential (Fuhad *et al.*, 2020; Lee *et al.*, 2021; Lozano *et al.*, 2020; Masud *et al.*, 2020).

Despite the promising results of these studies unavailability of a microscope and mRDT in some health facilities in constrained areas and the self-medication behaviour of some of the patients (Barber *et al.*, 2017; Bibin *et al.*, 2017; Madhu, 2020; Muthumbi *et al.*, 2019; Rajaraman *et al.*, 2018, 2019) remain the major challenge. On the other hand, several machine-

learning studies have used malaria symptoms, signs, and patient information to diagnose malaria. For example, the study done by Bria *et al.* (2021) used malaria symptoms and non-symptom factors to diagnose malaria. It showed potential good prediction accuracy if the combined significant features were identified. However, this study focused on showing the significance of the features in predicting malaria but failed to develop the model. Furthermore, other studies that used malaria symptoms to diagnose malaria used data mining techniques such as rule-based classification, which is considered weak in classification (Bbosa *et al.*, 2016; Oguntimilehin *et al.*, 2015). In Tanzania, most of the studies have been done in diagnostic testing (RDT and microscopy (Mpapalika & Matowo, 2020; Mwanga *et al.*, 2019). The summary of the studies that used machine learning to diagnose malaria is shown in Table 2.

Table 2: Summary of machine learning models for malaria diagnosis

Model Name	Purpose of the Model	Model Classifiers	Reference
The machine learning model for predicting malaria using clinical information	To predict malaria using parasite case reports	Support vector machine, random forest (RF), multilayered perceptron, AdaBoost, gradient boosting (GB), and CatBoost	Lee <i>et al.</i> (2021)
Predicting malaria epidemics in Burkina Faso with machine learning	To forecast the case rate of malaria cases for mitigation purposes.	Gaussian Process and Random Forest	Harvey <i>et al.</i> (2021)
Machine learning approaches classify clinical malaria outcomes based on haematological parameters	To classify malaria outcomes based on Haematological parameters.	Artificial Neural Networks	Morang'a <i>et al.</i> (2020)
Malaria detection using machine learning	To detect malaria parasites using automated image analysis	Support Vector Machine, Deep Transfer Learning and Conventional Neural Network	Kb <i>et al.</i> (2021)
Automated detection of malaria parasite Using Deep Learning Algorithms using microscopic images	To use a deep learning algorithm to learn microscopic images	Artificial Neural Network (ANN),	Fuhad <i>et al.</i> (2020), Kumar <i>et al.</i> (2021), Masud <i>et al.</i> (2020), Pan <i>et al.</i> (2018), Sriporn <i>et al.</i> (2020), Yang <i>et al.</i> (2019),
Diagnosis of malaria using patients' symptoms and signs	To use malaria symptoms and signs to diagnose malaria	Support Vector Machine, K- Nearest Neighbour, Decision Tree, Random Forest and Naïve Bayes	Bbosa <i>et al.</i> (2016), Bria <i>et al.</i> (2021), Chandramohan <i>et al.</i> (2002), Mariki <i>et al.</i> (2022) and Sapkota <i>et al.</i> (2010)

2.5 Research Gap

From the review showed that using machine learning to classify malaria patients using clinical symptoms and non-symptomatic features is a feasible approach since a similar approach has been successfully applied to other diseases. Even though there is a potential in using machine learning to predict malaria, there are some research gaps that this study is going to cover. Based on the review of previous studies on the use of machine learning for malaria diagnosis using patients' symptoms and non-symptomatic features, it has been identified that using only patients' symptoms cannot successfully classify malaria patients. Combining the patient's symptoms and non-symptomatic features is essential for accurately predicting malaria. Many studies also focused on automating malaria diagnostic tools such as microscopes. Still, the challenge is that some of these health facilities don't have this equipment and self-medication behaviour is common among patients. Also, using one machine learning classifier does not give satisfactory predictive power. Using combined machine learning algorithms will give the proposed model better predictive accuracy. The malaria diagnosis dataset used in this study is a unique dataset that covers the whole treatment procedure of the patient which makes it a good suit for clinical prediction of malaria. Therefore, there is a need to develop a machine-learning model using patients' symptoms and demographic features for malaria diagnosis in resource-poor countries like Tanzania. This study aimed to fill this vital gap in malaria research in Tanzania since the country has unavailable diagnostic tools in remote settings, and self-treatment is increasing over time.

2.6 Conceptual Framework of the Study

Figure 6 summarises all the components of the study's conceptual framework. The study started with data collected from malaria patients' files using a designed form. Then pre-processing of data, including data cleaning and transformation of variables, was performed. The modelling process started by screening features to be used in model development. The Important Features obtained were used in the next step of model development, where the performance of all the selected classifiers was evaluated. Finally, the classifiers used during model development were tested with a new dataset.

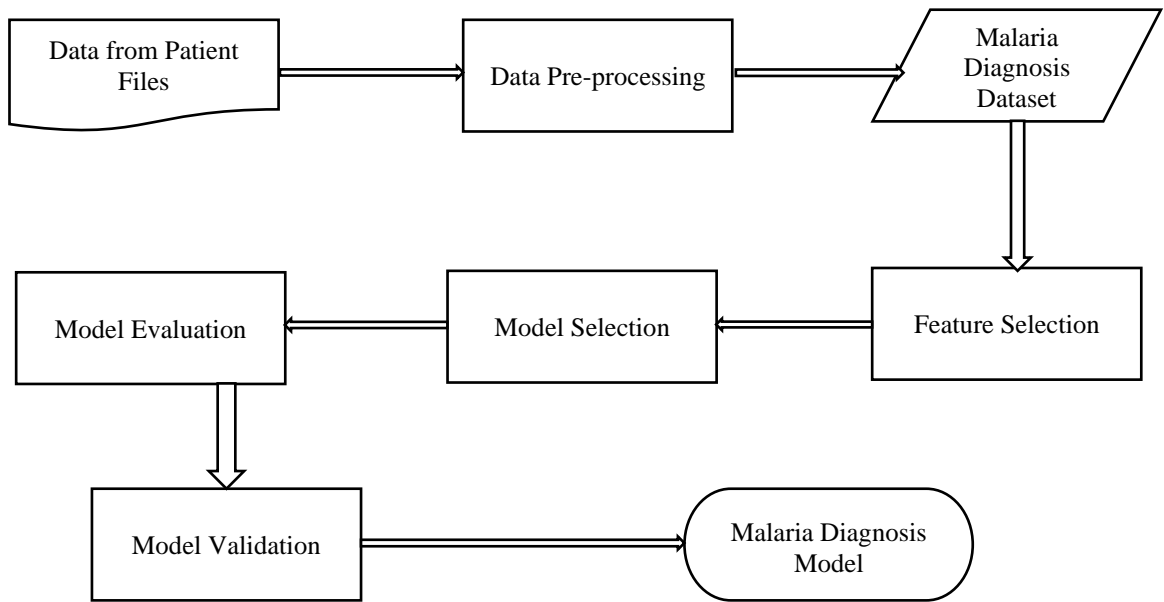


Figure 6: Conceptual framework for the study

CHAPTER THREE

CHARACTERISATION OF MALARIA DIAGNOSIS DATA IN HIGH AND LOW ENDEMIC AREAS OF TANZANIA

3.1 Abstract

Malaria remains a significant cause of morbidity and mortality, especially in the Sub-Saharan Region. Malaria is considered preventable and treatable; however, malaria has increased outpatient visits, hospitalisation, and death worldwide in recent years, with about 9% prevalence in Tanzania. With the massive number of patient records in the health facilities, this study aims to understand the key characteristics and trends of malaria diagnostics symptoms, testing and treatment data in Tanzania's high and low endemic regions.

This is a retrospective study with two phases designed. The primary data were collected from four facilities in two regions, i.e., Morogoro (high endemic) and Kilimanjaro (low endemic), Tanzania. Firstly, malaria patient records were extracted from malaria patients' files from 2015 to 2018. Data collected include: (a) the patient's demographic information, (b) the symptoms presented by the patient when consulting a doctor, (c) the tests taken and results, (d) diagnosis based on the laboratory results and (e) the treatment provided. Apart from that, we surveyed patients who visited the health facility with malaria-related symptoms to collect extra information such as travel history and the use of malaria control initiatives such as insecticide-treated nets. A descriptive analysis was generated to identify the frequency of responses. Correlation analysis Random effects logistic regression was performed to determine the association between malaria-related symptoms and malaria positivity. Significant differences of $p < 0.05$ (i.e., a confidence interval of 95%.) were accepted.

Of the 2556 records collected, 1527 (60%) were from the high endemic area, while 1029 (40%) were from the low endemic area. The most observed symptoms: for high endemic facilities were fever followed by headache, vomiting and body pain; for common endemic facilities, were high fever, sweating, fatigue and headache. A male with malaria symptoms had a higher chance of being diagnosed with malaria than a female. Most patients with fever had a high probability of being diagnosed with malaria. From the interview, 68% of the patients observed with malaria-related symptoms treated themselves without proper diagnosis. The malaria diagnosis data indicate that proper malaria diagnosis is a significant concern. The majority still self-medicate with anti-malaria drugs once they experience any malaria-related symptoms.

Therefore, future studies should explore this challenge and investigate the potentiality of using malaria diagnosis records to diagnose the disease.

3.2 Introduction

Globally, according to WHO's malaria report 2021, there is an estimated increase in malaria cases from 227 million in 2019 to 241 million in 2020, with most of this increase coming from countries in the WHO African Region (Chacko, 2021; WHO, 2022). In 2020, malaria deaths were reported to increase by 12% compared with 2019, to an estimated 627 000, end from 409 000 (Bremam, 2022; WHO, 2022). Tanzania said more than six million confirmed malaria cases in 2019 (Faria, 2022; WHO, 2022). The disease is one of the leading health issues in the country (Mlacha *et al.*, 2020). According to the source estimates, Tanzania accounted for three per cent of the global malaria cases that year (WHO, 2020).

Moreover, there were more than 2500 malaria deaths in 2021 compared to 1171 deaths in 2019 (WHO, 2022). Malaria is considered preventable and treatable. The global priority is to reduce the burden of disease and death while retaining the long-term vision of malaria eradication (Dhiman, 2019; Hemingway *et al.*, 2016; Patouillard *et al.*, 2017; Shretta *et al.*, 2017). Nevertheless, the number of malaria cases worldwide seems to be increasing due to increasing transmission risk in areas where malaria control has declined, the increasing prevalence of drug-resistant strains of parasites, and in relatively few cases, massive increases in international travel and migration (Tangpukdee *et al.*, 2009; WHO, 2015). In Tanzania malaria burden is still unacceptably high; with an overall prevalence of around 9% in mainland Tanzania (Aikambe & Mnyone, 2020). Self-medication has been described as a significant hindrance to proper disease management in many developing countries (Sissinto *et al.*, 2019; Chipwaza *et al.*, 2014; Sigonda *et al.*, 2005; Nsagha *et al.*, 2011). Recently, in Tanzania, the “not every fever is Malaria” campaign aims to educate people that not every fever episode experienced is a malaria case (Baltzell *et al.*, 2019). Other diseases such as typhoid, dengue, chikungunya, and urinary tract infections present the same symptoms as malaria (§Blanco *et al.*, 2021; Capeding *et al.*, 2013; Crump *et al.*, 2013; D’Acremont *et al.*, 2010, 2014; de Santis *et al.*, 2017; Goodyer, 2015). Therefore, proper management of malaria requires prompt and accurate diagnosis and treatment of the disease (Landier *et al.*, 2016).

Understanding the critical characteristics of malaria symptoms, testing and treatment are essential to controlling a disease that continues to pose a significant risk of morbidity and

mortality in the country, with evidence of a resurgence of the disease in recent years (Bali *et al.*, 2011; Krumholz *et al.*, 2006). Understanding the malaria diagnostic process will be essential to inform future case management strategies and guide programmes to improve adherence to national guidelines. Medical records track disease management history and offer information on diagnoses, lab test results, and treatment (Bali *et al.*, 2011; Gallay *et al.*, 2018; Graber *et al.*, 2017). In addition, medical records help us measure and analyse trends in healthcare use, patient characteristics, and quality of care (Graber *et al.*, 2017). Understanding malaria cases' elements are critical for evaluating the disease state. Therefore, this study aims to investigate the features of malaria diagnosis records and explore different variables that can influence malaria diagnosis.

3.3 Material and Methods

This is a mixed study with a retrospective chart review and survey methods. The first phase included retrieving malaria patient records from the health facilities to curate the malaria diagnosis dataset. The second phase engaged a semi-structured questionnaire, whereby questions were conducted to collect the relevant data showing the current malaria diagnosis process and records for the proposed malaria diagnosis model training and validation.

3.3.1 Study Design

This quantitative research used retrospective chart review methods which reviewed pre-recorded malaria patients' records and Patients Surveys to gain insight into malaria diagnosis and treatment practices among patients using a semi-structured questionnaire.

(i) Inclusion Criteria

For the retrospective chart review, only records of malaria cases diagnosed with either microscopy or mRDT and reported patients' symptoms, and the type of treatment given were included in the study. All the positive and negative diagnosis records of patients over five years old were included in the study. For the survey, only patients over five years who have visited the health facility with malaria-related symptoms were included in the study.

(ii) Exclusion Criteria

Any record that did not have complete treatment data was excluded from the study. Patients below five years were excluded since they could not explain their symptoms when they are sick.

3.3.2 Ethical Clearance

The study was approved by the National Institute for Medical Research (NIMR/HQ/R.8.c/Vol.I/1352) before the malaria patients' records were collected. Participants were recruited for the survey, permission to conduct the research was sought and granted by the medical officers in charge at the Regional, District, and health facility levels. Informed consent was obtained from all the patients (or accompanying parents/guardians of minors) who willingly signed the consent form after they were provided with information about the study's objectives. In addition, children over seven years verbally assented to that purpose. The study was of no greater than minimal risk and had no direct impact on patients' rights, welfare, or clinical care. Measures implemented to minimise the risk of confidentiality breaches during the study include anonymising data records and keeping data secured and accessible only to authorised persons.

3.3.3 Study Area and Scope of the Study

The present study was undertaken in two regions in Tanzania, Morogoro and Kilimanjaro, as illustrated in Fig. 7. Morogoro region is one of the regions in Tanzania with a high prevalence of Malaria. The region is situated in the coastal zone of Tanzania (6°49'S and 37°40'E) with a population of approximately 2.3 million at an average altitude of 522 m above mean sea level. The study site on the lower slopes of Uluguru Mountains experiences heavy rainfall from February to June with a total average annual precipitation of 783.5 mm, mean relative humidity of 72 %, minimum temperature of 22 °C, and maximum temperature of 33 °C during wet seasons (Nzobo *et al.*, 2015). Kilimanjaro is amongst the regions with a low malaria prevalence alongside Arusha in Tanzania. The region is located in the northern zone of Tanzania with a population of approximately 1.6 million with an altitude range of roughly 600–1800 m, including the significant municipality of Moshi at about 900 m above sea level. The area receives between 900 and 1200 mm of rainfall per year with two rainy seasons, the long rains from March to May and the short rainy season from November to December.

Four health facilities were selected from the two regions, two from each area. In addition, a regional hospital with the highest level of healthcare and a primary health centre was randomly chosen for each site. These health facilities were selected to represent patients of all levels. Mawenzi regional hospital and Majengo health centre in the Kilimanjaro region and Morogoro regional hospital and Mzumbe health centre in the Morogoro region. The choice of these regions was based on the prevalence of malaria, where Morogoro means areas with a high prevalence of (15.0%) and Kilimanjaro represents regions with a low prevalence of (1.0%).

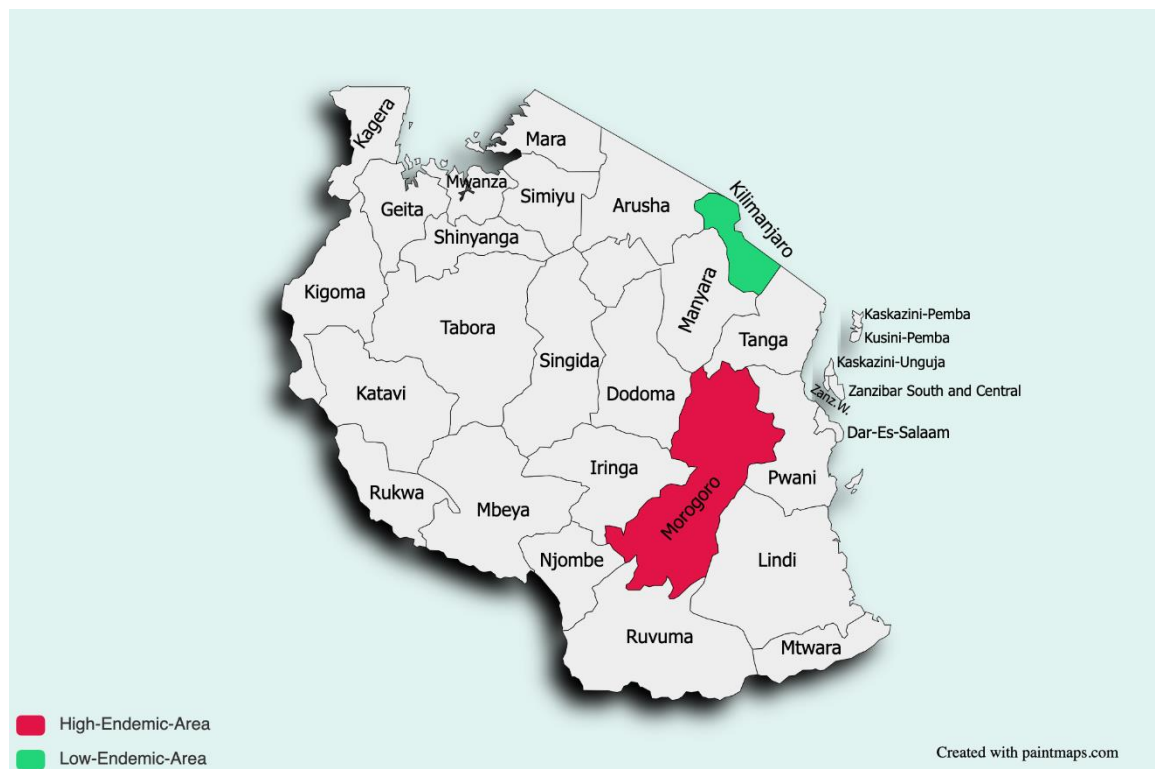


Figure 7: Study area

3.3.4 Study Population

This is a secondary data analysis study of routinely collected malaria data from hospitals and health facilities in the chosen regions. The study population for the retrospective chart review included the patients treated for malaria from 2015 to 2018 in the four selected health facilities. Therefore, only records of malaria case data diagnosed with either microscopy or mRDT and existed at the time of review and approval were accessed for review.

As for the survey, the patients over five years old who visited the health facilities for treatment with malaria-related symptoms were interviewed to gain more insight into malaria treatment

and diagnosis. The patients were selected based on their availability in the health facility for treatments with malaria-related symptoms.

3.3.5 Data Collection

The primary data for this study collected were: (a) malaria patients' records from patients' treatment files and (b) a survey of patients who visited the health facility with malaria-related symptoms. Two data collection tools were developed to collect data from the two groups. Firstly, the patient's records extraction form, as shown in Appendix 2, was designed based on the summary of the Ministry of Health (MoH) patient's file and the information collected when the patient visits the selected health facilities. The records were retrieved from the patients who had been treated for malaria from the year 2015 to 2018. The aim was to identify the past state of clinical malaria diagnosis in the local health facilities (Mawenzi regional hospital and Majengo health centre in Kilimanjaro and Morogoro regional hospital and Mzumbe health centre) Morogoro) and understand the standard practice in the procedure of malaria diagnosis and treatment. Data collected from the patient's files were: (a) the patient's demographic information, (b) the symptoms presented by the patient when consulting a doctor, (c) the tests taken and results, (d) diagnosis based on the laboratory results and (e) the treatment provided. Two trained nurses administered data collection in each health facility, and all participants provided written informed consent.

Secondly, the semi-structured questionnaire shown in Appendix 3 was administered to patients with malaria-related symptoms found in the health facility through an interview. The survey aimed to supplement information on the malaria patients' characteristics not captured in the patients' files, such as the significance of travel history. Also, the survey acted as a validation point of the common symptoms observed by the patients against symptoms recorded in the file.

3.3.6 Data Analysis

The collected malaria diagnosis data were entered in Redcap and obtained into a CSV file analysed in Anaconda (Jupyter Notebook) using Python 3.6. First, the data were coded and cleaned; then, descriptive analysis was generated to identify the frequency of responses to the question items. The investigation was grouped into patients' demographic information and malaria diagnosis procedures. Initial tabulations and univariate analysis examined the distribution of malaria symptoms, diagnosis and treatment overall and within categories.

We computed the association between observed malaria-related symptoms from the patient's records against malaria positivity. The aim was to learn the significance of each symptom and patient demographic information on malaria diagnosis. In addition, observe the likelihood of being malaria positive in a high or low endemic area. Correlation analysis was performed to determine the association between variables such as the age of the patient, the residence area, and age and travel history and signify the degree to which changes in the importance of a dependent variable (Y) increase or decrease in parallel with changes in the values of an independent variable (X). Random effects logistic regression assesses the adjusted impact of covariates on malaria-related symptoms and positivity and adjusts for correlation within hospitals. Significant differences between the dependent and independent variables were accepted at $p < 0.05$, i.e., a confidence interval of 95%. A simple linear regression model was used to determine how the number of malaria cases varied with years, season, age and sex.

3.4 Results

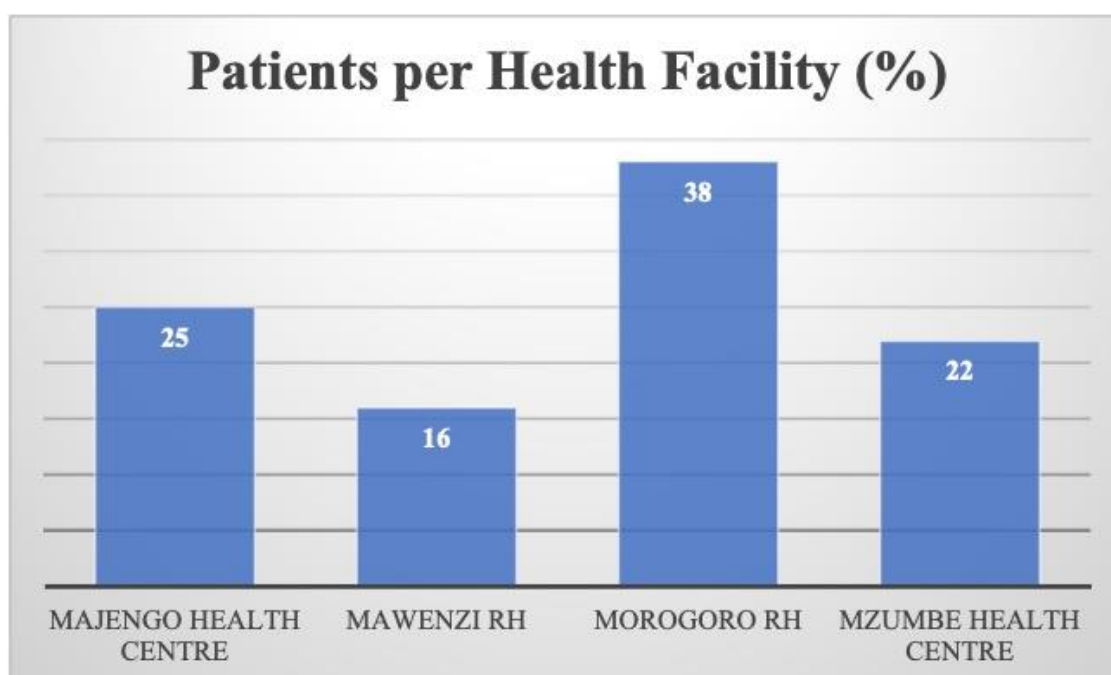
3.4.1 Document Review results

The documentary analysis method was used to identify, select, interpret, and synthesise information in the files of patients who suffered from malaria or presented with malaria-related symptoms. The documentary analysis identified 2556 patient records, of which 60% were from the Morogoro Region, and 40% were from the Kilimanjaro Region. The results also indicated that 61% and 39% of the selected records were female and male, respectively. These patients were of different age distributions, whereby 49.22% were aged between 5 to 24 years, 32.98% were between 25 to 44 years, and 17.78% were aged 45 years and above, as shown in Table 3.

Table 3: Reviewed malaria patients records preliminary information

S/N	Category	Frequency	Percentage (%)	
Malaria	Positive	Morogoro	495	69
		Kilimanjaro	227	31
Diagnosis	Negative	Morogoro	1024	56
		Kilimanjaro	802	44
Sex	Female	1561	61	
	Male	995	39	
Patient's Age	05-14	641	25	
	15-24	742	29	
	25-34	420	16	
	35-44	320	12	
	45-54	220	0.09	
	55-64	130	0.05	
	65+	93	0.04	

Apart from that, 69% of the patients diagnosed with malaria are from Morogoro, while 31% are from Kilimanjaro. While the month of April, May and August showed highest rate on hospital visits as depicted in Fig. 8 and Fig. 9, fever and headache are the most observed symptom, as shown in Fig. 10.

**Figure 8: Number of records per health facility**

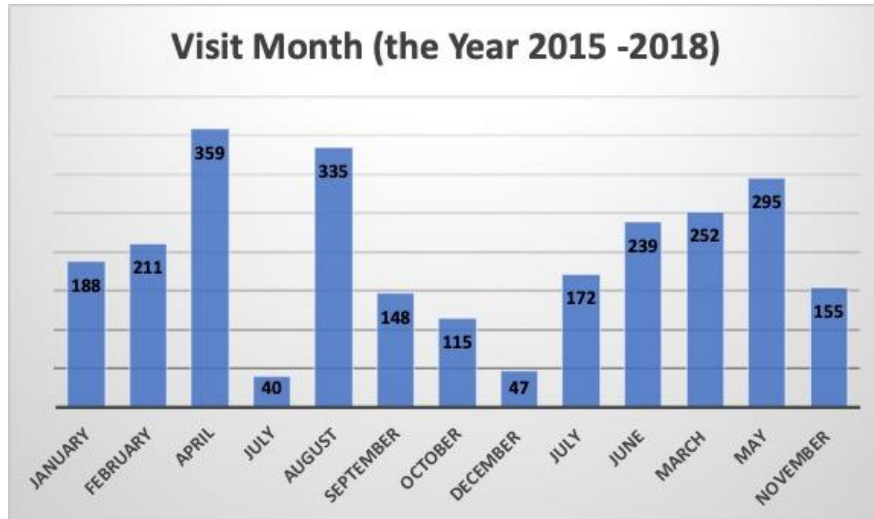


Figure 9: Number of patients per month visiting the health facility

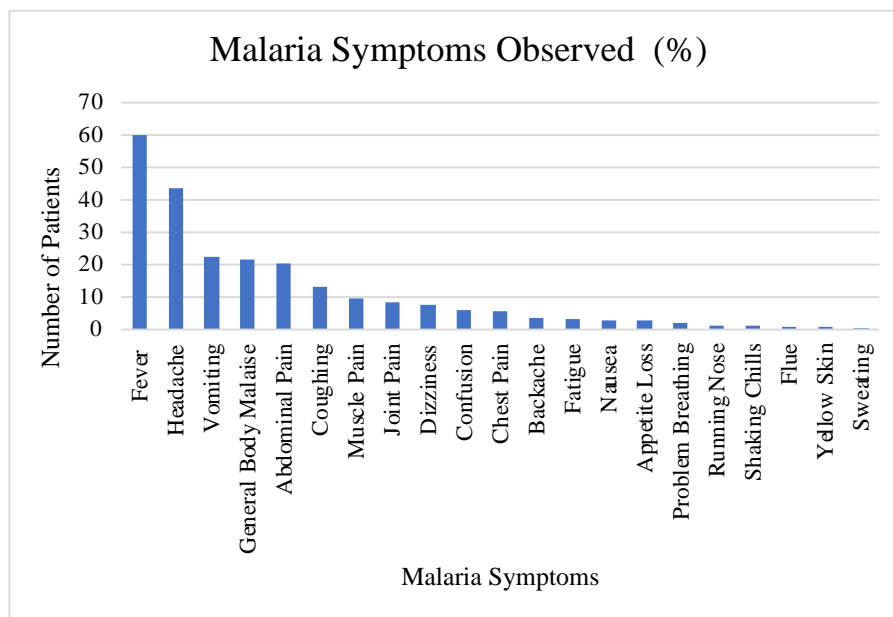


Figure 10: Frequency of malaria symptoms observed

(i) Significant malaria symptoms

This section computed the association using relative risk and odds ratio between observed malaria-related symptoms from the patient's records against malaria positivity. Relative risk is a ratio of the probability of an event occurring in the exposed group versus the probability of the event occurring in the non-exposed group. The association computation aimed to learn the significance of each sign and patient demographic information on malaria diagnosis. The aim is to see which symptoms, when observed, the likelihood of being malaria positive is high or low. The relative risk and odds ratio of different malaria-related symptoms and non-symptomatic features are discussed as shown in Table 4.

High Fever from 40oC

It was found that the magnitude of malaria among those with a high fever from 40°C is equal to 70.8%, and the extent of malaria among those without a high fever from 40°C is equal to 28.2%. The difference in the two proportions is statistically significant with ($p = 0.002$). Patient with a high fever of 40°C and above has a 40% risk of having malaria, while those without a high fever of 40°C have a 60% protection.

Abdominal Pain

The magnitude of malaria among those with Abdominal Pain is 22.2%, and the extent of malaria among those without Abdominal Pain is 77.7%. The difference in the two proportions is statistically significant with ($p = 0.046$). Patient without abdominal pain has twice the risk of about 180% compared to those with abdominal pain.

Vomiting

The patients with vomiting symptoms have a 100% risk of having malaria, while those who have not demonstrated vomiting symptoms have a 40% risk of having malaria. The magnitude of malaria among those with vomiting symptoms is 43.0%, and the extent of malaria among those without Abdominal Pain is 56.9%. The difference in the two proportions is statistically significant, with a p-value of 0.001.

Joint Pain

Patients with joint pain symptoms have a 100% risk of having malaria, while those who haven't shown any sign of joint pain have a 48% risk of having malaria. The difference in the two proportions is statistically significant, with a p-value of 0.049. The magnitude of malaria among those with joint pain is equal to 16.6%, and the extent of malaria among those without joint pain is equal to 83.3%

General Body Malaise

The analysis also observed that patients who have not observed body malaise have a 50% risk of having malaria compared to those presenting the sign of body malaise. The difference in the two proportions has shown statistical significance with a p-value of 0.015. The magnitude of

malaria among those with general body malaise is 36.1%, and the extent of malaria among those without general body malaise is 63.8%.

Location and Malaria Positivity - Health Facilities

In Table 3, 86% of the patients diagnosed with malaria are from Morogoro, while 13% are from Kilimanjaro. Of all patients with malaria positivity, 8% are from Mawenzi regional hospital and have a 50% risk of malaria. Also, the analysis shows that 31% of the patients diagnosed with malaria are from Morogoro regional hospital and have 3.7 times the chance of malaria. The differences in the two relationships have statistical significance. As for Mzumbe Health Centre, 54% of the patients diagnosed with malaria are from this health facility, and there is three times the chance of having malaria when from this facility.

Sex/Gender and Age

Male patients have twice the chance of malaria than female patients. Also, the research shows that age has no statistical significance in malaria positivity. However, general observation after the odds ratio analysis was done on a combination of different variables against malaria positivity is that patients that are from the facilities in Morogoro, male, with ages between 25-44 years and those who come with high fever, headache, abdominal pain, joint pain, body malaise, vomiting as symptoms have a statistical significance.

(ii) Variables predicting malaria positivity

General observation after the odds ratio analysis was done on different variables against malaria positivity is that patients that are from the facilities in Morogoro region, male, with age between 25-44 years and those who come with high fever, headache, abdominal pain, joint pain, body malaise, vomiting as symptoms have a statistical significance in malaria positivity as shown in Table 5. When the multivariate analysis was performed on significant variables, the results showed that the residence area of the patient and some of the symptoms remained highly correlated as seen in Table 6. The outcome variable “malaria positivity” was caused by many input factors. These factors include the area the patient is coming from, the age of the patient, the sex of the patient and the symptoms presented. All factors indicated positive correlations with malaria positivity, but the area of residence and symptoms (High Fever, Nausea, Joint pain and Body Malaise) had the strongest correlation compared with the other

factors. These values indicate that kin observation of the symptoms by the patient is very important in malaria diagnosis and in raising awareness in the community.

Table 4: Malaria symptoms observed with malaria positivity in document review

Symptoms Observed	Checked with Malaria	Checked with No Malaria	Unchecked with Malaria	Unchecked with No Malaria	P-Value for the symptom
High fever (≥ 40 °C)	51(70.8%)	149(50.3%)	21(29.2%)	147(49.6%)	0.002
Shaking chills	1(1.39%)	0(0%)	71(98.6%)	296(100%)	0.042
Profuse sweating	0(0%)	1(0.34%)	72(100%)	296(100%)	0.621
Fatigue	1(1.39%)	6(2.03%)	71(98.6%)	290(97.9%)	0.722
Headache	48(66.6%)	195(65.5)	24(33.3%)	101(34.1%)	0.899
Muscle aches/pain	2(2.7%)	8(2.7%)	70(97.2%)	288(97.3%)	0.972
Abdominal discomfort	16(22.2%)	102(34.4%)	56(77.7%)	194(65.5%)	0.046
Vomiting	31(43.0%)	69(23.3%)	41(56.9%)	227(76.6%)	0.001
Dizziness	7(9.7%)	33(11.5%)	65(90.2%)	263(88.8%)	0.727
Problem breathing	0(0%)	5(1.6%)	72(100%)	291(98.3%)	0.267
Seizure	0(0%)	1(0.3%)	72(100%)	295(99.6%)	0.621
Nausea	3(4.1%)	8(2.7%)	69(95.8%)	288(97.3%)	0.513
Joint Pain	12(16.6%)	26(8.7%)	60(83.3%)	270(91.2%)	0.049
General Body Malaise	26(36.1%)	66(22.3%)	46(63.8%)	230(77.7%)	0.015
Chest Pain	2(2.7%)	34(11.4%)	70(97.2%)	262(88.5%)	0.026
Coughing	7(9.7%)	40(13.5%)	65(90.2%)	256(86.4%)	0.387
Backache	2(2.7%)	43(14.5%)	70(97.2%)	253(85.4%)	0.006
Loss of consciousness	0(0%)	2(0.68%)	72(100%)	294(99.3%)	0.484

Table 5: Multivariate analysis of significant factors to malaria positivity results (a)

Malaria diagnosis factors	Odds Ratio	Std. Err.	z	P> z	Interval
Majengo Health Facility	13.6054	6.911744	5.14	0.000	5.026762 - 36.82428
Mzumbe Health Facility	7.641262	3.386888	4.59	0.000	3.205393 - 18.21582
Sex_(M)	1.065771	.3500778	0.19	0.846	.5598429 - 2.028903
Age					
25-44	.9478376	.3332518	-0.15	0.879	.4758374 - 1.888032
45+	.7296848	.3896629	-0.59	0.555	.2562008 - 2.078213
Symptoms Observed					
High Fever	.2761818	.0957055	-3.71	0.000	.1400321 - .5447065
Abdominal discomfort	1.646044	.6146237	1.33	0.182	.7917858 - 3.421964
Nausea	.5845159	.1916728	-1.64	0.102	.307378 - 1.111527
Joint Pain	.2416235	.1119603	-3.07	0.002	.0974363 .5991805
Body Malaise	.5119071	.1828989	-1.87	0.061	.2541358 1.031137
Chest Pain	2.280418	1.788837	1.05	0.293	.4901211 10.61025
Back pain	1.872374	1.528118	0.77	0.442	.3781757 9.270254
Cons	.1396885	.1867649	-1.47	0.141	.0101647 1.919665

Table 6: Multivariate analysis of significant factors to malaria positivity results (b)

Malaria diagnosis factors	Odds Ratio	Std. Err.	z	P> z	95% Conf. Interval
Health facility					
Majengo Health Facility	17.6626	8.62072	5.88	0.000	6.785819 45.97345
Mzumbe Health Facility	10.49589	4.305658	5.73	0.000	4.697174 23.45319
Symptoms observed					
High Fever	.2440872	.0822909	-4.18	0.000	.1260588 .4726252
Nausea	.5809043	.1855332	-1.70	0.089	.310629 1.086344
Joint pain	.2268551	.1025172	-3.28	0.001	.0935589 .5500621
Body Malaise	.4657251	.1559291	-2.28	0.022	.2416237 .8976764
Cons	.728074	.4287537	-0.54	0.590	.2295691 2.309072

3.4.2 Malaria Patients Survey

The overall observation from the patient survey was that of the 312 malaria patients questioned, 44.24% were from the Kilimanjaro region, and 55.76% were from the Morogoro region.

Among the 312 respondents, 65.58% were female, 34.42% were male, and 54.54% were between 15 -and 35 years. The results also indicated that 48.22% of the respondents have only primary school education, 33.65% have a secondary school education, 16.8% have a college education, and only 1.29% are uneducated, as shown in Table 7.

Table 7: Survey respondents' demographics information

Category	Frequency (N)	Percentage (%)
Residence area		
Morogoro	173	55.8
Kilimanjaro	138	44.2
Patients' education level		
Primary School Education	150	48.2
Secondary School Education	105	33.7
College Education	53	16.8
None	4	1.3
Patients sex		
Female	204	65.6
Male	108	34.4

(i) Malaria symptoms identified by the survey respondents

Symptoms Observed from the malaria patients survey of 312 participants found that headache (67.3%), high fever (up to 40°C) (43.9%), fatigue (feeling tired) (35.2%), muscle aches/pain (28.8%) and abdominal discomfort (14.42%) and nausea (14.42%) were highly observed symptoms in both the regions. Other symptoms are indicated as seen in Table 8.

(ii) Malaria diagnosis and treatment history

The survey results in Table 9 revealed that 61.5% were formally diagnosed with malaria in the period of three months of 2018, and among that, Kilimanjaro (54.5%) and Morogoro (45.5%), while 38.5% were not diagnosed with malaria. Amongst the 38.5% who were not diagnosed with malaria in Kilimanjaro, 31% and Morogoro value 69%. Also, the analysis showed that among the 38.5% of patients not diagnosed with malaria, 66.7% observed malaria symptoms, and 56% mainly from Morogoro self-medicated with antimalaria drugs. In addition, 40% of the patients diagnosed with malaria have a travelling history to the high endemic areas in the past three months. These results are shown tabled in Table 9.

Table 8: Malaria symptoms identified by the survey respondents

Symptoms observed	Patients Survey (N=312)	
	Frequency	Percentage
High fever (from 40 °C)	137	43.9%
Shaking chills	23	7.4%
Profuse sweating	8	2.6%
Fatigue	110	35.2%
Headache	210	67.3%
Muscle aches/pain	90	28.8%
Abdominal discomfort	45	14.4%
Nausea	42	14.4%
Vomiting	33	10.6%
Dizziness	36	11.5%
Delirium and confusion.	1	0.32%
Problem breathing	1	0.32%
Severe anaemia	2	0.6%
Seizure	1	0.32%

(iii) The use of malaria control initiatives

As illustrated in Table 9, most respondents (88%) used Treated Nets, followed by Insecticides Spray 6.57%. Malaria control initiatives were introduced by the WHO and administered by countries to control malaria cases (Finda *et al.*, 2020; Matowo *et al.*, 2017; Russell *et al.*, 2011). Malaria vaccination showed an inferior adaptation with only 0.64%. Few respondents (5.12%) do not use any malaria control initiative. The reasons were that the area has few or no mosquitoes and the current insecticide-treated nets are worn out.

Table 9: Malaria diagnosis and treatment history

S/N	Questions	Feedback	n (%)
1.	Being Diagnosed with Malaria in the past three months (N=312)	Yes	192 (61.5%)
		No	120(38.5%)
2.	Observed Malaria related symptoms in the past three months (N=120)	Yes	80(66.6%)
		No	40(33.4%)
3.	The number of times you have been diagnosed with malaria or observed malaria-related symptoms in the past three months N=192	Once (One time)	60(31.3%)
		More than once	132 (68.7%)
4.	Did you get any treatment for such self-observation of malaria-related symptoms?	Yes	186 (68.3%)
		No	86(31.7%)
5.	Use of malaria control initiatives	• Treated Nets	275(88%)
		• Insecticides Spray	19 (6.5%)
		• Malaria Vaccination	2 (0.64%)
		• Non-use of Malaria Control Initiative (MCI)	16 (5.12%)
6.	Reason for not using any MCI	• Minimal amount of mosquitos	10 (62%)
		• Tear and wear of the current Net	6 (38%)

3.5 Discussion

This chapter aims to; (a) explore different variables that can influence malaria diagnosis and (b) create a dataset from malaria patients' records that can be used for training and validation of the machine learning model to improve malaria diagnosis in a resource-poor country like Tanzania. Overall, it was found that half of the patients who observed malaria-related symptoms treated themselves with anti-malaria drugs without any proper diagnosis from the health facility. This signifies that self-medication is still a challenge. Similar findings were also observed in the studies done in Kenya, Benin and Ghana, where self-medication is still practised in these counties and Tanzania is no different (Sissinto *et al.*, 2019; Quaresima *et al.*, 2021). Furthermore, we found that patients from high endemic facilities, who are male, and those who come with high fever, headache, abdominal pain, joint pain, general body malaise, and vomiting symptoms have a high chance of being diagnosed with malaria. This finding aligns with the Tanzania malaria diagnosis guideline, where the guideline also identifies the symptoms observed in this study (WHO-Guidelines, 2015). As for the male gender, the 2022 study by Okiring in Uganda also found that males had a higher probability than females of

testing positive for malaria, and this makes the general lifestyle and economic activities of male to be in question (Okiring *et al.*, 2022). Also, the same study observed that those aged between 15 and 39 are at risk of being diagnosed with malaria, as found in this study, where ages between 25 and 44 years are more likely to have malaria than other age ranges.

The findings also revealed that the risk of malaria among males is high due to the high participation rate in social activities at night and some economic activities such as agriculture. Supporting these findings is the study done in East Africa under the Gates Foundation, where it was noted that Men often face the risk of exposure through their occupations, such as fishing, mining, forestry, or agriculture, when these activities are conducted during peak biting times (Katz & Hartley, 2020). Apart from that, it was found that a lack of awareness of the effects of self-medication was described as a significant source of self-medication, as supported by the studies of Bria *et al.* (2021). Apart from self-medication has been described as contributing factor to drug resistance, developing chronic diseases, and even death, sometimes to untreated infections, assuming they have malaria (Mboera *et al.*, 2007). There are several reasons why self-medication is more practised; the study by Ngasala *et al.* (2008) has shown that even though over 80% of Tanzanians live within 5 km of a health facility providing malaria treatment, treatment is often inadequate due to a lack of standard malaria treatment guidelines (Ngasala *et al.*, 2008). Another study by Yeka *et al.* (2012). has shown that financial constraints have caused inappropriate drug usage to seek the full treatment procedure and sometimes inherited behaviour among community members. It was also found that residence area, High Fever, Nausea, Joint pain, and Body Malaise had the strongest correlation with malaria positivity compared with the other symptoms. This indicates that kin observation of both non-symptoms, such as where the patients live and their sex, are significant in observing the patient malaria diagnosis and raising awareness in the community (Bria *et al.*, 2021).

With all that has been observed developing a tool that can give patients the probability of being malaria positive when observing any malaria-related symptoms might be a possible solution to reduce the rate of self-medication (Bria *et al.*, 2021). Prediction models are among those tools that can improve the diagnosis and awareness of the patient's state before buying over-the-counter medication (Deepthi *et al.*, 2020). The model can relate patients' history of the diseases and integrate symptoms and signs presented to physicians (Bria *et al.*, 2021; Deepthi *et al.*, 2020). The limitations of this study are the following: firstly, our study population was based only in two regions which cannot generalise our findings to the entire country. Secondly, this

study only described the dataset without demonstrating the development and implementation of machine learning models in Tanzania. The study's strength is comparing the data from two regions representing the country's higher and low endemic areas. In addition, we analysed both medical history records and recent data obtained through the survey.

3.6 Conclusion

The malaria diagnosis data in this study indicate that proper malaria diagnosis is a significant concern. As the majority still self-medicate with anti-malaria drugs once they experience malaria-related symptoms, future studies should explore this challenge and investigate the potentiality of using malaria diagnosis records to diagnose the disease. Furthermore, although microscopic blood slides and rapid diagnostic tests are widely available, several challenges were identified, including self-medication with anti-malaria drugs and presumptive treatment of malaria. Therefore, it is recommended that better methods of malaria diagnosis should be imposed in society to reduce the effects.

CHAPTER FOUR

DEVELOPMENT OF A MACHINE LEARNING MODEL FOR MALARIA PREDICTION

4.1 Abstract

Presumptive treatment and self-medication for malaria have been used in limited-resource countries. However, these approaches have been considered unreliable due to the unnecessary use of malaria medication. This study aimed to demonstrate supervised machine learning models in diagnosing malaria using patient symptoms and demographic features. The malaria diagnosis dataset was collected in two regions of Tanzania: Morogoro and Kilimanjaro. Regional-based features were obtained to improve model performance and reduce processing time. Machine learning classifiers with the k-fold cross-validation method were used to train and validate the model. The dataset developed a machine-learning model for malaria diagnosis using patient symptoms and demographic features. A malaria diagnosis dataset of 2556 patients' records with 36 features was used. It was observed that the ranking of features differs among regions and when combined dataset. The feature ranking indicated that fever is universally the most noteworthy feature for predicting malaria, followed by general body malaise, vomiting and headache. The features identified comply with malaria diagnosis and treatment guidelines provided by WHO and Tanzania Mainland that indicate that in situations such as rural areas where there is no parasitological test available within 2 hours of presenting for treatment in medical centres, medical doctors can provide a prognosis using a clinical examination and physical examination to treat suspected patients. The compliance is observed to produce a prediction model that will fit in the current healthcare provision system.

Random Forest was the best classifier, with an accuracy of 95% in Kilimanjaro, 87% in Morogoro and 82% in the combined dataset. Based on clinical symptoms and demographic features, a regional-specific malaria predictive model was developed to demonstrate relevant machine-learning classifiers. Important features are useful in making the disease prediction.

4.2 Introduction

Machine learning (ML) is an emerging approach that is effective in making decisions and predictions from the large quantity of data produced by the healthcare industry. It learns from experience and detects valuable patterns from large, unstructured, complex datasets to predict

future incidences. Today, the biggest challenge in front of the healthcare industry is diagnosing diseases with accuracy and at affordable costs. A massive amount of complex data is available with the hospitals that can be used to extract useful information for diagnosis. This data can be used for future predictions with the help of data mining. The healthcare field generates massive data about clinical assessment, patient records, disease treatment, clinical follow-ups, and medication (Fatima & Pasha, 2017; Iyer *et al.*, 2015). This massive data can improve healthcare delivery when incorporated with machine learning techniques. Patient care and illness management improvements may result from more precise clinical outcome prediction. The correct prediction of which patients should be provided with malaria treatment and should have future check-ups, for instance, may reduce the needless administration of malaria medications in managing malaria (Menard & Dondorp, 2017; Mwai *et al.*, 2009). Apart from that, a lack of proper diagnosis might result in the mismanagement of other diseases that have related symptoms to malaria. Common behaviour on self-medication with malaria drugs and challenges in the health system in most low-income countries like Tanzania necessitate a machine learning-based diagnosis model. In addition, the model can assist in correctly diagnosing malaria for patients who cannot get a laboratory-based diagnosis.

The use of ML for malaria diagnosis is not necessarily the right solution. For example, a better solution would be to have rapid malaria diagnostics tests at pharmacies to ensure only malaria patients or those with an anti-malaria prescription are given anti-malarial drugs. However, the rapid tests would be costly for pharmacies and require administration by trained pharmacists or personnel, who may not be available in rural/remote areas. A cheap but effective tool for determining possible malarial status is therefore needed. The ML-based diagnostic tool could be one such tool. Different studies have shown how machine learning assisted other areas of the health care system (Davenport & Kalakota, 2019; Khare *et al.*, 2017; Shailaja *et al.*, 2018; Sidey-Gibbons, 2019; Triantafyllidis & Tsanas, 2019). Recently, supervised learning algorithms have been applied in various studies to diagnose malaria (Fuhad *et al.*, 2020; Madhu, 2020; Masud *et al.*, 2020; Muthumbi *et al.*, 2019; Poostchi *et al.*, 2018; Yang *et al.*, 2019). While machine learning has been successfully used in illness management, most applications ignore that most health institutions do not have a microscope. Patients treat themselves by relying on home tests instead. Machine learning is a reliable and efficient non-invasive way of distinguishing between healthy and malaria-infected individuals. Though previous research has cast doubt on the viability of utilising clinical symptoms in malaria prediction, this study's trials demonstrate that it is possible to use clinical symptoms alongside

patient demographics to predict malaria using machine learning classifiers (Bibin *et al.*, 2017; Das *et al.*, 2013; Femi Aminu *et al.*, 2016; Fuhad *et al.*, 2020; Madhu, 2020; Masud *et al.*, 2020; Patil *et al.*, 2018; Pillay *et al.*, 2019; Rajaraman *et al.*, 2018, 2019; Shekalaghe *et al.*, 2013; Van Driel, 2020).

4.2.1 Related Works

Malaria shares similar symptoms with other febrile diseases such as dengue, typhoid, common cold, respiratory tract infection, dyspepsia, and pneumonia (Abba *et al.*, 2011; Crump *et al.*, 2017; Nadjm *et al.*, 2010). Parasitological tests, like microscopic and rapid diagnostic tests (RDT), are the recommended and standard tools for diagnosing malaria (WHO, 2019, 2020, 2021). However, in areas where parasitological tests for malaria are not readily available, the complexity of malaria diagnosis may lead to misdiagnosis, overdiagnosis, and inappropriate presumptive treatment (Gosling *et al.*, 2008; Graz *et al.*, 2011; Isiguzo *et al.*, 2014; UM, 2016; V D'Acremont, 2009). As specified by WHO, in situations such as rural areas where there is no parasitological test available within two hours of presenting for treatment in medical centres, medical doctors can provide a prognosis using a clinical examination and physical examination to treat suspected patients (WHO, 2019, 2021; WHO-Guidelines, 2015). Consequently, suspected patients would be presumptively treated. Malaria is traditionally diagnosed clinically by doctors. This is the least expensive and most widely used approach. Presumptive treatment is a clinical diagnosis based on the patient's indications and symptoms as well as physical findings at the examination. Malaria's initial symptoms are vague and include fever, headache, bodily weakness, chills, dizziness, abdominal discomfort, diarrhoea, nausea, vomiting, anorexia, and itching. Misdiagnosis is possible with clinical diagnosis of malaria due to a lack of appropriate understanding regarding important malaria symptoms (other than shivering, fever, and sweating) and non-malaria-related variables (Bria *et al.*, 2021). Presumptive treatment could increase the use of unnecessary anti-malarial drugs, which have side effects and increase the spread of resistance to the drugs (Sissinto *et al.*, 2019; Chipwaza *et al.*, 2014; Kajeguka *et al.*, 2017; Hertz *et al.*, 2019; Kazaura, 2017; Mwita *et al.*, 2019).

Machine learning has been utilised in malaria diagnosis, from diagnostic tools to predicting illness presence based on patient symptoms and indicators. Malaria research has been conducted throughout the last decade in the areas of diagnostic testing (RDT) and microscopy, specifically the automation of these tools (Brown *et al.*, 2020; Dharap & Raimbault, 2020; Ford *et al.*, 2020; Ravalji *et al.*, 2020; Shekalaghe *et al.*, 2013). These studies elicited how machine

learning can assist in reading microscopic blood smear images to diagnose malaria and automate the complete blood count, which is the test that screens for infection in the blood. The performance of machine learning in the automation of these tools has improved, and classifier prediction accuracy has shown potential (Fuhad *et al.*, 2020; Lee *et al.*, 2021; Masud *et al.*, 2020; Van Driel, 2020). Despite the promising results of these studies, the unavailability of a microscope and mRDT in some of the health facilities in constrained areas and the self-medication behaviour of some of the remain the major challenge patients (Bibin *et al.*, 2017; Das *et al.*, 2013; Liang *et al.*, 2017; Madhu, 2020; Masud *et al.*, 2020b; Muthumbi *et al.*, 2019; Poostchi *et al.*, 2018; Rajaraman *et al.*, 2018, 2019).

On the other hand, several machine-learning studies have used malaria symptoms, signs, and patient information to diagnose malaria. For example, the study done by Bria *et al.* (2021) used malaria symptoms and non-symptom factors to diagnose malaria. It showed potential good prediction accuracy if the combined significant features were identified. However, these studies do not specifically identify significant symptoms, notwithstanding their contribution to malaria diagnosis improvement. Furthermore, other studies that used malaria symptoms to diagnose malaria used data mining techniques such as rule-based classification, which are considered weak in classification (Bbosa *et al.*, 2016). In Tanzania, most of the studies have been done in malaria diagnosis (Mpapalika & Matowo, 2020; Mwanga & Mapua, 2019; Mwanga & Minja, 2019). A malaria diagnosis study using symptoms and patients' demographic features has never been done in Tanzania. This study aimed to fill this important gap in malaria research in Tanzania since the country has settings where diagnostic tools are unavailable, and self-treatment is over the chart. This study's findings can raise public awareness of the potentiality of using machine learning in classifying malaria patients by developing a simple tool that will be used before administering anti-malaria drugs. The study will raise public awareness of significant malaria symptoms and patient features in diagnosing malaria at early stages within Tanzanian societies vulnerable to malaria and reduce the rate of self-medication and presumptive treatment in the country.

4.2.2 Theoretical Background

This study uses the most common supervised machine learning classifiers to build a malaria diagnosis model. The popular machine learning classifiers for disease diagnosis are Logistics Regression (LR), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision

Tree (DT), and Random Forest (RF), which were used in the model development (Ibarra *et al.*, 2021) as explained in 2.3.4.

4.3 Materials and Methods

This chapter aimed to develop a machine learning-based model to classify patients with and without malaria using their symptoms and non-symptoms factors. The machine learning-based model for malaria diagnosis development was structured in six stages, namely: (a) Dataset description and pre-processing, (b) Features selection, (c) Machine learning classifiers, (d) Cross-Validation methods, (e) Classifier performance evaluation and (e) Development of regional-specific malaria diagnosis model as shown in Fig. 11.

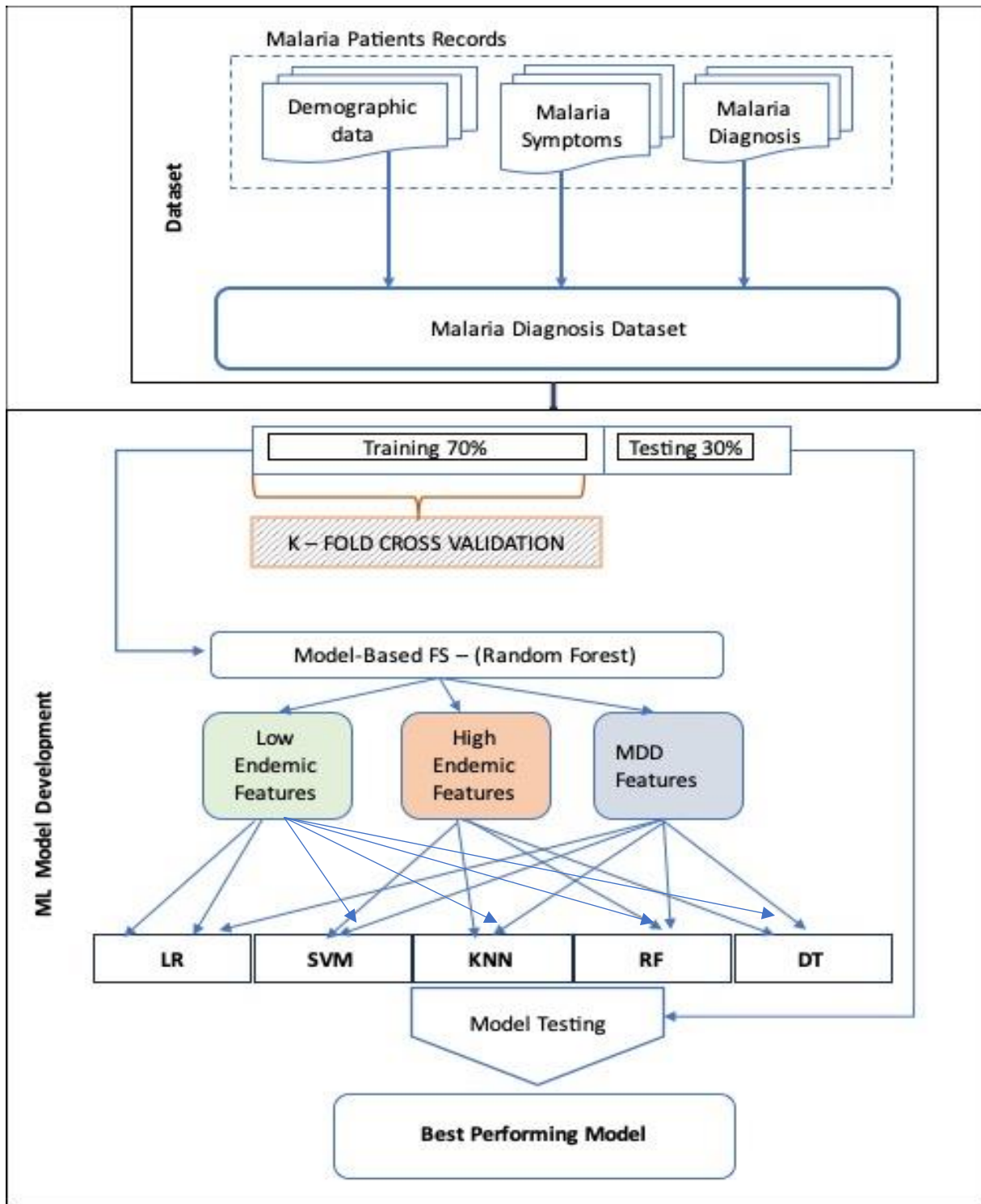


Figure 11: Machine learning framework employed in features selection, model development and validation

4.3.1 Dataset Collection and Description

(i) Study Area

Data were collected from four hospitals in two regions in Tanzania: Morogoro and Kilimanjaro (Fig. 7). The four health facilities are Mawenzi regional hospital, Majengo health centre in the Kilimanjaro region, Morogoro regional hospital, and Mzumbe health centre in the Morogoro region. The dataset represents the patients who live in the areas with low malaria transmission, represented by the Kilimanjaro region and those who live in the areas with high malaria transmission, represented by the Morogoro region. The choice of these regions was based on the prevalence of malaria, where Morogoro represents regions with a high prevalence of (15.0%), and Kilimanjaro represents regions with a low prevalence (1.0%) of malaria.

(ii) The method used and Participants

A malaria patient's records extraction form was designed to summarize the MoH patient's file and the data collected when visiting the health facility. The records were retrieved from the patient's files who have been treated for malaria from 2015 to 2019. The aim was to identify the past state of clinical malaria diagnosis in the local health facilities and understand the standard practice in malaria diagnosis and treatment. The critical information collected was: (a) the patient's demographic information, (b) the symptoms presented by the patient when consulting a doctor, (c) the tests taken and results, (d) diagnosis based on the laboratory results and (e) the treatment provided. Trained nurses administered data collection, and all participants provided written informed consent.

(iii) Ethical clearance

Participants were recruited, and data were gathered after receiving approval from the National Institute for Medical Research in Tanzania (NIMR). Informed written consent was obtained from all individuals before their inclusion in the study. Medical centres followed NIMR's guidance in obtaining patients' permission to share their medical records.

4.3.2 Dataset Descriptions and Pre-processing

Data cleaning, transformation and reduction were performed on the dataset. For data cleaning, missing values were handled by removing the tuples with the missing values. Next, concept hierarchy generalisation (Han *et al.*, 2012; Velliangiri *et al.*, 2019) was used to transform the

patient's residence area variable by grouping the residence area to the hospital the patient attended. Finally, feature selection as a one-dimensionality reduction technique (Masud *et al.*, 2020b) was applied to reduce the number of subset attributes insignificant to the target variable. A target variable whose values are modelled and predicted by other variables. In this case, the target variable is malaria diagnosis, which can either be positive or negative.

The malaria diagnosis dataset was used in this study to develop a machine-learning model for malaria diagnosis. The dataset was obtained by extracting malaria patients' diagnosis records from the Tanzania Ministry of Health's patient files in two regions in Tanzania: Morogoro and Kilimanjaro. The original Malaria diagnosis dataset has a sample size of 2556 patients' records with 36 features, as shown in Table 10. The targeted output variable has two classes representing patients with malaria (tested positive) and those without malaria (tested negative). Instances that could lead to individual patients being located or identified were removed to maintain the confidentiality of the patient and ethical practice. Nominal features were encoded to conform to Scikit-learn and coded 1 for patients with malaria and 0 for patients without malaria (health people). The output of this section was a malaria diagnosis dataset used for malaria diagnosis model development.

Table 10: Malaria diagnosis dataset features description

S/N	Feature Name	Data Type	Description	Domain of Values
1	Residence Area	Categorical	1 = MajengoHC, 2 = MorogoroRH, 3 = MawenziRH, 4 = MzumbeHC	1, 2, 3, 4
2	Visit Date	Categorical	Date in Months	1,2,3,4,5,6,7,8,9,10 ,11,12
3	Age	Categorical	Age in Years	> 5 age <95
4	Gender	Categorical	Male = 1, Female = 0	1, 0
5	Fever	Integer	Yes = 1, No = 0	1, 0
6	Sweating	Integer	Yes = 1, No = 0	1, 0
7	Fatigue	Integer	Yes = 1, No = 0	1, 0
8	Headache	Integer	Yes = 1, No = 0	1, 0
9	Shaking & Chills	Integer	Yes = 1, No = 0	1, 0
10	Muscle Pain	Integer	Yes = 1, No = 0	1, 0
11	Joint Pain	Integer	Yes = 1, No = 0	1, 0
12	General Malaise	Body Integer	Yes = 1, No = 0	1, 0
13	Chest Pain	Integer	Yes = 1, No = 0	1, 0
14	Abdominal Pain	Integer	Yes = 1, No = 0	1, 0
15	Nausea	Integer	Yes = 1, No = 0	1, 0
16	Vomiting	Integer	Yes = 1, No = 0	1, 0
17	Coughing	Integer	Yes = 1, No = 0	1, 0
18	Dizziness	Integer	Yes = 1, No = 0	1, 0
19	Confusion	Integer	Yes = 1, No = 0	1, 0
20	Backache	Integer	Yes = 1, No = 0	1, 0
21	Restless	Integer	Yes = 1, No = 0	1, 0
22	Flue	Integer	Yes = 1, No = 0	1, 0
23	Problem breathing	Integer	Yes = 1, No = 0	1, 0
24	Anemia	Integer	Yes = 1, No = 0	1, 0
25	Yellow skin	Integer	Yes = 1, No = 0	1, 0
26	Bloody stool	Integer	Yes = 1, No = 0	1, 0
27	Appetite loss	Integer	Yes = 1, No = 0	1, 0
28	Conversion	Integer	Yes = 1, No = 0	1, 0
29	Dehydration	Integer	Yes = 1, No = 0	1, 0
30	Pale	Integer	Yes = 1, No = 0	1, 0
21	Running Nose	Integer	Yes = 1, No = 0	1, 0
32	Blurred vision	Integer	Yes = 1, No = 0	1, 0
33	Pain in urination	Integer	Yes = 1, No = 0	1, 0
34	Palpation	Integer	Yes = 1, No = 0	1, 0
35	Diarrhea	Integer	Yes = 1, No = 0	1, 0
36	Diagnosis	Categorical	Positive = 1, Negative = 0	1,0

4.3.3 Feature Selection

Feature selection is one of the vital processes for machine learning model development because it includes irrelevant features affect the classification performance of the machine learning

model. Identifying features (variables) associated with malaria diagnosis and treatment is vital in achieving successful malaria prediction. Feature selection is an efficient data pre-processing technique in data mining to reduce data dimensionality (Jain & Singh, 2018). It is essential to identify the most important risk factors related to the disease in medical diagnosis. Relevant feature identification helps remove unnecessary, redundant attributes from the disease dataset, giving quick and better prediction results (Spencer *et al.*, 2020). This section aimed to identify significant features for malaria diagnosis both in low and high-endemic areas of Tanzania. To achieve the primary goal, the two questions were answered. First, the features and their importance would vary for high and low-endemic regions.

The malaria diagnosis dataset was used to produce three different feature sets. The first feature set was derived by applying the features selection to a dataset consisting entirely of patients from the Kilimanjaro region (low endemic area), the second from a dataset consisting entirely of patients from the Morogoro region (high endemic area), and the third from a dataset consisting of patients from both the Morogoro and Kilimanjaro regions (combined areas). This study employed a model-based feature selection approach to narrow the dataset to the most relevant features in diagnosing malaria. This technique relies on supervised machine learning algorithms to evaluate the significance of each feature. Model-based feature selection has two approaches: feature importance and selection from the model to select the most significant features (Brodersen *et al.*, 2011). The random forest used the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node, which means a decrease in impurity over all trees (Lozano *et al.*, 2021; Liang *et al.*, 2020). This approach improved the purity of the node while naturally ranking and using tree-based tactics. The impurity decreases most noticeably at the nodes at the beginning of the trees and least noticeably at the nodes at the ends of the trees. Thus, pruning the trees below a specific node produced a subset of the most crucial traits.

To minimize the complexity and improve the model's performance, the top 10 important features were selected for the regional datasets and 15 important features for the combined malaria dataset, as shown in Table 11. Both features were obtained from the feature selection methods and were employed for the models' development. The evaluation criteria applied is if the accuracy of the model trained using the dataset with the important features is higher than the full features dataset. The selected important features are considered significant for the classification of malaria and will be used for the malaria prediction model development.

After that the healthcare workers from all the study sites (Morogoro Regional Hospital, Mzumbe Health Centre, Mawenzi Regional Hospital and Majengo Health Centre) were consulted to assess and give their perspective on the important features of malaria diagnosis selected by the model. Apart from that, the healthcare workers were also asked on the feasibility of using the malaria diagnosis model in their work settings. As shown in Appendix 4, the six medical officers used the questionnaires to get a deep understanding of the knowledge of malaria diagnosis and disease management in general. This assessment focused on the evaluation of main, supporting, and severe symptoms of malaria and non-symptom-related factors that could contribute to malaria diagnosis.

4.3.4 Prediction Classifiers

After the dataset was described and pre-processed, features were selected based on different machine learning algorithms and the importance of every feature in the predictive variable was done. Then, machine learning classification algorithms were used to classify the patients with malaria and those who do not have malaria. The popular disease diagnosis machine learning classifiers, which are Logistics Regression (LR), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF), were used in model development. Finally, the machine-learning classifiers' performance for malaria diagnosis and feature selection was computed and compared to obtain the best performing model.

4.3.5 Machine Learning Classifiers Validation

The study used the repeated K-fold cross-validation (CV) method and four performance evaluation metrics. In repeated k-fold cross-validation, the data set was divided into k equal size of parts. The $k - 1$ group was used to train the classifiers, and the remaining portion was used to check the outperformance in each step. The execution was repeated a number of times to attain the optimum results. The process of validation was repeated k times. The classifier performance was computed based on k results. For CV, different values of k were selected. In this experiment, $k = 10$ was used because of its good performance and recommendations in many pieces of literature. In the 10-fold CV process, 70% of data were used for training, and 30% were used for testing purposes. The process was repeated ten times for each fold of the process. All training and test groups instances were randomly divided over the whole dataset before selecting and testing new sets for the news cycle. At the end of the 10-fold process, averages of all performance metrics were computed.

4.3.6 Machine Learning Model Performance Evaluation

Various performance evaluation metrics were used in this study to check the performance of the classifiers. First, a confusion matrix was used, and every observation in the testing set was predicted in precisely one box Table 1. Two matrix approach was deployed because there were two classes which were malaria positive (1) and malaria negative (0). Moreover, it gives two types of correct predictions of the classifier and two classifiers of incorrect prediction. Apart from that classification report was computed to get the classification accuracy, precision, recall and F1 score of the classifiers. From the confusion matrix, TP: predicted output as true positive (TP), it was concluded that the positive malaria subject is correctly classified and subjects have malaria. TN: predicted output as true negative (TN); it was supposed that a negative malaria subject is correctly classified and healthy. Predicted output as false positive (FP), it was concluded that a negative malaria subject is incorrectly classified as having malaria (a type 1 error). FN: predicted output as false negative (FN), it was concluded that a positive malaria subject is incorrectly classified as the subject does not have malaria as the subject is healthy (a type 2 error).

4.3.7 Development of Regional-Specific Malaria Diagnosis Models

To develop a model that fitted the dataset and attained high prediction accuracy and the algorithm that works for all the regions were the factors that were considered in selecting the algorithm that can be used in feature selection and malaria diagnosis model development. The final regional specific model was developed using the best performing machine learning classifier. The model's performance with the selected important features was evaluated and presented.

4.4 Result

4.4.1 Feature Selection Results

(i) Important features for high endemic area dataset

For the Morogoro dataset (high endemic area), the most important features were the patient's age, fever, abdominal pain, visit date, dizziness, vomiting, headache, sex of the patient, general body malaise, and confusion, as shown in Fig. 12.

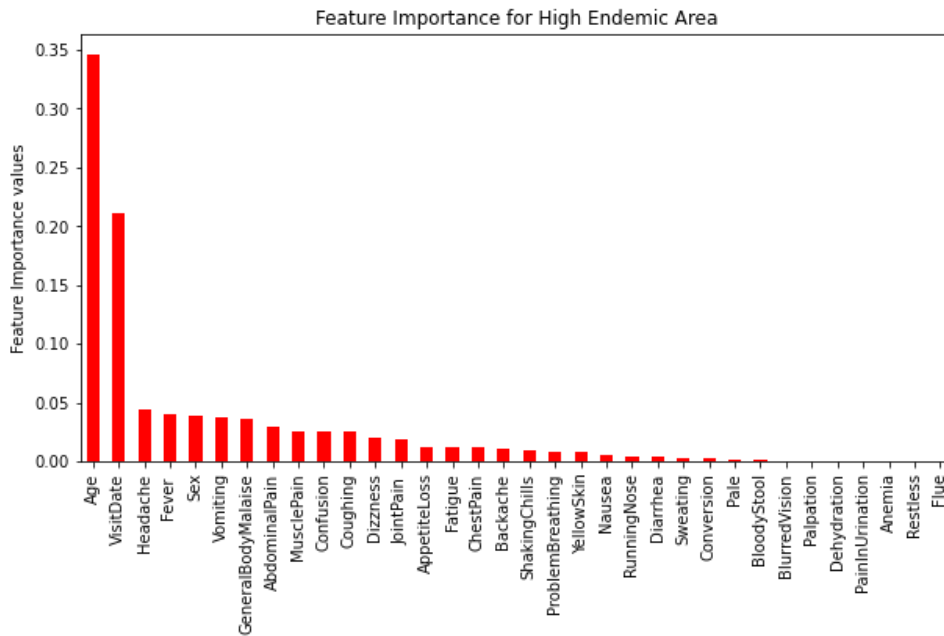


Figure 12: Important Features with Random Forest in High Endemic Area (Morogoro)

(ii) Important features for low endemic area dataset

Headache, age, vomiting, visiting date, fever, general body malaise, joint pain, coughing, abdominal pain, and sex in the corresponding hierarchy as depicted in Fig. 13, were the most important features in the low endemic areas as represented by the Kilimanjaro dataset.

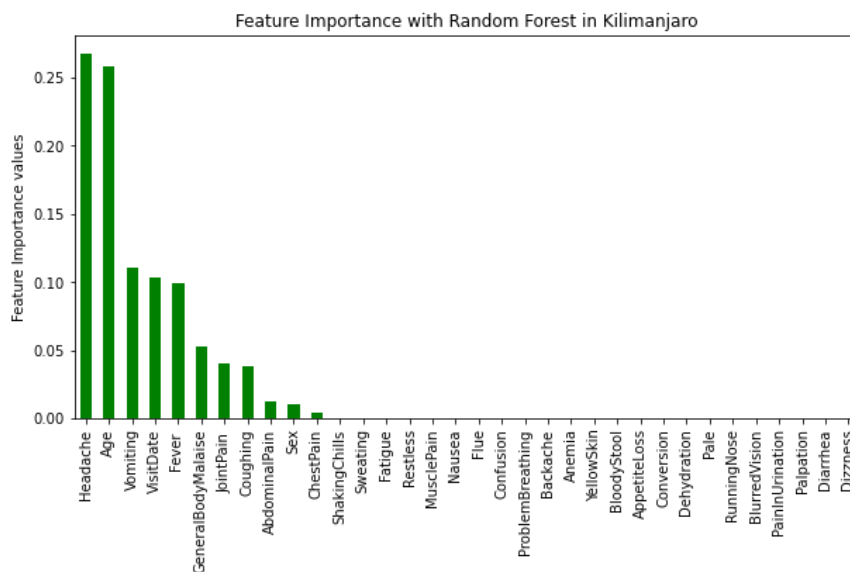


Figure 13: Important features with random forest in low endemic area (Kilimanjaro)

(iii) Important Features for combined areas dataset

From the malaria diagnosis combined dataset, the most important features are residence area of a patient, fever, age of the patient, general body malaise, visit date, headache, abdominal pain, backache, chest pain, sex of a patient, vomiting, confusion, dizziness, coughing and joint pain as shown in Fig. 14.

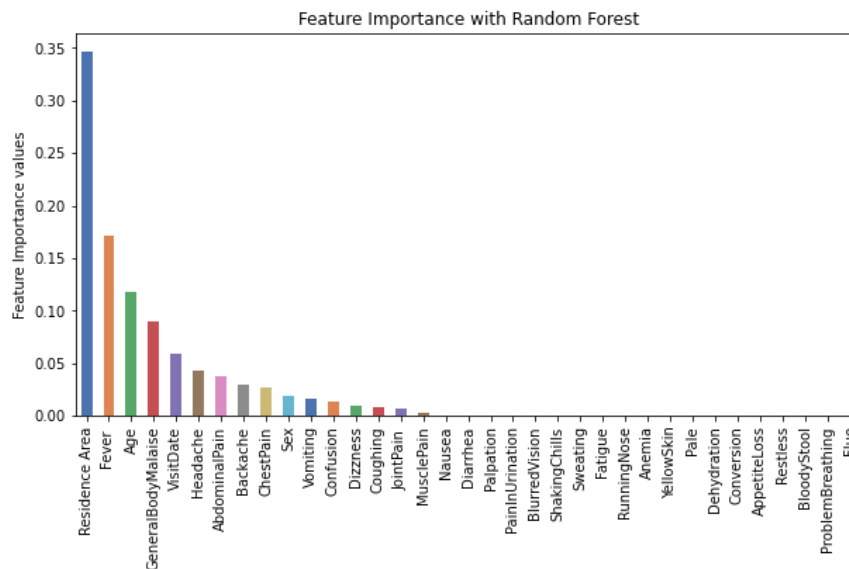


Figure 14: Important features with random forest in combined dataset

(iv) Categorical features correlation

From the important features selected by tree-based methods, categorical features were treated numerically after encoding them. The important features selected are the patient's residence area, visit date, sex and age. Then, the significance of each feature and the subset of these features to the target were computed using correlation analysis. While the sex of the patient shows no importance in diagnosing malaria in all the datasets, the residence area of the patients from the two regions showed a high significance in diagnosing malaria. Apart from that, the visit date is significant in diagnosing malaria. Therefore, the visit date variable was included in the dataset to identify if the time of the visit has any significance in the diagnosis of malaria. Basically, to confirm if there is seasonal malaria. For the Kilimanjaro region, January, February, May and August were more significant, while Morogoro, April, July, and October were more significant than other months. Furthermore, for the combined malaria dataset, January, April, May, August and October were significant to malaria diagnosis. Ages of 12 and

55 years showed significance in the malaria dataset, while ages of 2,15 and 55 years were more significant in Morogoro and Kilimanjaro regions.

Table 11: Regional based important features in malaria diagnosis

Ranking	Full Dataset	High Endemic Dataset	Low Endemic Dataset
1.	Residence Area	Headache	Age
2.	Fever	Age	Fever
3.	Age	Vomiting	Abdominal Pain
4.	General Body Malaise	Visit Date	Visit Date
5.	Visit Date	Fever	Dizziness
6.	Headache	General Body Malaise	Vomiting
7.	Abdominal Pain	Joint Pain	Headache
8.	Backache	Coughing	Sex
9.	Chest Pain	Abdominal Pain	General Body Malaise
10.	Sex	Sex	Confusion
11.	Vomiting		
12.	Confusion		
13.	Dizziness		
14.	Coughing		
15.	Joint Pain		

Nevertheless, it was observed that the ranking of these features was different among datasets where some features which were considered to be the most significant to one region were not as substantial to another region, as shown in Table 11. Apart from that, features specific to a particular region, for example, Joint Pain and Dizziness symptoms, were only significant in the Kilimanjaro region. Muscle Pain and Confusion were only important in the Morogoro region. From the malaria diagnosis combined dataset, the most important features are the residence area of a patient, fever, age of the patient, general body malaise, visit date, headache, abdominal pain, backache, chest pain, sex of a patient, vomiting, confusion, dizziness, coughing and joint pain.

4.4.2 Important Features Validation: Healthcare Worker's Perspective

The study revealed that medical doctors commonly use the main symptoms for malaria clinical diagnosis: fever, shivering, and headache. Furthermore, the doctors use nausea and vomiting, dizziness, loss of appetite, diarrhoea, joint pain, limpness, abdominal pain, and heartburn as the supporting symptoms for the clinical diagnosis of malaria. Moreover, according to medical doctors, malaria can also be identified using severe symptoms such as loss of consciousness, anaemia, jaundice, enlarged spleen, seizures, and shortness of breath if the disease is considered severe. The study finds that the more experienced doctors understand the disease more than the new doctors in the respective area of work. Also, severe malaria is clinically presumed when intensity increases among the observed symptoms. The residence area was significant among patients, and knowing the patient's travel history from low-endemic to high-endemic areas is essential in diagnosing malaria. Table 12 and Table 13 summarises the medical doctor's perspective on malaria symptoms and factors for malaria diagnosis.

The medical doctors also identified excessive vomiting, coca cola urine (urine with brown colour), confusion and loss of consciousness as some of the severe. All the doctors agree that age is significant, especially for children under five and adults below 35. The doctors also agreed that knowing if there was a family member of the patient who was diagnosed with malaria was important since it could have been for the disease to have been transferred to them by the insect. Seasonal malaria was approved by all the doctors as one of the important features in diagnosing malaria in both low and high-endemic areas. Also, the patient's occupation is important since some work environments are more prone to mosquitos and hence a danger to malaria.

Table 12: Medical doctors' perspective on malaria diagnosis symptoms

	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Doctor 5	Doctor 6
Working Experience	3 years	5 years	7 years	4 years	6 years	5 years
Regions practiced	Kilimanjaro Dar es salaam	Tanga Kilimanjaro Shinyanga	Mwanza Tanga Morogoro	Morogoro Dodoma Shinyanga	Arusha Morogoro Mbeya	Kilimanjaro Dar es salaam Tanga
Main Symptoms	Headache Backache Vomiting	Headache, Vomiting, High fever, Body Pain, Diarrhoea	Fever, Headache Vomiting, Nausea	Fever, Body pain, Headache, Vomiting	Fever, Headache Sweating, Body pain	Headache, Backache, Vomiting
Severe Symptoms	Dizziness Anaemia Confusion	Excessive vomiting Yellow fever Paleness	High fever, Vomiting, Anaemia, Conversion, Loss of consciousness, Coca- Cola urine	Confusion fainting Kidney failure	Confusion, Loss of conscious	Dizziness, Anaemia, Confusion
Residence Area	Yes	Yes (warm areas support parasite growth)	Yes (awareness differ, the use of control initiative differs)	Yes	Yes	Yes
Age	Yes (children under 5)	No	Yes (kids <5)	Yes (<35)	Yes	Yes (children under 5)
Sex	No	Yes (pregnancy)	Yes (pregnancy)	No	No	No

Table 13: Medical doctors' perspective on other factors to be considered for malaria diagnosis

Other Factors	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Doctor 5	Doctor 6
History of Travelling	Yes (low prevalence area)	Yes	Yes (from high to low)	Yes	Yes	Yes (low prevalence area)
Using Control initiatives	Yes	No	Yes	yes	Yes	Yes
Family member being sick	Yes	Yes	Yes	Yes	Yes	Yes
Distance from the health facility	Yes	No	Yes	Yes	Yes	Yes
Yearly season	Yes (rainy season)	No	Yes	Yes (planting season)	Yes	Yes (rainy season)
Occupation	Yes (work during the night, farmers)	Yes (eg drivers)	Yes	yes	Yes	Yes (work during the night, farmers)
Malaria prediction Model feasibility	Yes, with concerns	Yes, with concerns	Yes, with concerns	Yes, with concerns	Yes, with concerns	Yes, with concerns
Reasons for Concerns	Substitute the lab confirmation of the disease	Substitute the lab confirmation of the disease	Substitute the lab confirmation of the disease	Substitute the lab confirmation of the disease	Substitute the lab confirmation of the disease	Substitute the lab confirmation of the disease

4.4.3 Machine Learning Classifiers Performance with Important Features

(i) Classifiers Performance on Full Features with K-Fold Cross-Validation

In this experiment, the five-machine learning classifiers were checked with 10-fold cross-validation methods in full 35 features of the complete malaria diagnosis dataset as described in Table 10. While different parameter values were passed through classifiers, the mean of 10-fold methods was computed.

Table 14: 10-fold CV classification performance evaluation of different classifiers on malaria diagnosis dataset on full features

Predictive Model	Classifiers performance evaluation metrics (%)				
	Accuracy	AUC	Sensitivity	Specificity	Precision
Logistic Regression	75	76	77	57	74
K Nearest Neighbour	72	69	78	49	71
Random Forest	79	80	82	69	71
Support Vector Machine	73	75	74	61	71
Decision Tree	72	72	85	58	77

From this experiment with full features on a full malaria diagnosis dataset, Random Forest classifier showed overall good performance among other classifiers with a classification accuracy of 79%, AUC of 80%, Sensitivity of 82%, Specificity of 69%, Precision of 71% and recall of 76% as shown in Table 4. The specificity value of Random Forest was 69% showing the probability that a diagnostic test was negative and the person does not have malaria. The decision tree classifier has demonstrated exemplary performance on Sensitivity of 85%, precision of 77% and recall of 76%. The K-Nearest Neighbour classifier has underperformed on the Specificity of 49% and AUC of 69% but scores the Sensitivity of 78%, precision of 71% and accuracy of 72%. The Support Vector Machine achieved an accuracy of 73%, specificity of 61%, precision of 71%, AUC of 74% and Sensitivity of 74%. Apart from that, the Logistic regression classifier achieved an accuracy of 75%, specificity of 57%, precision of 74%, AUC of 76% and Sensitivity of 77%. The performance comparison on AUC, Specificity and Sensitivity among the classifiers is shown in Fig. 15.

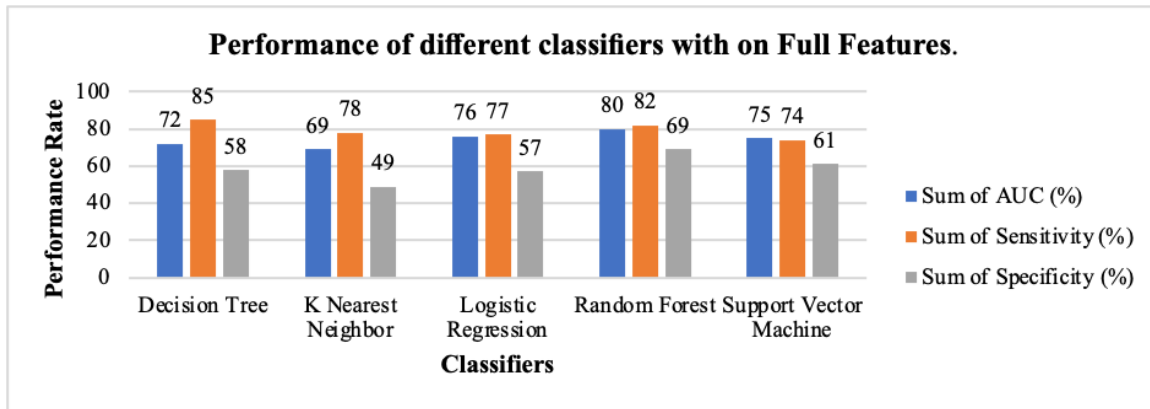


Figure 15: AUC, Sensitivity and Specificity performance of different classifiers on full features dataset

(ii) Classifiers Performance on Selected Important Features with 10 -Fold Cross-Validation

The experiment was performed using only ten important features selected during the feature engineering process. In this experiment, all classifiers had high performance in all metrics compared to when the full features were used Table 15. For the Accuracy and AUC, the Random Forest classifier has the best performance with an accuracy of 82% and AUC of 83%, followed by the Logistic Regression classifier with an accuracy of 76% and AUC of 78%. Random forest and Decision Tree classifiers have the best 81% and 76% precision, respectively. These models confidently predict true negatives that 81% of the negative malaria prediction were healthy (with no malaria).

Table 15: 10-fold CV classification performance evaluation of different classifiers on malaria diagnosis dataset ten important features

Predictive Model	Classifiers performance evaluation metrics				
	Accuracy	AUC	Sensitivity	Specificity	Precision
Logistic Regression	75	73	76	63	73
K Nearest Neighbour	72	70	80	60	71
Random Forest	82	83	84	74	81
Support Vector Machine	74	75	75	58	71
Decision Tree	74	73	85	54	76

For the classification of confident true positive that does not classify a sick patient as a healthy person, Decision Tree performed well with a Sensitivity of 85%, followed by Random Forest with Sensitivity of 84%. In this dataset, Random Forest had an F1 score of 81%. Support Vector

Machine had the best performance on Specificity by 74%, while the KNN classifier performed the least in all aspects with the score of 72% accuracy, 70% AUC, 80 % sensitivity, 60% specificity and 71% precision. It was also established that the Logistic Regression classifier's accuracy and AUC dropped after selecting the important features. The average accuracy and AUC dropped from 76% and 75% to 75% and 73%, respectively, as shown in Fig. 16. This signifies that the dropped features dominated the predictive capacity of this classifier.

(iii) Classifiers Performance on Selected 10 Important Features on Regional Datasets

The ten selected important features from every regional dataset were checked on five machine learning classifiers with a 10-fold cross-validation method. The average AUC, Sensitivity and Specificity results for the Kilimanjaro dataset were presented in Fig. 17. The machine learning classifiers were trained and tested in phases with different features to see features that would bring the best performance. First, the classifiers trained and tested the three most important features. Then three important features were added, and the last four important features were fed. It was observed that the performance of the classifiers was good at the ten important features. Results of classification accuracy, AUC, Specificity, Sensitivity, Precision and F-1 score on different graphs were used for better demonstration. These performance metrics were computed automatically.

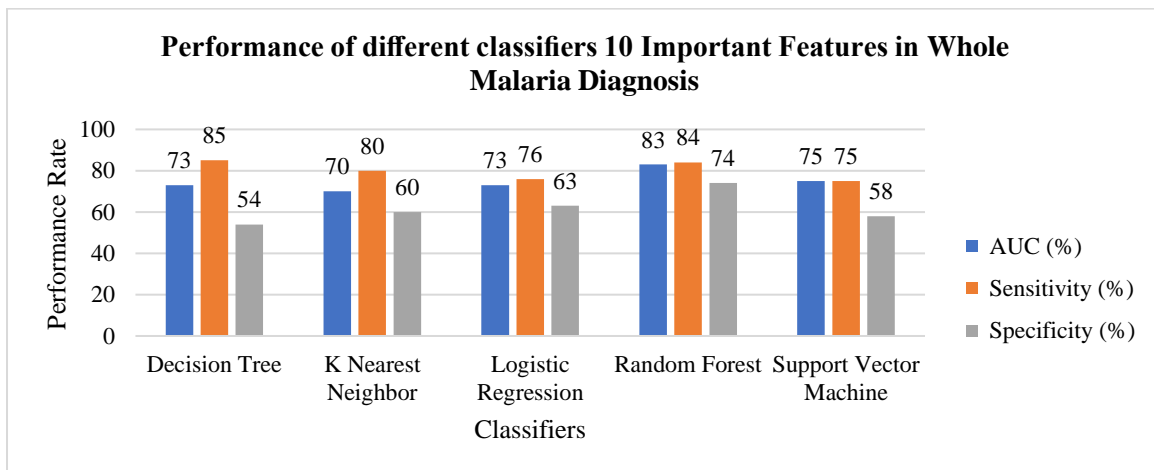


Figure 16: AUC, Sensitivity and Specificity performance of different ML classifiers on important features of the whole Malaria diagnosis dataset

In both experiments, Random Forest classifier has shown outstanding performance with 95% and 87% classification accuracy, 96% and 85% Sensitivity, 92% and 78% Specificity, 92% and

80% Precision, 97% and 86% AUC for Kilimanjaro and Morogoro respectively. This classifier has outperformed all the other classifiers in all performance metrics. The Decision Tree classifier performed second best to Random Forest, and its performance in the Kilimanjaro dataset is better than in the Morogoro dataset. While the classifier archived well with 92% classification accuracy, 91% Sensitivity, 80% Specificity and 80% Precision in the Kilimanjaro dataset, its Specificity and Precision was poor by 67% and 68% in the Morogoro dataset.

For the Logistic Regression classifier, the classification accuracy scores, AUC and Sensitivity were good by 81%, 82% and 85%, respectively, for the Kilimanjaro dataset and 76%, 77% and 74% for the Morogoro dataset, respectively. On the other hand, the classifier had an unsatisfactory performance on Specificity (65%) and Precision (65%) in Kilimanjaro dataset and 68% Specificity, 67% Precision for Morogoro dataset. K-Nearest Neighbour performed well on the same metrics as Logistic Regression in all the datasets. Unlike Logistic Regression and KNN classifiers, Support Vector Machine classifier had a pretty good performance in all metrics for all the datasets, as shown in Table 16.

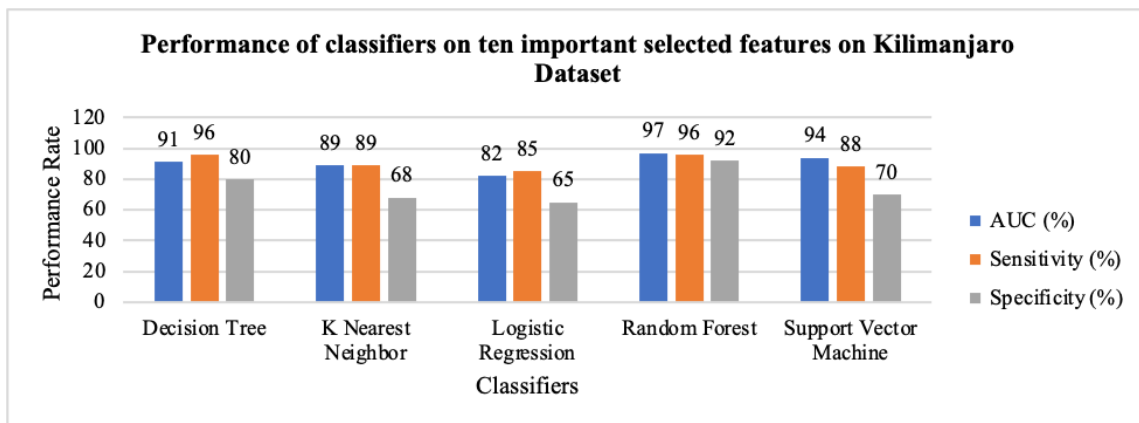


Figure 17: AUC, Sensitivity and Specificity performance of classifiers on ten important features on Kilimanjaro dataset

The main aim of conducting these experiments was to create a machine learning model that can classify patients correctly with malaria from healthy patients based on the symptoms presented and some of the patient's demographic information. When the classification accuracy of the classifiers was compared between the regional datasets, Random Forest was found to be the best classifier with 95% accuracy for the Kilimanjaro dataset and 87% accuracy for the Morogoro dataset, as shown in Fig. 19.

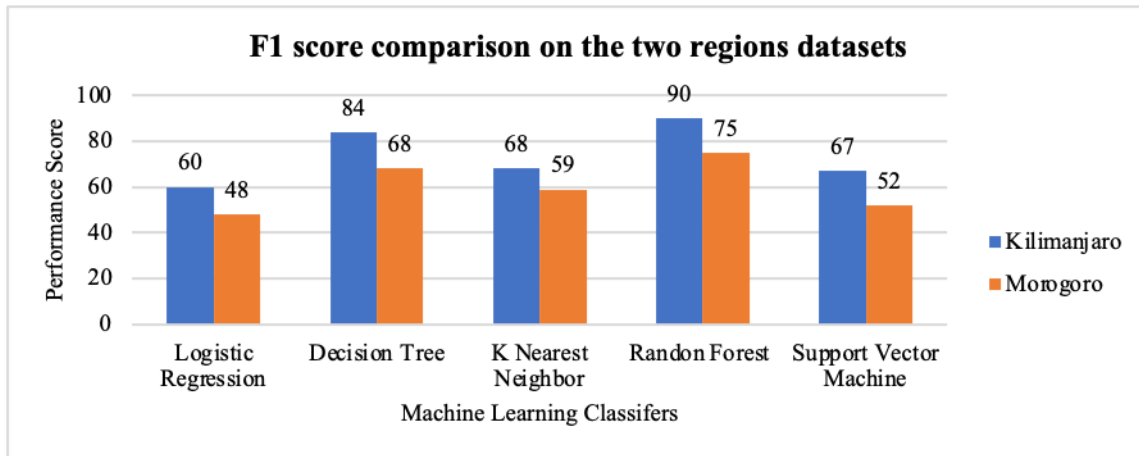


Figure 18: F1 score comparison on the two regions' datasets

The Sensitivity score of the classifiers in each dataset is shown in Fig. 20. Random Forests and Decision Trees classifiers showed a 96% sensitivity performance in Kilimanjaro. Morogoro's Random Forest classifier showed a good performance of 85% Sensitivity. The experiment also identified the harmonic mean between Precision and Recall (F1 score), which tells how precise and robust the classifier incorrectly classified the true negative and truly positive. As shown in Fig. 18, the Random Forest classifier performed with a 90% F1 score in the Kilimanjaro region dataset and a 75% F1 score in the Morogoro region dataset, as also shown in the ROC plot in Fig. 21. The summary of performance of classifiers are presented in Table 16 and Table 17. The summary of excellent performance metrics results and best classifiers are presented in Table 18.

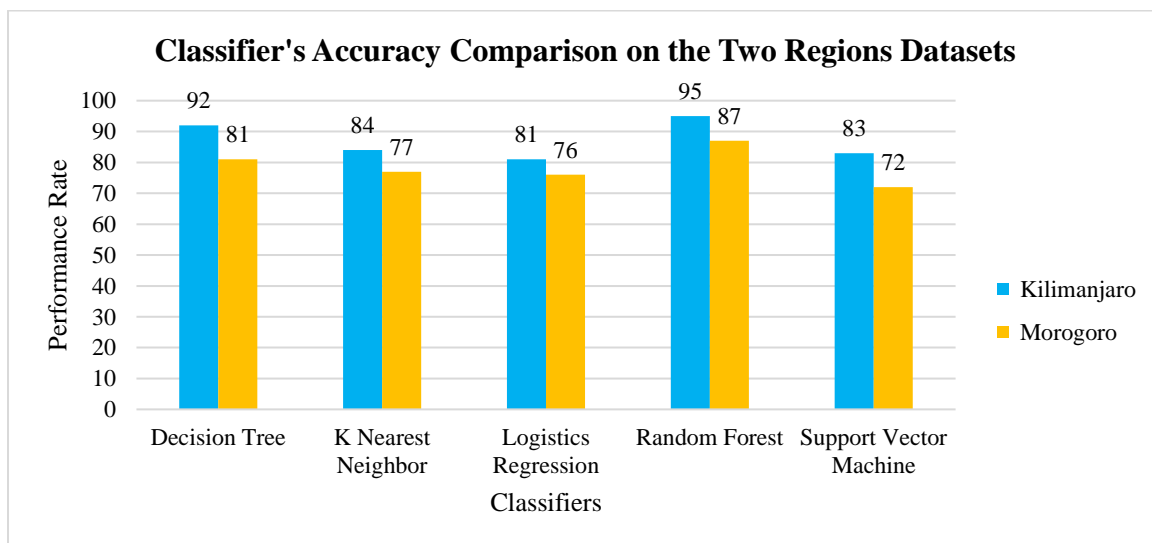


Figure 19: Classification Accuracy comparison in two regions dataset

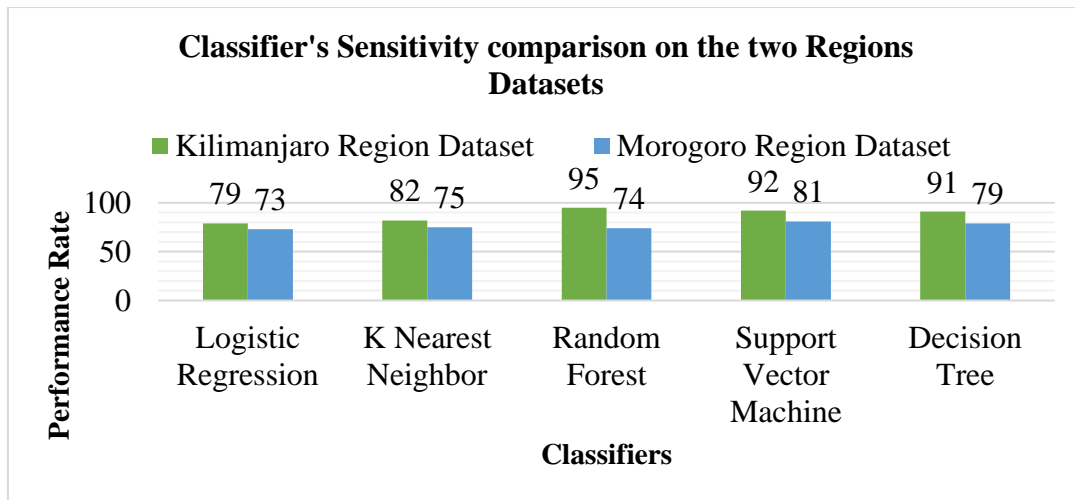


Figure 20: Classifier's Sensitivity comparison on the two regions' datasets

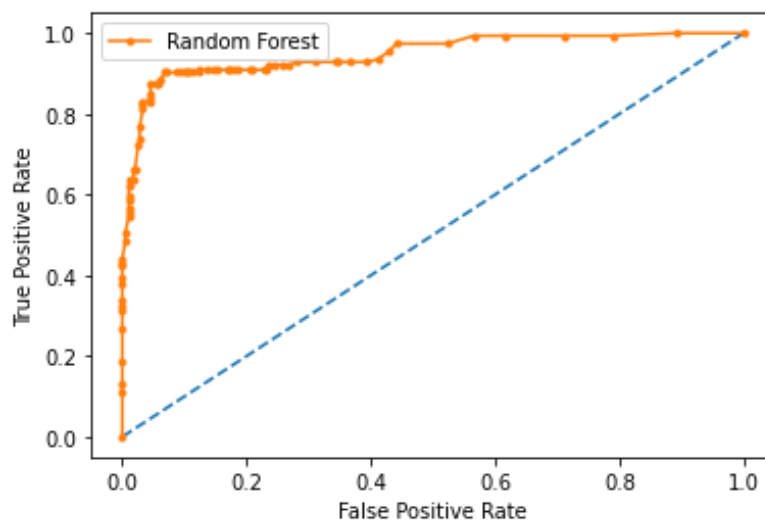


Figure 21: ROC plot for Random Forest performance evaluation

Table 16: Summary of classifiers performance on Morogoro dataset (%)

Predictive Model	Accuracy	AUC	Sensitivity	Specificity	Precision	Recall	F 1
Logistic Regression	76	77	74	68	67	38	73
K Nearest Neighbour	77	76	77	62	63	56	75
Random Forest	87	86	85	78	80	71	74
Support Vector Machine	72	80	75	77	76	40	81
Decision Tree	81	77	82	67	68	68	79

Table 17: Summary of classifiers performance on Kilimanjaro dataset (%)

Predictive Model	Accuracy	AUC	Sensitivity	Specificity	Precision	Recall	F 1
Logistic Regression	81	82	85	65	65	56	79
K Nearest Neighbour	84	89	89	68	68	69	82
Random Forest	95	97	96	92	92	89	95
Support Vector Machine	83	94	88	70	70	64	92
Decision Tree	92	91	96	80	80	89	91

Table 18: Excellent performance metrics results and best classifiers

Datasets	Accuracy (%) and the best classifier	Sensitivity (%) and the best classifier	Specificity (%) and the best classifier	Precision (%) and the best classifier	AUC (%) and the best classifier	F 1 score and the best classifier
Kilimanjaro Dataset	95% Random Forest	96% Random Forest and Decision Tree	92% Random Forest	92% Random Forest	97% Random Forest	95% Random Forest
Morogoro Dataset	87% Random Forest	85% Random Forest	78% Random Forest	80% Random Forest	86% Random Forest	81% Support Vector Machine
Combined Malaria Diagnosis Dataset	82% Random Forest	85% Decision Tree	74% Random Forest	81% Random Forest	83% Random Forest	79% Random Forest

4.4.4 Development of Final, Regional-Specific Malaria Diagnosis Models

The final regional-specific model was able to classify the previously unseen malaria diagnosis sets of data (testing sets) at an accuracy of 86% for the combined dataset, 84% for low endemic and 88% for high endemic areas. Although almost all the models attained high accuracy, as described in the preceding paragraphs, the study adopted a combination of RF and DT as a

feature selection method and model development since these two algorithms showed the best prediction accuracy, respectively. Random forest was selected for feature selection because it was robust and had high performance. Random forest is considered a robust model based on its ability to do an intensive search of features that can maximize prediction accuracy. Random Forest and Decision Tree were adopted for model selection because this study aimed at modelling the decision-making process for the patients who have presented malaria. Decision Tree naturally has been constructed specifically for modelling decisions. The aim of this study is to make the decision on whether the patient has malaria or not based on the symptoms and non-symptomatic features presented therefore, we found that DT is coherent in presenting the relevant information and so relevant to our study to make the decision needed. These two models' representation style gives a decisionmaker alternative solutions and possible choices, making it easier to make a well-informed choice. Decision Tree uses 'what if' thoughts for decision-makers to scrutinise certain choices' possible risks and benefits. Additionally, both DT and RF accommodate nonlinear relationships compared to other models. Therefore, the DT was used to depict the variables or combination of variables with the most predictive power in malaria diagnosis. The study used Gini Index value way of splitting a decision tree. The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly (Afzali & Karnon, 2014; Kingsford & Salzberg, 2008; Puspitasari *et al.*, 2022). This measures the impurity value of a split condition (Tangirala, 2020). The Gini value ranges from 0 (highest purity) to 0.5 (high impurity). The lower the Gini Index, the better the lower the likelihood of misclassification (Kingsford & Salzberg, 2008). Formula for calculating Gini index $Gini=1-\sum_{i=1}^n(p_i)^2$ where: 'pi' is the probability of an object being classified to a particular class.

(i) Models to Predict Malaria in Low Endemic Areas (Kilimanjaro Dataset)

In Kilimanjaro, out of 36 variables, ten (10) were selected by a Random Forest algorithm to build a model with a prediction accuracy of 84%. The final form of the decision tree model with a maximum depth of four (4) is shown in Fig. 22. Overall, the age of the patient, having presented headache and fever symptoms, the month of the year when the symptoms were observed (May and January) and abdominal pain showed the highest predictive power compared to other features in Kilimanjaro. The results showed that only a specific combination of features can determine whether a patient has malaria. For example, patients who were <5 years, had headaches and visited the health facility in May are malaria negative while patients

who are above 5 years, and presented the signs of headache and their visit date is not May are malaria positive. Also, the tree shows that the patients will be malaria positive if they present the signs of headache and observe these symptoms in any month but May and September. Another positive diagnosis is observed in patients with general body malaise with a not-so-strong Gini value of 0.4.

(ii) Models to Predict Malaria in High Endemic Areas (Morogoro Dataset)

The same as the Kilimanjaro dataset, in this dataset, 10 important features were selected with an RF algorithm to build a model which obtained a prediction accuracy of 88%. The final form of the decision tree model with a maximum depth of 4 is shown in Fig. 23. Fever, age, the month of visit (April and February), muscle pain and vomiting showed the highest predictive power compared to other features in Morogoro. The decision tree showed that only a specific combination of features could determine whether a patient has malaria rather than a single symptom. The tree shows the Fever and age of the patient and the month the symptoms were observed to be dominant to malaria-positive patients. Other symptoms are Muscle pain, vomiting and abdominal pain. For example, patients who presented signs of fever, aged between 5 and 22 years and visited the health facilities in January and April are likely to be malaria positive.

(iii) Models to Predict Malaria using a Combined Malaria Diagnosis Dataset (Morogoro and Kilimanjaro)

As for the combined dataset, 15 important features were selected using the same algorithm to build a malaria diagnosis model. The malaria diagnosis model acquired a prediction accuracy of 86%. The combined dataset was computed to represent the country in general since it carries both low and high endemic areas. Figure 24, shows the final form of the decision tree model with a maximum depth of 4. The residence area of the patients, fever, abdominal pain, back pain symptoms and age of the patients are the features that showed the highest predictive power in the diagnosis of malaria in the combined malaria diagnosis dataset. For example, some of the rules that were observed from the tree are for patients that: (a) Reside in Majengo, they are of age 40 and above, and they have neither backache nor fever (Gini value 0.114), (b) Reside in Majengo or Mzumbe, they are age 40 and below, and they don't have fever (Gini value 0.2) (c) Reside in Majengo, with fever, no abdominal pain, and they age below 55 (Gini value 0.145).

(iv) Malaria Prediction Model Performance

The dataset performance was also computed using a Random Forest Algorithm since the algorithm has been widely used to improve the accuracy of predictive models (Geldof *et al.*, 2020; Shaikhina *et al.*, 2019). Table 19 summarises the results obtained on the models' performance. As stated above, models were built and tested based on the top 15 significant features for the combined dataset and 10 significant features for the regional dataset selected by features selection methods. In predicting malaria, the two models had a high performance with prediction accuracy of 96%, 99% and 98% for the Kilimanjaro, Combined and Morogoro datasets, respectively.

Table 19: Models' performance for predicting malaria. The results are accuracies obtained by models developed (Decision tree (DT) and Random Forest (RF)) using different sets of important features selected (%)

Datasets	Random Forest (RF)	Decision Tree (DT)
Low Endemic Area (Kilimanjaro)	96	84
High Endemic Area (Morogoro)	98	88
Combined (Kilimanjaro & Morogoro)	99	86

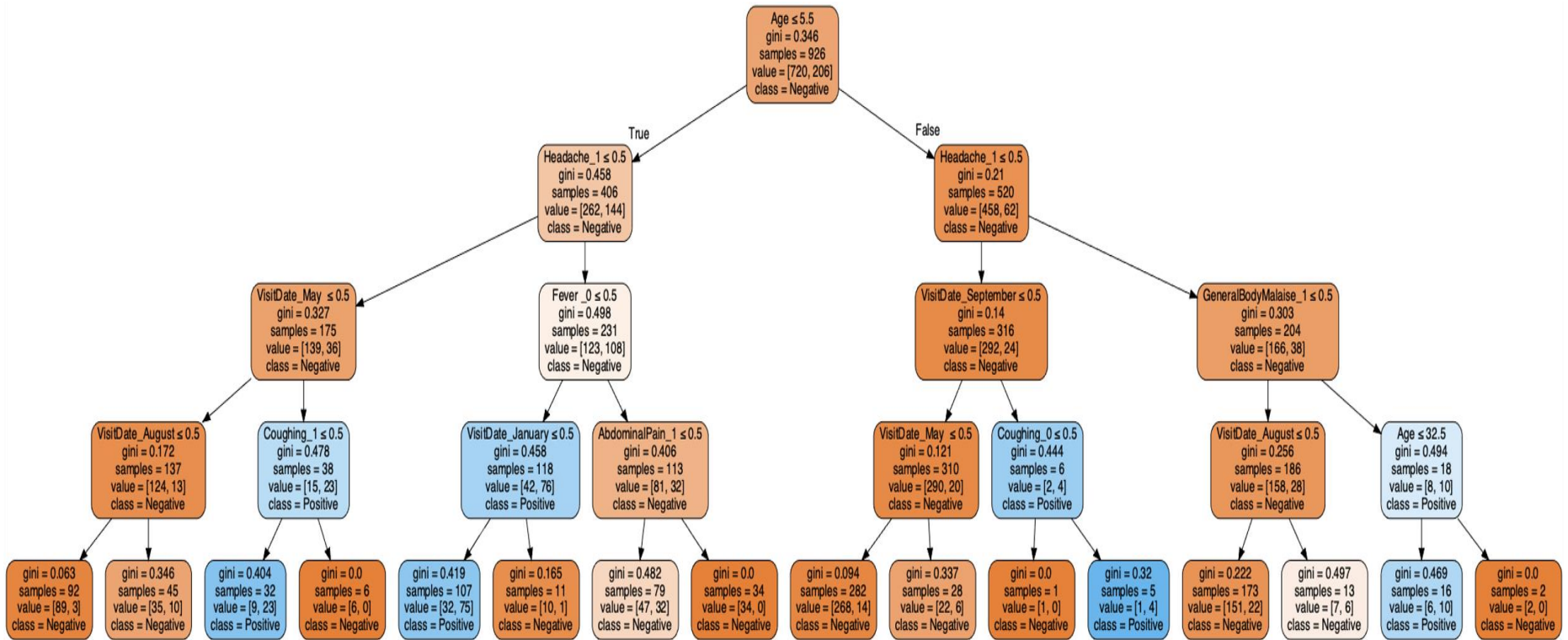


Figure 22: Decision tree for low endemic area (Kilimanjaro)

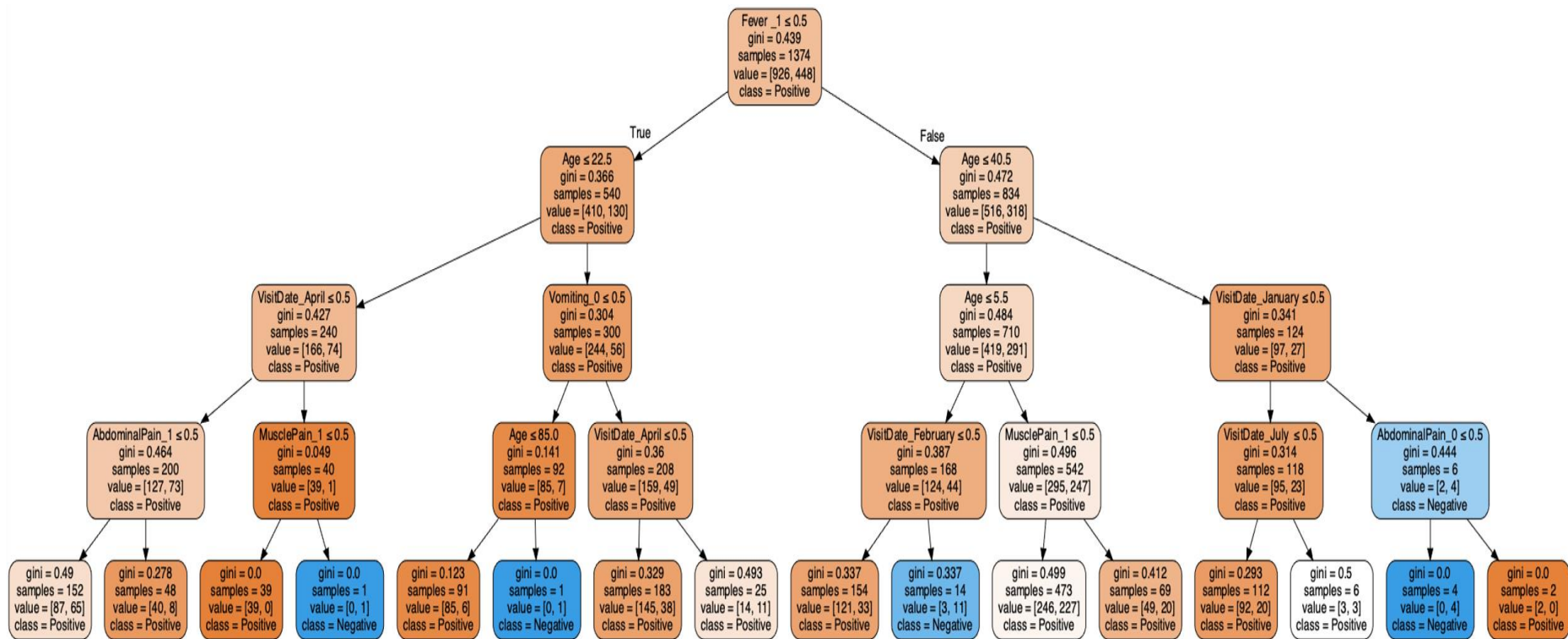


Figure 23: Decision tree for high endemic area (Morogoro)

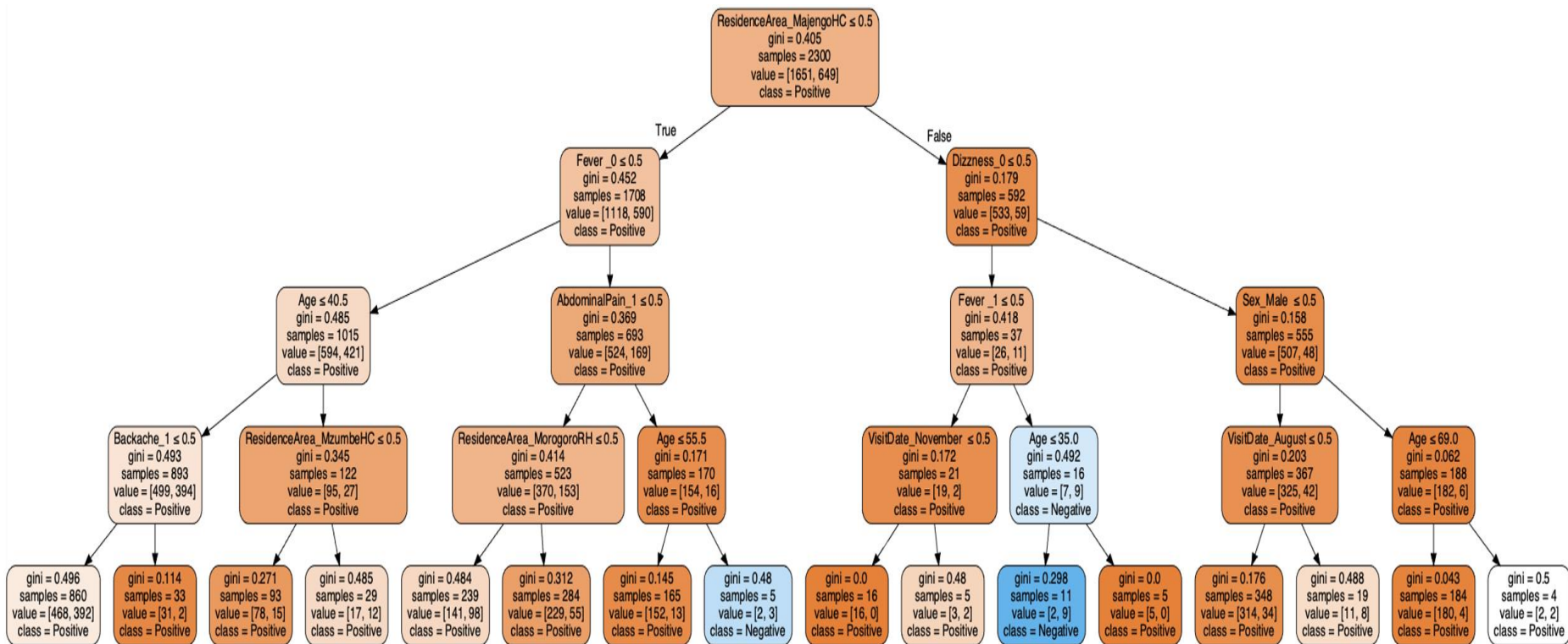


Figure 24: Decision tree for combined malaria diagnosis dataset

4.5 Discussion

Accurate, affordable, timely diagnosis is vital for properly managing any disease. In most developing countries, adequate diagnosis of malaria has been a challenge due to the lack of testing equipment, insufficient personnel to run diagnostic tests, and patients' self-medicating (Sissinto *et al.*, 2019; Gutema *et al.*, 2011; Kajeguka *et al.*, 2017). Using patient symptoms and demographic characteristics, this study illustrates the efficacy of supervised learning models in diagnosing malaria in Tanzania's low and high-endemic areas. The study identified important features that have predictive power in malaria diagnosis, and these features align with the signs and symptoms identified in Tanzania's malaria treatment guidelines. The study also showed that the machine-learning model exhibited high predictive accuracy in diagnosing malaria using the identified important features.

The study discovered that not only some of the symptoms have significance in the diagnosis of malaria but also non-symptoms such as the residential area of the patient, sex and age have significance in the diagnosis of malaria. As depicted in Table 11, the difference in the level of important features for different regions signifies that each region is unique even though they're in the same country and should not be treated the same. The difference can be due to geographical location, which can enhance the rate of disease transmission. Apart from the difference in the level of importance, the experiments showed significant features in one region but no significance in the other. For example, coughing and joint pain are significant for malaria diagnosis in Morogoro. Still, they have zero significance in Kilimanjaro, while dizziness and confusion are important in diagnosing malaria in Kilimanjaro and not in Morogoro. It has also been noted that certain months of the year are particularly important in malaria diagnosis, as this is when most patients seek medical attention for malaria-related symptoms. We focus on the months surrounding and including the rainy season. The World Health Organization's guidelines on how to prevent the spread of malaria align with the results of this study. Climate factors such as rainfall, temperature, and humidity can all impact mosquito populations and, thus, transmission. According to research conducted by Chandramohan *et al.* (2002), Ngasala *et al.* (2008) and Nkumama *et al.* (2017) in many regions, transmission is seasonal, with a surge during and soon after the rainy season. It is possible for malaria epidemics to break out in locations where there is little to no immunity to the disease if the weather or other environmental factors suddenly favour transmission (WHO, 2020). The World Health Organization (WHO) and the malaria treatment guidelines for Tanzania's mainland both

recommend using parasitological confirmation of suspected malaria cases as a diagnostic criterion for patients of all ages exhibiting fever, headache, joint pains, malaise, vomiting, diarrhoea, body ache, weakness, poor appetite, pallor, and an enlarged spleen (Michael & Mkunde, 2017).

The healthcare workers revealed that the severity of the symptoms observed and the duration to which the symptom was observed is very important in knowing the intensity of the disease. It was also revealed that knowing the travel history of the patient is important especially for the patients who live in the low endemic area if they have travelled to the area with high malaria rate since the chances of contracting the disease high. Apart from that knowing if there was a family member of the patient who was diagnosed with malaria was important since it could have been for the disease to have been transferred to them by the insect bite. Seasonal malaria was approved by all the doctors as one of the important features in diagnosing malaria in both low and high-endemic areas. Also, the patient's occupation is important since some work environments are more prone to mosquitos and hence a danger to malaria. With these discoveries, gives an opportunity to improve the attributes of the malaria diagnosis features. For example, adding the travel history and the intensity of the observed symptoms might improve the malaria prediction accuracy.

Six well-known machine learning classifiers, such as Logistics Regression (LR), K-Nearest Neighbour (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF), were used in cooperation with RF a feature selection classifier. In regional and combined datasets, Random Forest showed overall good performance compared to other classifiers with an accuracy of 79%, AUC of 80%, Sensitivity of 82%, Specificity of 69%, Precision of 71% and recall of 76%, followed by Decision Tree. Furthermore, all models performed highly with selected important features except the Logistic Regression recorded lower AUC and accuracy. In addition, the Random Forest classifier has shown strong performance in predicting malaria in both regions. Although the random forest algorithm is considered a black box because the information is hidden inside the model structure, this study adopted it as a feature selection algorithm due to its robustness, execution speed, and intensive searching procedure. Similar findings were described in the studies conducted in Senegal and Burkina Faso, indicating that random forest is a promising classifier with high accuracy in forecasting malaria in the named settings (Harvey *et al.*, 2021; Yadav *et al.*, 2021). Based on the test performance of the models, which in this study were DT and RF, demonstrated that

despite the malaria diagnosis's complexity, machine learning could capture features and model malaria diagnosis differently. The good prediction accuracy attained by classification models signifies that all the features selected among the regions and model development algorithms were useful in their rights. However, differences in the selected features for each dataset signify that these regions are different even if they are in the same country. Hence, searching for the appropriate features per region and understanding the domain science behind disease management is also important.

This study adopted a combination of RF and DT for feature selection and model development. Decision Tree algorithm was used to define classification rules for modelling malaria diagnosis decisions led this study to identify very useful patterns. The results showed that decision tree rules developed in this study showed that a specific combination of features could determine whether a patient has malaria rather than a single symptom. The successful prediction of malaria using a decision tree aligns with the studies by Ilyas *et al.* (2021), Serpen (2016) and Tekale *et al.* (2018) which attained a good prediction accuracy in classifying chronic kidney disease. Another study by Kajungu *et al.* (2012) used a decision tree successfully to model drug prescription practices in Tanzania. Based on the RF classification results obtained, the model has proven to be the most efficient algorithm for the classification of heart disease and therefore it is used in the proposed system. The study by Paul *et al.* (2022) and Polan *et al.* (2021) used the RF model to predict heart disease and other diseases. In general, RF and DT which are tree-based classifiers are considered to be the best classifiers to make medical decisions and for these classifiers performance in this study proves that.

Based on the studies done in Tanzania by Ngasala in 2019 had 95% accuracy of microscope in diagnosing malaria. Even though the microscope has been considered to be the golden standard for testing malaria parasite and is expected to have a diagnosis accuracy of 100% some studies have reported the drop in accuracy in different settings. Compared to the developed model which had a similar accuracy performance in predicting malaria. The results obtained from the two models proved that machine learning could relieve this burden of presumptive treatment of malaria by providing a high-accuracy disease prediction model that doesn't require expensive equipment or trained personnel, just patients' signs and symptoms for the patients that seek treatment at facilities without the recommended equipment or personnel for parasitology tests and those visiting pharmacies for medication without testing can be better assessed for the probability of disease before treatment.

4.6 Conclusion

This study developed a regional-specific malaria predictive model used in malaria diagnosis based on clinical symptoms and demographic data. Our study demonstrated that clinicians could use the model to detect new malaria cases in a clinical setting, provided patients' symptoms and demographic features are available when parasitology testing is unavailable. Furthermore, the study demonstrated that using the right machine learning classifiers and important features for each dataset is useful in predicting the disease correctly. For future studies, results from this study will be a necessary step in designing a decision support system through the developed model, which will be more suitable for people who cannot access laboratory-based diagnosis tools or the health facility before any treatment. The scope of this study is restricted to identifying the key indicators for malaria prediction. For instance, if patient data for the attributes utilised are available, a clinician's pharmacist can use the model developed in further study and with various people to detect malaria in new patients.

CHAPTER FIVE

MALARIA PREDICTION MODEL VALIDATION

5.1 Introduction

The best-performed model in this study was trained using an unseen dataset collected from the same health facilities as the training dataset to validate the malaria diagnosis model. Model validation in machine learning is comparing a trained model to a testing data set to determine whether the model successfully achieves its intended goal (Wang & Zheng, 2013). After model training, model validation is done to assess and identify the best-performing model. There are two basic methods for model validation: (a) in-sample validation, which tests data from the same dataset used to develop the model, and (b) out-of-sample validation, which tests data from a new dataset not previously used to build the model (Gill, 2022). The dataset used to create the training set was the same one from which the testing dataset was created. The major goal of using the testing data set is to evaluate a trained model's capacity for generalisation. The model's deployment forecast accuracy was examined after the malaria diagnosis model was successfully trained and demonstrated to provide accurate data. Models that have been thoroughly validated are resilient enough to accommodate different real-world scenarios (Datatron, 2022; Wang & Zheng, 2013).

5.2 Validation Process with unseen malaria diagnosis dataset

The model validation process was performed using a new set of unseen datasets of the patients' records treated with malaria in 2019 in the study's two regions (Kilimanjaro and Morogoro). The unseen dataset can be defined as all types of data that a model has never learned before (Alhamid, 2020). The validation dataset had a total of 700 patient records with 36 variables where 459 records are from Morogoro region and 223 are from Kilimanjaro region. The dataset was curated from the four health facilities which included a similar set of variables to the malaria diagnosis datasets used to develop and train the developed model as shown in Table 10. The datasets for model validation were also pre-processed as explained in preceding sections. The best-performed algorithms that were selected to develop models. The well-performed models were tested in a validation dataset to assess the models' robustness and reduce the model's overfitting. Decision Tree and Random Forest were used for modelling. The model development process followed all procedures required for model development as elaborated in preceding sections 4.3.2 to 4.3.7. The model was then evaluated based on its

prediction accuracy, sensitivity, specificity, and F1-Score. The main aim of this second model validation process was to check the robustness of the developed model through the use of other datasets with similar characteristics and also to validate if the classifiers can be referred to and adaptable enough to develop other models with new datasets.

5.3 Validation Results

5.3.1 Description of Unseen Malaria Diagnosis Dataset

Malaria patients' distribution according to the laboratory-confirmed results of positive or negative as obtained in the patient's records, residence area, age of patients, patients visiting month and sex from the unseen malaria diagnosis dataset is as seen in Fig. 25, Fig. 26 and Fig. 27. This description makes a better understanding of the dataset.

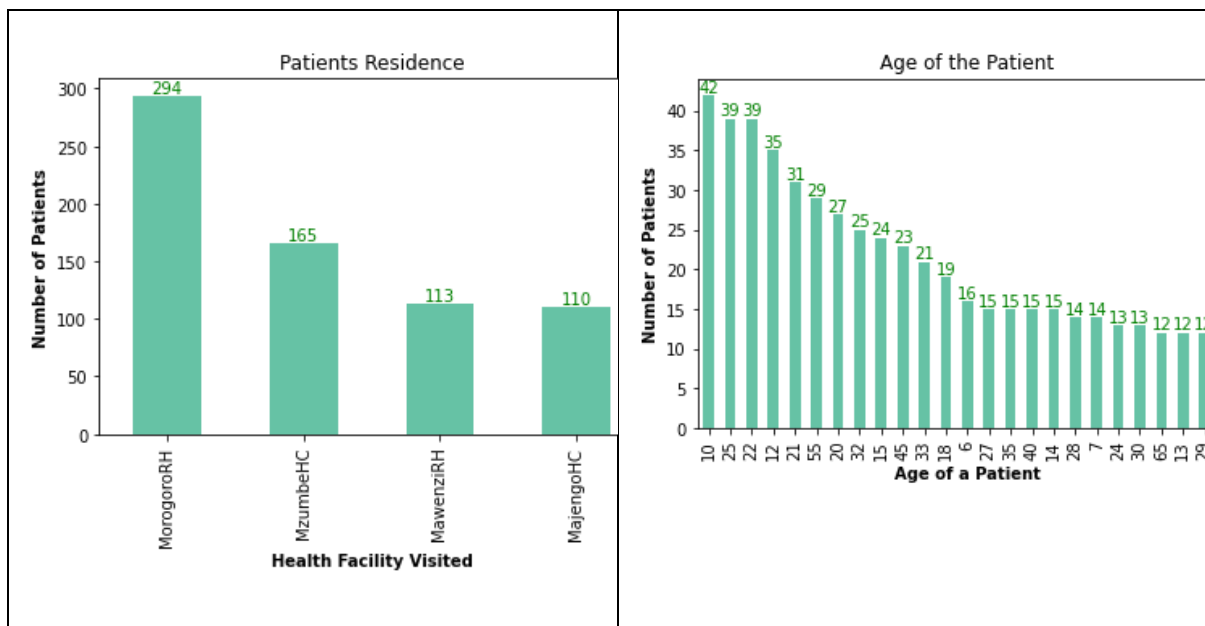


Figure 25: Frequency of residence area and age of the patient

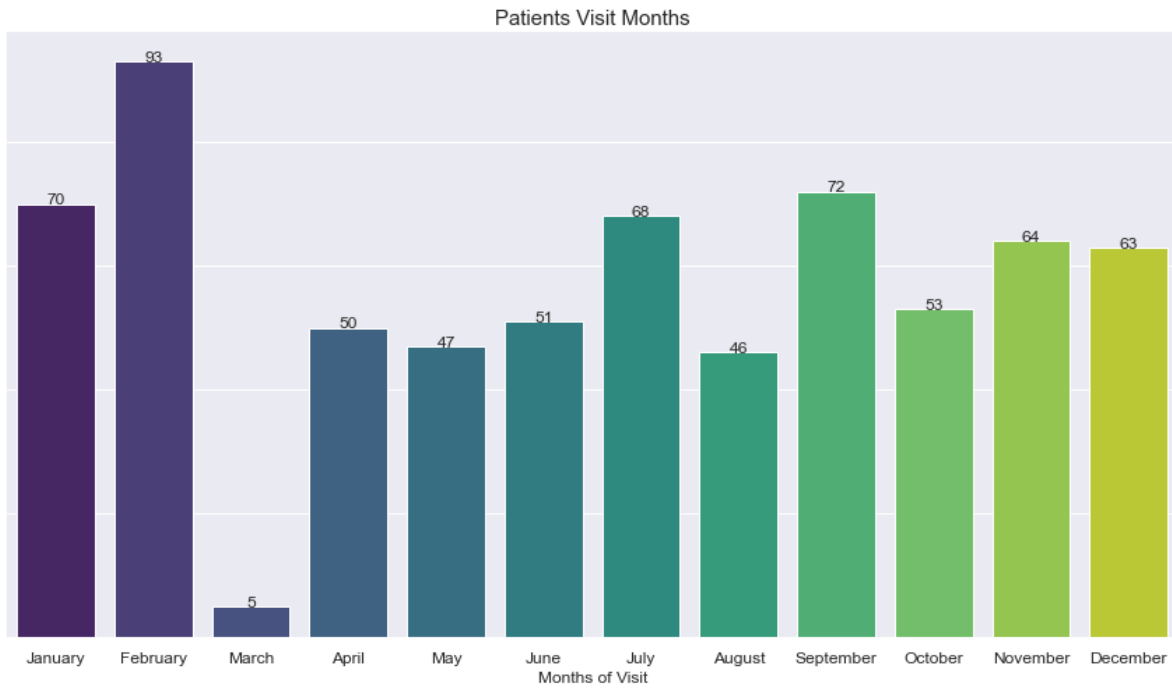


Figure 26: Frequency of patients based on their visit month

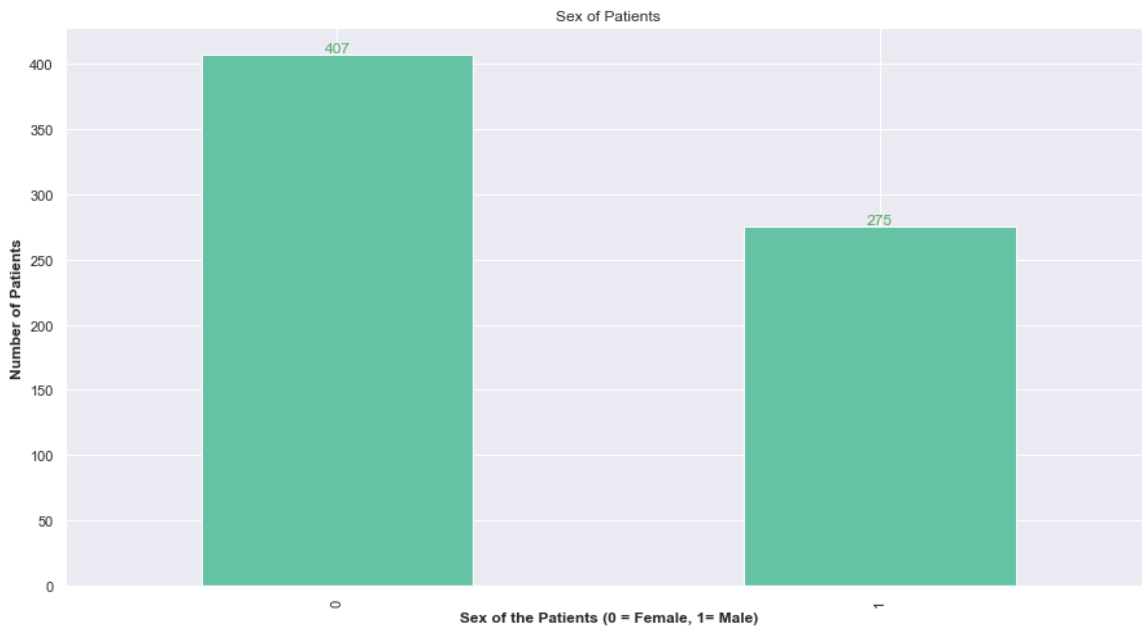


Figure 27: Patients distribution based on their Sex

5.3.2 The validated important features

The malaria diagnosis dataset was divided into a low-endemic area, a high-endemic area and a combined dataset. The model selects the most important features for each subset of the validation dataset as shown in Fig. 28, Fig. 29 and Fig. 30. The random forest model selected

almost the same features but the level of importance differs in each region. For example, in the combined dataset; visit date, age, fever and abdominal pain were the most important features in the diagnosis of malaria while age, visit date, residence area and sex of the patient are important in high endemic areas. The model scored 99% accuracy on training data but the score dropped a little bit on a testing data.

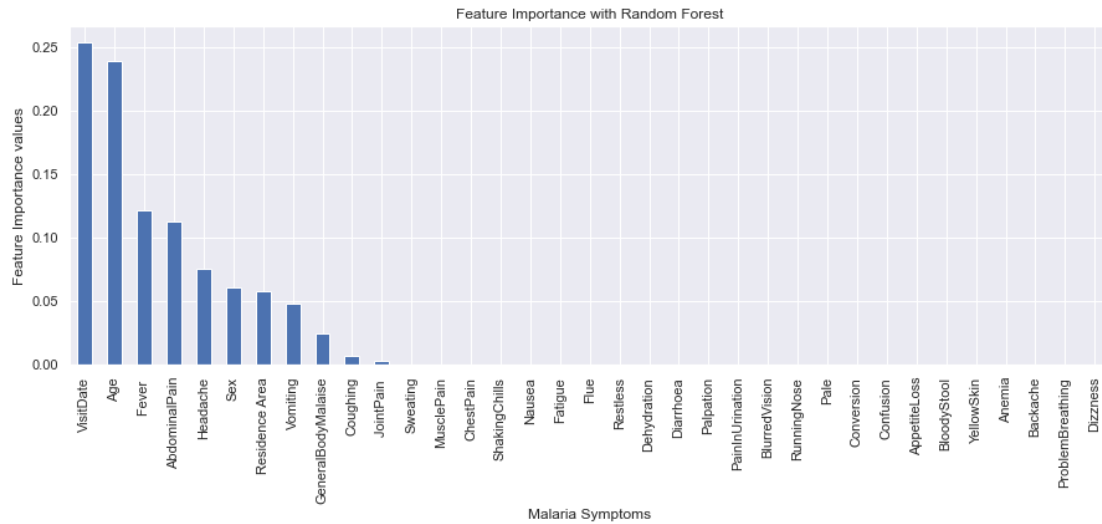


Figure 28: Important features in the two regions

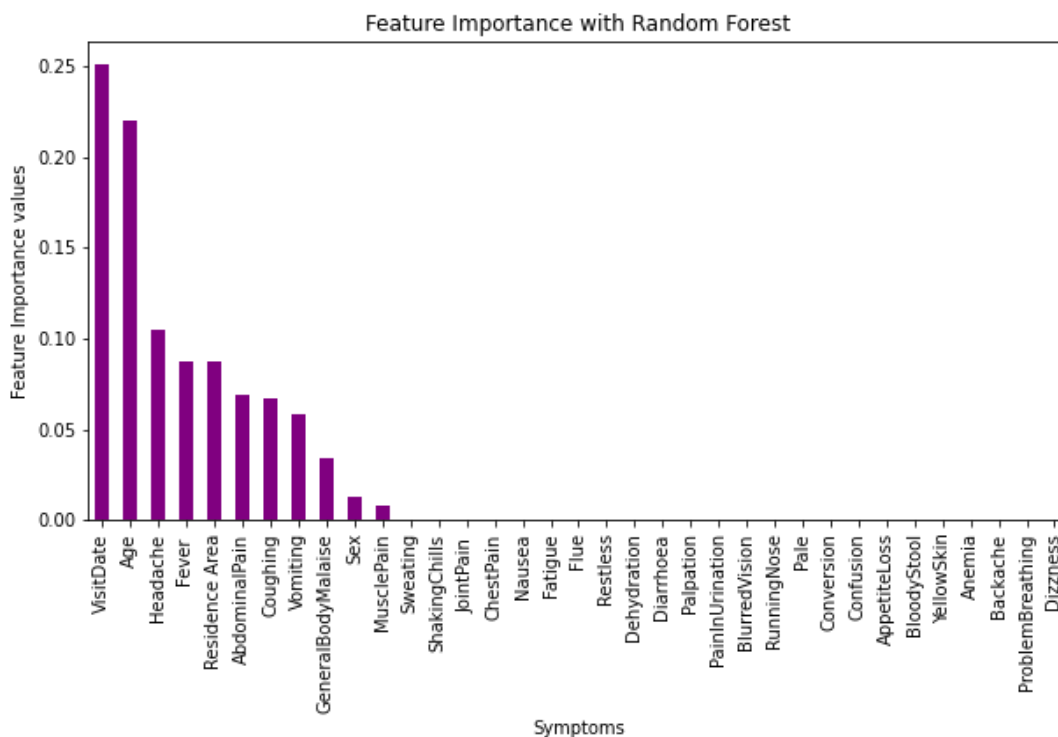


Figure 29: Important features in low endemic area

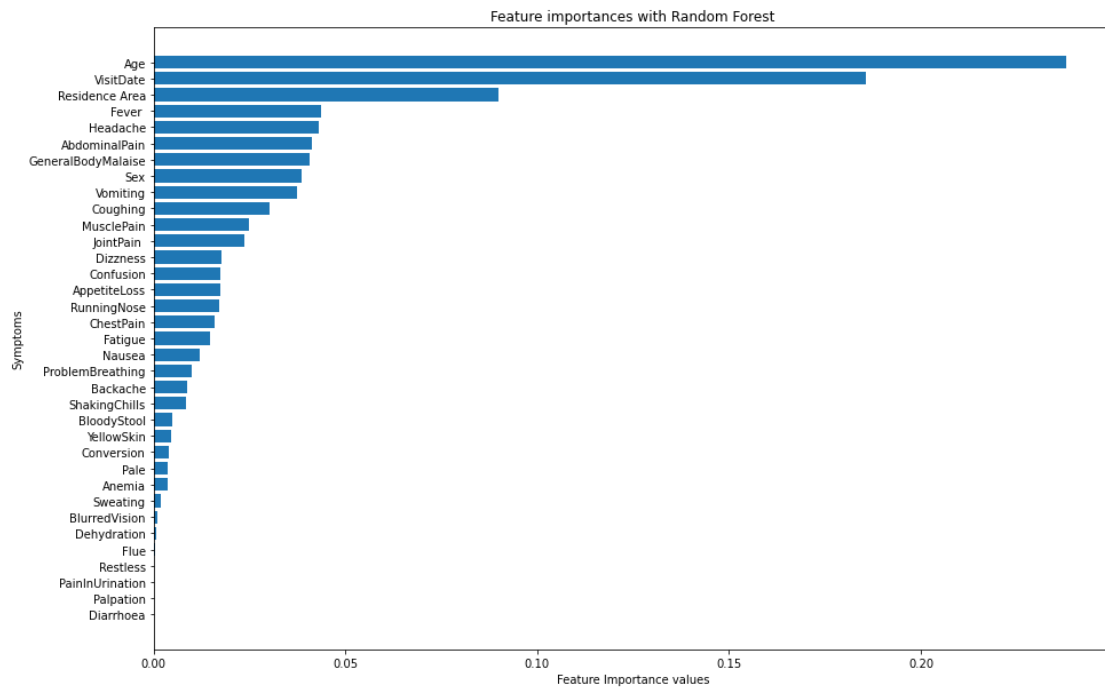


Figure 30: Important features in high endemic areas

It was observed that in the high endemic area age of the patient, visit date and the residence area of the patient had high significance in malaria diagnosis. Other important features are fever, headache, abdominal pain, general body malaise, gender of the patient and vomiting.

5.3.3 Model performance on validation dataset

The validation process used the best-performing classifiers which are Random Forest (RF) and Decision Tree (DT) to test and validate its robustness using the unseen malaria diagnosis dataset. Both Random Forest and Decision Tree classifiers performed well with the unseen dataset. Random Forest attained a prediction accuracy of 74%, Precision of 75%, Recall score of 84%, roc_auc of 83% and F1 score of 79%. Decision Tree scored a prediction accuracy of 72%, precision of 77%, recall of 77% roc_auc of 71% and f1 score of 76% as shown in Fig. 31 and Fig. 32.

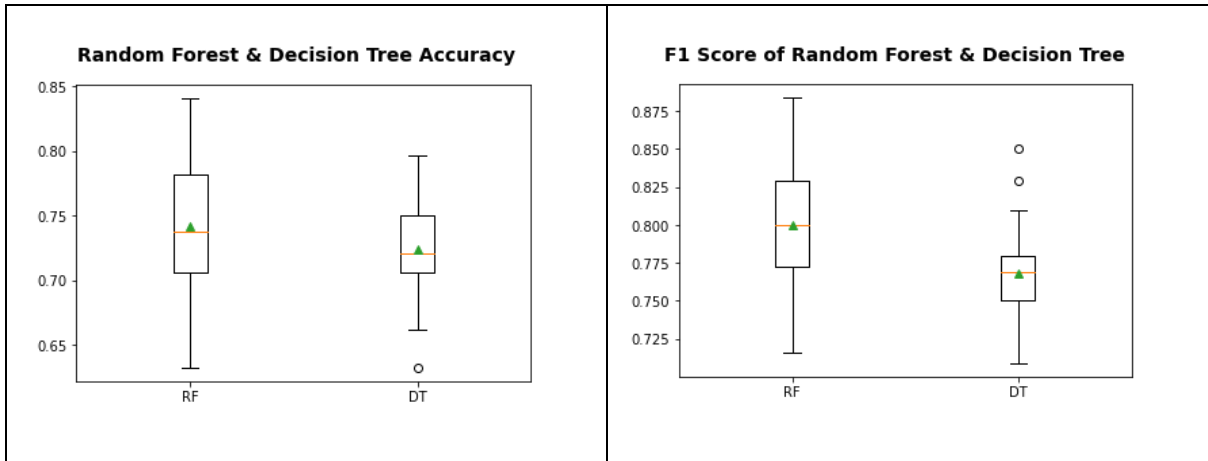


Figure 31: Prediction Accuracy and F1 score for the validation dataset

The classifiers performed well in predicting that the test results will be positive when the disease is present (true positive rate) by 93% and 91% for RF and DT respectively. Apart from that the probability that a test result will be negative when the disease is not present (true negative rate)

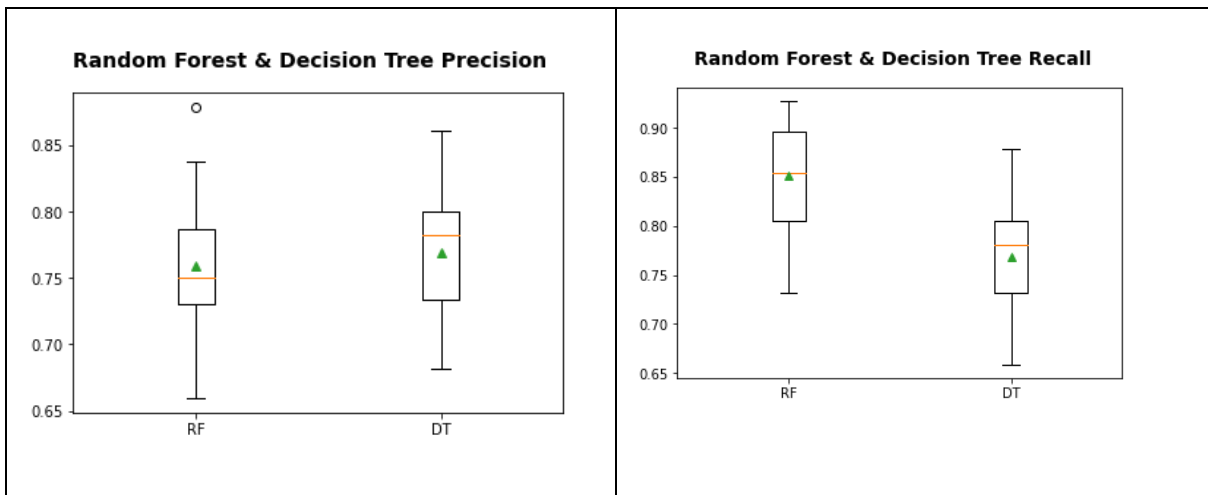


Figure 32: Precision and recall scores for the validation dataset

Generally, the models showed a good performance in classifying sick patients from healthy ones. DT had a 98% of true positive classification while RF has a 98.5% of true positive classification as depicted in Fig. 33.

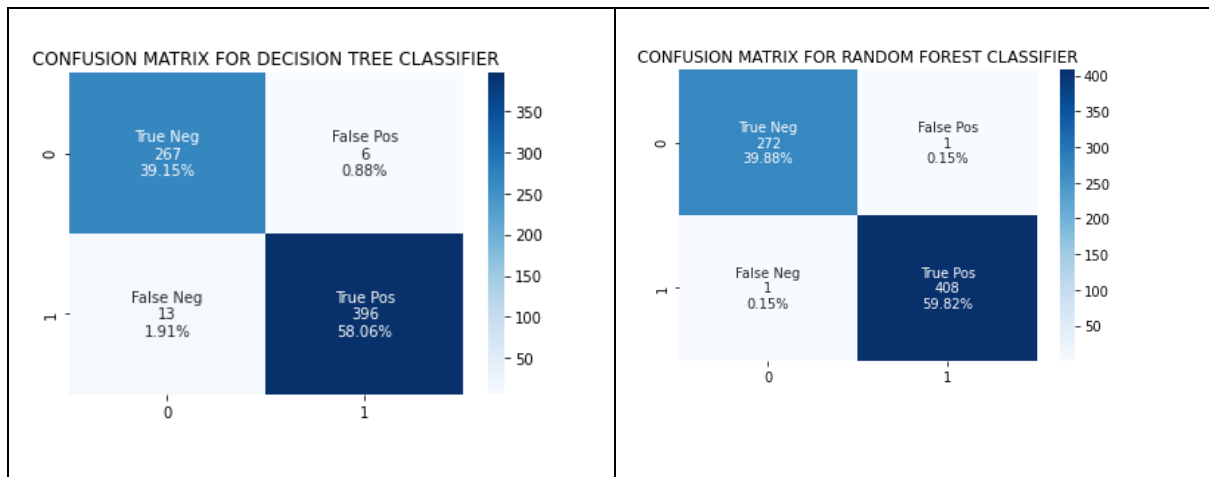


Figure 33: Confusion matrix of DT and RF on a validation dataset

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

In the previous chapters, the results of the study objective have been given, discussed, interpreted and compared with relevant literature. This chapter presents the concluding remarks from the research findings, study contribution, and future research directions. This study aimed to develop a malaria diagnosis machine learning model using patients' symptoms and demographic features in a resource-poor country like Tanzania, where self-medication and presumptive treatment are highly practised. Furthermore, the study provided scientific evidence of the important features or indicators in the malaria diagnosis drawn from the malaria diagnosis dataset, that malaria diagnosis features vary from place to place based on their level of endemicity. Therefore, the significance of one feature in one place is not the same in the other place. Additionally, this study provides insight that machine learning can diagnose malaria using the patients' clinical symptoms and demographic features. The significant findings, outcomes and recommendations of all three specific objectives of the study are summarised in this chapter.

6.2 Summary of the Study Findings

6.2.1 Characterisation of malaria diagnosis records

Characterisation of the malaria diagnosis dataset was the first objective of this study which aimed to investigate the state of malaria diagnosis records and explore malaria patient records to identify different variables that are significant in malaria diagnosis and create a malaria diagnosis dataset from malaria patient records that can be used in training, testing and validation of the machine learning-based malaria diagnosis model. Self-medication with anti-malaria drugs is significant when malaria-related symptoms are observed, especially in high-endemic areas. Malaria case management in Tanzania is typically insufficient due to a shortage of diagnostic equipment like microscopes and non-compliance with standard malaria treatment standards, even though more than 80% of the population lives within 5 kilometres of a health centre providing malaria treatment. When malaria-related symptoms are noticed, it is crucial to investigate the patient's travel history and whether or not there is a visitor in the home who recently returned from a high-endemic location. Finally, the malaria diagnosis datasets were

created from the curated malaria patients' records. The datasets included the Low endemic dataset (Kilimanjaro region), high endemic dataset (Morogoro region) and a combined dataset (Kilimanjaro region & Morogoro region).

6.2.2 Development of machine learning model for malaria diagnosis

The second objective intended to develop machine learning for malaria diagnosis using patients' symptoms and non-symptomatic features. To attain high prediction accuracy for this model the identification of features with predictive power was done. It was discovered that not only some of the symptoms have significance in the diagnosis of malaria but also non-symptoms such as the residential area of the patient, sex and age, have significance in the diagnosis of malaria. The difference in the level of importance of the malaria diagnosis features for different regions signifies that each region is unique even though they're in the same country, and their patients should not be observed the same. The identified malaria diagnosis criteria by WHO match the features identified in this study, proving that the model that will be developed will support the given malaria treatment guideline. The trained models attained the highest performance accuracy with the selected important features. Overall, the features' ranking differed among the regional datasets due to geographical location, which enhances the rate of disease transmission.

After selecting these important features for malaria diagnosis, the second objective demonstrated the success of using supervised learning models in diagnosing malaria using patient symptoms and demographic features. The machine learning classifiers in this study had different performances when exposed to the three classifiers. In the regional datasets, it was observed that Random Forest was the best classifier, with 95% prediction accuracy for low endemic datasets and with about 87% of prediction accuracy for a high endemic dataset. In the ability of a classifier to correctly identify patients with a disease, both Decision Trees and Random Forests classifiers attained an equal performance of 96% in low endemic areas. But in high endemic areas, only Random Forest classifier performance well with an accuracy of 85%.

In both datasets, Random Forest (RF) showed overall good performance compared to other classifiers. RF had a correct predictions rate of 79%, Precision of 71% and recall of 76%. Apart from that this model distinguish between the positive and negative patient by 80% and it has 82% of true positive rate. Furthermore, all models had a high performance with selected important features except the Logistic Regression recorded lower AUC and accuracy. In

important feature selection, Random Forest and Decision Tree confidently predict true negatives that 81% of the negative malaria prediction were healthy (with no malaria).

With the two high-performing machine learning classifiers (RF and DT) the malaria diagnosis model was developed. Decision Tree was used to depict the decision rules towards the classification of malaria patients and RF was used to improve the prediction accuracy since it uses many decision trees to predict. The results showed that only a specific combination of features can determine whether a patient has malaria rather than just one feature. In predicting malaria, the two models had a high performance with prediction accuracy of 96%, 99% and 98% for the Kilimanjaro, Combined and Morogoro datasets, respectively.

With these results the regional specific model for malaria diagnosis was developed and believed that patients who are not sick from malaria will be correctly identified and won't be prescribed with antimalaria drugs even though they have observed malaria related symptoms.

6.2.3 Validation of the developed machine learning model for malaria diagnosis

Lastly, the study validated the obtained results by subjecting the best performed models which are Random Forest and Decision Tree to an unseen malaria diagnosis dataset. Generally two models obtained high accuracy in predicting malaria. The models performance attained, proved that the algorithm used for modelling in this study were robust enough even to a new set of data. Apart from that medical doctors were interviewed to assess on the feasibility, performance, computational complexity, impact and awareness of using machine learning based model to improve malaria diagnosis in Tanzania. The important features obtained in this study align with the most significant symptoms and non symptoms identified by the medical doctors. This shows a great feasibility of using this model for malaria diagnosis.

6.3 Contributions of the Study

6.3.1 Scientific Contributions

Machine learning can relieve malaria mismanagement by providing a high-accuracy disease prediction tool that doesn't require expensive equipment or trained personnel. Although microscopic blood slides and rapid diagnostic tests are widely available, several challenges were identified, including self-medication and presumptive with antimalaria drugs and presumptive treatment of malaria. To get the best prediction accuracy of a model, applying

feature selection methods to the malaria diagnosis dataset is a key procedure to perform. To have a regional-specific malaria predictive model used in malaria diagnosis based on clinical symptoms and demographic data is essential since every region have a specific characteristics and for precise prediction of the disease.

6.3.2 New knowledge added from the study

The study has added value to the body of knowledge of machine learning, e-Health and ICT on dealing with proper disease management in resource-poor settings using ML techniques. While several studies have suggested that using clinical symptoms in prediction of malaria is not a practical idea, the experiments performed in this study proved the feasibility of using clinical symptoms and patients' demographic information to predict malaria using machine learning classifiers.

6.4 Recommendations

6.4.1 To the government and policymakers

Towards the efforts of reducing drug resistance, the results of this study can be used by the policymakers and the Ministry of Health for the better management of malaria by developing a simple tool that will be used to predict if a patient is sick before administering anti-malaria drugs. The tool can be very useful to health facilities that lack testing tools and drug dispensing outlets such as pharmacies and drug stores. Furthermore, this study's findings can inform the government of the potentiality of using machine learning in managing other diseases and predicting different health situations for early intervention. Also, the study can raise public awareness of significant malaria symptoms and patient features in diagnosing malaria at early stages within Tanzanian societies vulnerable to malaria and reduce the rate of self-medication and presumptive treatment in the country. Apart from that, the study recommends the establishment of an electronic patient records management system so that data curation and accessibility are easy and reliable.

6.4.2 To the practitioners

For the medical doctors this model will assist them on their diagnosis process, especially in the areas where laboratory confirmation is unavailable. In a clinical setting, this study demonstrated that clinicians can use the model to detect new malaria cases provided that

patients' symptoms and demographic features are available when the laboratory test is unavailable. The model developed can ensure that patients who seek treatment at facilities without the recommended equipment or personnel for parasitology tests get confirmation of having malaria or not before getting any treatment. Furthermore, for the drug dispensing outlets this model which will be more suitable for people who seeks for drugs but cannot access the laboratory-based diagnosis tools or access the health facility before any treatment. To the researchers this study paves a way to explore malaria diagnosis data available in the health facilities

6.5 Limitation of the Study

Results of this study, however, are subject to certain limitations. First, our sample is restricted to patients' records extracted from the patient's files in the selected health facilities. The additional potential limitation is that the developed models were based on the data obtained in the four health facilities in two regions. Therefore, we can not generalise our results to the entire country's population. The second potential limitation is that the developed models were based on the data collected in 2015-2019, where the recommendation given in this study may not reflect current practices. Apart from that the use of symptoms and no-symptomatic features is the symptoms depends on how th and that can be challenging if they are not reported correctly. Apart from that to the severity of the symptoms is another important aspect in distinguishing severe malaria which is not recorded well in the patients files. Despite these limitations and recommendations, the achieved outcomes remain significant to the study area and the collected malaria diagnosis dataset.

6.6 Future Research

Based on what has been done in this study, research on machine learning for malaria diagnosis can be extended to provide a broader understanding of dynamics related to malaria diagnosis and machine learning. More studies to include more regions and enlarge the dataset for improving the model's performance and inclusivity. The study developed a malaria prediction model using Random Forest and decision tree which are both tree-based models, in future studies could combine the two models or include more models to improve the performance of the prediction of malaria. Also in the future studies, results from this study will be necessary step in designing a malaria diagnosis decision support system through the developed model. Apart from that malaria diagnosis models need to be updated with the current malaria diagnosis

datasets since the state of the disease changes overtime such as the rate of endemicity within regions. The dataset should also include more fetures such as weather information. Furthermore future researches can extend to other diseases that are related to malaria and have high rate of self medication and presumptive treatment.

REFERENCES

- Abba, K., Deeks, J. J., Olliaro, P. L., Naing, M., Jackson, S. M., Takwoingi, Y., Donegan, S., & Garner, P. (2011). Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries. *Cochrane Database of Systematic Reviews*. <https://doi.org/10.1002/14651858.CD008122.PUB2>.
- Aikambe, J. N., & Mnyone, L. L. (2020). Retrospective analysis of malaria cases in a potentially high endemic area of Morogoro rural district, Eastern Tanzania. *Research and Reports in Tropical Medicine*, *11*, 37. <https://doi.org/10.2147/ RRTM.S254577>
- Alefan, Q., & Halboup, A. (2016). Pharmacy practice in Jordan. *Pharmacy Practice in Developing Countries: Achievements and Challenges*, 211–232. <https://doi.org/10.1016/ B978-0-12-801714-2.00011-3>.
- Alghanim, S. A. (2011). Self-medication practice among patients in a public health care system عام صحي نظام وجود ظل يف امريض لدى الذاتية املداواة ممارسة. *Eastern Mediterranean Health Journal*, *17*(5).
- Altaras, R., Nuwa, A., Agaba, B., Streat, E., Tibenderana, J. K., & Strachan, C. E. (2016). Why do health workers give anti-malarials to patients with negative rapid test results? A qualitative study at rural health facilities in western Uganda. *Malaria Journal*, *15*(1), 23. <https://doi.org/10.1186/S12936-015-1020-9>
- Andrade, B. B., Reis-Filho, A., Barros, A. M., Souza-Neto, S. M., Nogueira, L. L., Fukutani, K. F., Camargo, E. P., Camargo, L. M., Barral, A., Duarte, N., & Barral-Netto, M. (2010). Towards a precise test for malaria diagnosis in the Brazilian Amazon: Comparison among field microscopy, a rapid diagnostic test, nested PCR, and a computational expert system based on artificial neural networks. *Malaria Journal*, *9*(1), 1–11. <https://doi.org/10.1186/1475-2875-9-117/TABLES/6>.
- Ansari, M. (2018). Sociobehavioral Aspects of medicines use in developing countries. *Social and Administrative Aspects of Pharmacy in Low-and Middle-Income Countries: Present Challenges and Future Solutions*, 15–33. <https://doi.org/10.1016/B978-0-12-811228-1.00002-9>.

- Ansumana, R., Jacobsen, K. H., Gbakima, A. A., Hodges, M. H., Lamin, J. M., Leski, T. A., Malanoski, A. P., Lin, B., Bockarie, M. J., & Stenger, D. A. (2013). Presumptive self-diagnosis of malaria and other febrile illnesses in Sierra Leone. *The Pan African Medical Journal*, *15*. <https://doi.org/10.11604/PAMJ.2013.15.34.2291>.
- Attinsounon, C. A., Sissinto, Y., Avokpaho, E., Alassani, A., Sanni, M., & Zannou, M. (2019). Self-medication practice against malaria and associated factors in the city of Parakou in northern Benin: Results of a population survey in 2017. *Advances in Infectious Diseases*, *09*(03), 263–275. <https://doi.org/10.4236/aid.2019.93020>
- Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465–473. <https://doi.org/10.1016/J.CMPB.2013.11.004>
- Bali, A., Bali, D., Iyer, N., & Iyer, M. (2011). Management of medical records: facts and figures for surgeons. *Journal of Maxillofacial & Oral Surgery*, *10*(3), 199. <https://doi.org/10.1007/S12663-011-0219-8>
- Baltzell, K., Kortz, T. B., Scarr, E., Blair, A., Mguntha, A., Bandawe, G., Schell, E., & Rankin, S. (2019). 'Not All Fevers are Malaria': A mixed methods study of non-malarial fever management in rural Southern Malawi. *Rural and Remote Health*, *19*(2). <https://doi.org/10.22605/RRH4818>
- Barber, R. M., Fullman, N., Sorensen, R. J. D., Bollyky, T., McKee, M., Nolte, E., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., Abdurahman, A. A., Abera, S. F., Abraham, B., Abreha, G. F., Adane, K., Adelekan, A. L., Adetifa, I. M. O., ... Murray, C. J. L. (2017). Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: A novel analysis from the global burden of disease study 2015. *Lancet*, *390*(10091), 231–266. [https://doi.org/10.1016/s0140-6736\(17\)30818-8](https://doi.org/10.1016/s0140-6736(17)30818-8)
- Bbosa, F., Wesonga, R., & Jehopio, P. (2016). Clinical malaria diagnosis: rule-based classification statistical prototype. *SpringerPlus*, *5*(1), 1–14. <https://doi.org/10.1186/S40064-016-2628-0/TABLES/7>

- Belachew, G. G., Alemayehu, G. D., Fikadu, B. D., Hadgu, B. A., Ghezu, H. M., Solomon, H. G., Gebresamuel, A. N., Yarlagadda, R., Wondimu, D. A., & Abebe, K. Z. (2011). Self-medication practices among health sciences students: The case of Mekelle University. *Journal of Applied Pharmaceutical Science*, 2011(10), 183–189.
- Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms* 2023, 16(2), 88. <https://doi.org/10.3390/A16020088>
- Bibin, D., Nair, M. S., & Punitha, P. (2017). Malaria parasite detection from peripheral blood smear images using deep belief networks. *IEEE Access*, 5, 9099–9108. <https://doi.org/10.1109/ACCESS.2017.2705642>
- Boillat, B. N., Mbarack, Z., Samaka, J., Mlaganile, T., Kazimoto, T., Mamin, A., Genton, B., Kaiser, L., & D’Acremont, V. (2021). Causes of fever in Tanzanian adults attending outpatient clinics: A prospective cohort study. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 27(6), 913.e1–913.e7. <https://doi.org/10.1016/J.CMI.2020.08.031>
- Bria, Y. P., Yeh, C. H., & Bedingfield, S. (2021). Significant symptoms and non-symptom-related factors for malaria diagnosis in endemic regions of Indonesia. *International Journal of Infectious Diseases*, 103, 194–200. <https://doi.org/10.1016/j.ijid.2020.11.177>
- Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Tittgemeyer, M., Buhmann, J. M., Weber, B., & Stephan, K. E. (2011). Model-based feature construction for multivariate decoding. *Neuro Image*, 56(2), 601–615. <https://doi.org/10.1016/j.neuroimage.2010.04.036>
- Brown, B. J., Manescu, P., Przybylski, A. A., Caccioli, F., Oyinloye, G., Elmi, M., Shaw, M. J., Pawar, V., Claveau, R., Shawe-Taylor, J., Srinivasan, M. A., Afolabi, N. K., Rees, G., Orimadegun, A. E., Ajetunmobi, W. A., Akinkunmi, F., Kowobari, O., Osinusi, K., Akinbami, F. O., ... Fernandez-Reyes, D. (2020). Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-Saharan West Africa. *Scientific Reports*, 10(1). <https://doi.org/10.1038/S41598-020-72575-6>

- Budimu, A., Emidi, B., Mkumbaye, S., & Kajeguka, D. C. (2020). Adherence, awareness, access, and use of standard diagnosis and treatment guideline for malaria case management among healthcare workers in Meatu, Tanzania. *Journal of Tropical Medicine*, 2020. <https://doi.org/10.1155/2020/1918583>
- Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., Stenlund, H., Martens, P., & Lloyd, S. J. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9), 3286–3291. <https://doi.org/10.1073/pnas.1302089111>
- Capeding, M. R., Chua, M. N., Hadinegoro, S. R., Hussain, I. I. H. M., Nallusamy, R., Pitisuttithum, P., Rusmil, K., Thisyakorn, U., Thomas, S. J., Huu Tran, N., Wirawan, D. N., Yoon, I. K., Bouckennooghe, A., Hutagalung, Y., Laot, T., & Wartel, T. A. (2013). Dengue and other common causes of acute febrile illness in Asia: An active surveillance study in children. *PLoS Neglected Tropical Diseases*, 7(7), e2331. <https://doi.org/10.1371/journal.pntd.0002331>
- CDC. (2021). Malaria Worldwide - Impact of malaria. *Centres for Diseases Control and Prevention*. https://www.cdc.gov/malaria/malaria_worldwide/impact.html
- Chacko, S. (2021). Global malaria response suffered due to COVID-19: World Malaria Report 2021. <https://www.downtoearth.org.in/news/health/global-malaria-response-suffered-due-to-covid-19-world-malaria-report-2021-80585>
- Chandramohan, D., Jaffar, S., & Greenwood, B. (2002). Use of clinical algorithms for diagnosing malaria. *Tropical Medicine and International Health*, 7(1), 45–52. <https://doi.org/10.1046/j.1365-3156.2002.00827.x>
- Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, 2, 100016. <https://doi.org/10.1016/J.HEALTH.2022.100016>
- Chen, P. H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. *Nature Materials* 2019 18:5, 18(5), 410–414. <https://doi.org/10.1038/s41563-019-0345-0>

- Chipwaza, B., Mugasa, J. P., Mayumana, I., Amuri, M., Makungu, C., & Gwakisa, P. S. (2014). Self-medication with anti-malarials is a common practice in rural communities of Kilosa district in Tanzania despite the reported decline of malaria. *Malaria Journal*, *13*(1), 252. <https://doi.org/10.1186/1475-2875-13-252>
- Chirombo, J., Ceccato, P., Lowe, R., Terlouw, D. J., Thomson, M. C., Gumbo, A., Diggle, P. J., & Read, J. M. (2020). Childhood malaria case incidence in Malawi between 2004 and 2017: Spatio-temporal modelling of climate and non-climate factors. *Malaria Journal* *2020 19:1*, *19*(1), 1–13. <https://doi.org/10.1186/S12936-019-3097-Z>
- Clair, S., Pitt, B., Bakeera, S., McCall, N., Lukolyo, H., Arnold, D., Audcent, T., Batra, M., Chan, K., Jacquet, A., Schutze, E., & Butteris, S. (2017). Global health: Preparation for working in resource-limited settings. *Paediatrics*, *140*(5). <https://doi.org/10.1542/PEDS.2016-3783>
- Crump, J. A., Morrissey, A. B., Nicholson, W. L., Massung, R. F., Stoddard, R. A., Galloway, R. L., Ooi, E. E., Maro, V. P., Saganda, W., Kinabo, G. D., Muiruri, C., & Bartlett, J. A. (2013). Etiology of severe non-malaria febrile illness in Northern Tanzania: a prospective cohort study. *PLoS Neglected Tropical Diseases*, *7*(7), e2324. <https://doi.org/10.1371/journal.pntd.0002324>
- Crump, J. A., Newton, P. N., Baird, S. J., & Lubell, Y. (2017). Febrile illness in adolescents and adults. *Disease Control Priorities, Third Edition (Volume 6): Major Infectious Diseases*, 365–385. https://doi.org/10.1596/978-1-4648-0524-0_CH14
- D'Acremont, V., Kilowoko, M., Kyungu, E., Philipina, S., Sangu, W., Kahama-Maró, J., Lengeler, C., Cherpillod, P., Kaiser, L., & Genton, B. (2014). Beyond malaria causes of fever in outpatient Tanzanian children. *New England Journal of Medicine*, *370*(9), 809–817. <https://doi.org/10.1056/nejmoa1214482>
- D'Acremont, V., Lengeler, C., & Genton, B. (2010). Reduction in the proportion of fevers associated with *Plasmodium falciparum* parasitaemia in Africa: A systematic review. *Malaria Journal*, *9*(1). <https://doi.org/10.1186/1475-2875-9-240>

- D'Acremont, V., Lengeler, C., Mshinda, H., Mtasiwa, D., Tanner, M., & Genton, B. (2009). Time to move from presumptive malaria treatment to laboratory-confirmed diagnosis and treatment in African children with fever. *PLoS Med.*, *6*(1), e252. <https://doi.org/10.1371/journal.pmed.0050252>
- Das, D. K., Ghosh, M., Pal, M., Maiti, A. K., & Chakraborty, C. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, *45*, 97–106. <https://doi.org/10.1016/J.MICRON.2012.11.002>
- Dash, S. S., Nayak, S. K., & Mishra, D. (2021). A review on machine learning algorithms. *Smart Innovation, Systems and Technologies*, *153*, 495–507. https://doi.org/10.1007/978-981-15-6202-0_51
- Datatron. (2022). What is model validation and why is it important? <https://datatron.com/what-is-model-validation-and-why-is-it-important/>
- Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. [/pmc/articles/PMC6616181/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/31222121/)
- de Santis, O., Kilowoko, M., Kyungu, E., Sangu, W., Cherpillod, P., Kaiser, L., Genton, B., & D'Acremont, V. (2017). Predictive value of clinical and laboratory features for the main febrile diseases in children living in Tanzania: A prospective observational study. *PLoS ONE*, *12*(5). <https://doi.org/10.1371/journal.pone.0173314>
- Deepthi, Y., Kalyan, K. P., Vyas, M., Radhika, K., Babu, D. K., & Krishna Rao, N. V. (2020). Disease prediction based on symptoms using machine learning. *Lecture Notes in Electrical Engineering*, *664*, 561–569. https://doi.org/10.1007/978-981-15-5089-8_55
- Dharap, P., & Raimbault, S. (2020). Performance evaluation of machine learning-based infectious screening flags on the HORIBA Medical Yumizen H550 Haematology Analyzer for vivax malaria and dengue fever. *Malaria Journal* *2020 19:1*, *19*(1), 1–10. <https://doi.org/10.1186/S12936-020-03502-3>

- Dhiman, S. (2019). Are malaria elimination efforts on right track? An analysis of gains achieved and challenges ahead. *Infectious Diseases of Poverty* 8:1, 8(1), 1–19. <https://doi.org/10.1186/S40249-019-0524-X>
- Dwyer, B.D., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, 14, 91–118. <https://doi.org/10.1146/ANNUREV-CLINPSY-032816-045037>
- Faria, J. (2022). Tanzania: Number of malaria cases Statista <https://www.statista.com/statistics/1239926/number-of-malaria-cases-in-Tanzania/>
- Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 09(01), 1–16.
- Femi, E., Onyebuchi, E., & Shehi, I. (2016). A predictive symptoms-based system using support vector machines to enhanced classification accuracy of malaria and typhoid coinfection. *International Journal of Mathematical Sciences and Computing*, 4, 54–66. <https://doi.org/10.5815/ijmsc.2016.04.06>
- Fernandez, C., Hervella, P., Mato, V., Rodríguez, M., Suárez, S., López, I., Estany, A., Sobrino, T., Campos, F., Castillo, J., Rodríguez, S., & Iglesias, R. (2021). Random forest-based prediction of stroke outcome. *Scientific Reports* 2021 11:1, 11(1), 1–12. <https://doi.org/10.1038/s41598-021-89434-7>
- Finda, M. F., Christofides, N., Lezaun, J., Tarimo, B., Chaki, P., Kelly, A. H., Kapologwe, N., Kazyoba, P., Emidi, B., & Okumu, F. O. (2020). Opinions of key stakeholders on alternative interventions for malaria control and elimination in Tanzania. *Malaria Journal*, 19(1), 1–13. <https://doi.org/10.1186/S12936-020-03239-Z/TABLES/2>
- Ford, C. T., Alemayehu, G., Blackburn, K., Lopez, K., Dieng, C. C., Lo, E., Golassa, L., & Janies, D. (2020). Modeling plasmodium falciparum diagnostic test sensitivity using machine learning with histidine-rich protein 2 variants. *MedRxiv*, 2020.05.27.20114785. <https://doi.org/10.1101/2020.05.27.20114785>
- Frøkjær, B., Bolvig, T., Griese, N., Herborg, H., & Rossing, C. (2012). Prevalence of drug-related problems in self-medication in Danish community pharmacies. *Innovations* 3(4), <https://pdfs.semanticscholar.org/46fb/216e1eaff1d1f7cc>

4cd5e36688ac4e191f6f.pdf?_ga=2.178764772.1138905733.158653354018158538
36.158 2188226

- Fuhad, K. M. F., Tuba, J. F., Sarker, M. R. A., Momen, S., Mohammed, N., & Rahman, T. (2020b). Deep learning based automatic malaria parasite detection from blood smear and its smartphone-based application. *Diagnostics*, *10*(5). <https://doi.org/10.3390/diagnostics10050329>
- Gallay, J., Mosha, D., Lutahakana, E., Mazuguni, F., Zuakulu, M., Decosterd, L. A., Genton, B., & Pothin, E. (2018). Appropriateness of malaria diagnosis and treatment for fever episodes according to patient history and anti-malarial blood measurement: A cross-sectional survey from Tanzania. *Malaria Journal*, *17*(1), 1–13. <https://doi.org/10.1186/S12936-018-2357-7/TABLES/2>
- Ganiger, S., & Rajashekharaiyah, K. M. M. (2018). Chronic diseases diagnosis using machine learning. *2018 International Conference on Circuits and Systems in Digital Enterprise Technology, ICCSDET 2018*. <https://doi.org/10.1109/ICCSDET.2018.8821235>
- Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012046. <https://doi.org/10.1088/1757899X/1022/1/012046>
- Geiling, J., Burkle, F. M., Jr, Amundson, D., Dominguez-Cherit, G., Gomersall, C. D., Lim, M. L., Luyckx, V., Sarani, B., Uyeki, T. M., West, T. E., Christian, M. D., Devereaux, A. v., Dichter, J. R., & Kissoon, N. (2014). Resource-poor settings: Infrastructure and capacity building: Care of the critically ill and injured during pandemics and disasters: Chest consensus statement. *Chest*, *146*(4), e156S. <https://doi.org/10.1378/CHEST.14-0744>
- Geldof, T., van Damme, N., Huys, I., & van Dyck, W. (2020). Patient-level effectiveness prediction modelling for glioblastoma using classification trees. *Frontiers in Pharmacology*, *10*. <https://doi.org/10.3389/FPHAR.2019.01665/FULL>

- Ghumbre, S. U., & Ghatol, A. A. (2012). Heart disease diagnosis using machine learning algorithm. *Advances in Intelligent and Soft Computing, 132 AISC*, 217–225. https://doi.org/10.1007/978-3-642-27443-5_25
- Gillet, P., Scheirlinck, A., Stokx, J., de Weggheleire, A., Chaúque, H. S., Canhanga, O. D., Tadeu, B. T., Mosse, C. D., Tiago, A., Mabunda, S., Bruggeman, C., Bottieau, E., & Jacobs, J. (2011). Prozone in malaria rapid diagnostics tests: How many cases are missed? *Malaria Journal, 10*(5). <https://doi.org/10.1186/1475-2875-10-166>
- Gil-Rivas, V., & McWhorter, L. (2013). Self-medication. *Principles of Addiction*, 235–241. <https://doi.org/10.1016/B978-0-12-398336-7.00024-3>
- Goodyer, L. (2015). Dengue fever and chikungunya: Identification in travellers. *Clinical Pharmacist, 7*(4). <https://doi.org/10.1211/cp.2015.20068429>
- Gosling, R. D., Drakeley, C. J., Mwitwa, A., & Chandramohan, D. (2008). Presumptive treatment of fever cases as malaria: Help or hindrance for malaria control? *Malaria Journal 7*(1), 132. BioMed Central. <https://doi.org/10.1186/1475-2875-7-132>
- Graber, M. L., Byrne, C., & Johnston, D. (2017). The impact of electronic health records on diagnosis. *Diagnosis, 4*(4), 211–223. <https://doi.org/10.1515/dx-2017-0012>
- Graz, B., Willcox, M., Szeless, T., & Rougemont, A. (2011). Test and treat or presumptive treatment for malaria in high transmission situations? A reflection on the latest WHO guidelines. *Malaria Journal, 10*(1), 1–8. <https://doi.org/10.1186/1475-2875-10-136>
- Grobusch, M. P., & Schlagenhauf, P. (2019). 16 – Self-Diagnosis and Self-Treatment of Malaria by the Traveler. In *Travel Medicine*, 169–178. <https://doi.org/10.1016/B978-0-323-54696-6.00016-1>
- Group, A., Michael, D., & Mkunde, S. P. (2017). Malaria Journal The malaria testing and treatment landscape in mainland Tanzania, 2016. *Malar Journal, 16*, 202. <https://doi.org/10.1186/s12936-017-1819-7>
- Habib, P. T., Alsamman, A. M., Hassnein, S. E., Shereif, G. A., & Hamwieh, A. (2018). *Assessment of Machine Learning Algorithms for Prediction of Breast Cancer*

Malignancy Based on Mammogram Numeric Data. <https://doi.org/10.1101/2020.01.08.20016949>

- Hagenlocher, M., & Castro, M. (2015). Mapping malaria risk and vulnerability in the United Republic of Tanzania: A spatial explicit model. *Population Health Metrics*, 13(1), 2. <https://doi.org/10.1186/s12963-015-00362>
- Haji Ali Afzali, H., & Karnon, J. (2014). Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies. *Encyclopedia of Health Economics*, 340–347. <https://doi.org/10.1016/B978-0-12-375678-7.01401-2>
- Han, J., Kamber, M., & Pei, J. (2012). Data pre-processing. *Data Mining*, 83–124. <https://doi.org/10.1016/B978-0-12-381479-1.00003-4>
- Harvey, D., Valkenburg, W., & Amara, A. (2021). Predicting malaria epidemics in Burkina Faso with machine learning. *PLoS ONE*, 16(6). <https://doi.org/10.1371/JOURNAL.PONE.0253302>
- Hawkes, M., & Kain, K. C. (2014). Advances in malaria diagnosis. *Expert Review of Anti-Infective Therapy*, 5(3), 485–495. <https://doi.org/10.1586/14787210.5.3.485>
- Hemingway, J., Shretta, R., Wells, T. N. C., Bell, D., Djimdé, A. A., Achee, N., & Qi, G. (2016). Tools and strategies for malaria control and elimination: What do we need to achieve a grand convergence in malaria? *PLoS Biol*, 14(3). <https://doi.org/10.1371/journal.pbio.1002380>
- Hertz, J. T., Madut, D. B., Tesha, R. A., William, G., Simmons, R. A., Galson, S. W., Maro, V. P., Crump, J. A., & Rubach, M. P. (2019). Self-medication with non-prescribed pharmaceutical agents in an area of low malaria transmission in northern Tanzania: A community-based survey. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 113(4), 183–188. <https://doi.org/10.1093/trstmh/try138>
- Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., & Malik, M. H. (2021). Chronic kidney disease diagnosis using decision tree algorithms. *BMC Nephrology*, 22(1), 1–11. <https://doi.org/10.1186/S12882-021-02474-Z/FIGURES/5>

- Imran, A., Amin, M. N., & Johora, F. T. (2019). Classification of chronic kidney disease using logistic regression, feedforward neural network and wide deep learning. *2018 International Conference on Innovation in Engineering and Technology, ICIET 2018*. <https://doi.org/10.1109/CIET.2018.8660844>
- Isiguzo, C., Anyanti, J., Ujuju, C., Nwokolo, E., de La Cruz, A., Schatzkin, E., Modrek, S., Montagu, D., & Liu, J. (2014). Presumptive treatment of malaria from formal and informal drug vendors in Nigeria. *PLOS ONE*, *9*(10), e110361. <https://doi.org/10.1371/JOURNAL.PONE.0110361>
- Isiguzo, C., Anyanti, J., Ujuju, C., Nwokolo, E., De La Cruz, A., Schatzkin, E., Modrek, S., Montagu, D., & Liu, J. (2014). Presumptive treatment of Malaria from formal and informal drug vendors in Nigeria. *PLoS ONE*, *9*(10). <https://doi.org/10.1371/journal.pone.0110361>
- Iyer, A., S, J., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process*, *5*(1), 01–14. <http://www.airconline.com/ijdkp/V5N1/5115ijdkp01.pdf>
- Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal* *19*(3), 179–189. Elsevier B.V. <https://doi.org/10.1016/j.eij.2018.03.002>
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. <https://doi.org/10.1136/SVN-2017-000101>
- Jović, A., Brkić, K., & Bogunović, N. (2020). *A review of feature selection methods with applications*.
- Kahama, J., D'Acremont, V., Mtasiwa, D., Genton, B., & Lengeler, C. (2011). Low quality of routine microscopy for malaria at different levels of the health system in Dar es Salaam. *Malaria Journal* *2011* *10*:1, *10*(1), 1–10. <https://doi.org/10.1186/1475-2875-10-332>
- Kajungu, D. K., Selemani, M., Masanja, I., Baraka, A., Njozi, M., Khatib, R., Dodoo, A. N., Binka, F., Macq, J., D'Alessandro, U., & Speybroeck, N. (2012). Using

classification tree modelling to investigate drug prescription practices at health facilities in rural Tanzania. *Malaria Journal*, 11(1), 1–11. <https://doi.org/10.1186/1475-2875-11-311/FIGURES/2>

Kajeguka, D., & Moses, E. (2017). Self-medication practices and predictors for self-medication with antibiotics and antimalarials among community in Mbeya City, Tanzania. *Tanzania Journal of Health Research*, 19(4). <https://doi.org/10.4314/THRB.V19I4>

Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. *Computational and Mathematical Methods in Medicine*, 2022. <https://doi.org/10.1155/2022/6517716>

Kasturiwale, H., Karhe, R., & Kale, S. N. (2022). Machine learning approach for medical diagnosis based on prediction model. *Medical Imaging and Health Informatics*, 1–21. <https://doi.org/10.1002/9781119819165.CH1>

Katz, E., & Hartley, A. (2020). Gender and Malaria Evidence Review. https://www.gatesgenderequalitytoolbox.org/wpcontent/uploads/BMGF_Malaria-Review_FC.pdf

Kazaura, M. R. (2017). Level and correlates of self-medication among adults in a rural setting of mainland Tanzania. *Indian Journal of Pharmaceutical Sciences*, 79(3), 451–457. <https://doi.org/10.4172/pharmaceutical-sciences.1000248>

Khare, A., Jeon, M., Sethi, I. K., & Xu, B. (2017). Machine learning theory and applications for healthcare. In *Journal of Healthcare Engineering*, 2017. Hindawi Limited. <https://doi.org/10.1155/2017/5263570>

Kim, J. K., Choo, Y. J., & Chang, M. C. (2021). Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models. *Journal of Stroke and Cerebrovascular Diseases*, 30(8). <https://doi.org/10.1016/J.JSTROKECEREBROVASDIS.2021.105856>

Kim, Y., Ratnam, J. v., Doi, T., Morioka, Y., Behera, S., Tsuzuki, A., Minakawa, N., Sweijd, N., Kruger, P., Maharaj, R., Imai, C. C., Ng, C. F. S., Chung, Y., &

- Hashizume, M. (2019). Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model. *Scientific Reports*, 9(1), 1–10. <https://doi.org/10.1038/s41598-019-53838-3>
- Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 26(9), 1011. <https://doi.org/10.1038/NBT0908-1011>
- Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019). Prediction of coronary heart disease using supervised machine learning algorithms. *IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2019-October*, 367–372. <https://doi.org/10.1109/TENCON.2019.8929434>
- Krumholz, H. M., Currie, P. M., Riegel, B., Phillips, C. O., Peterson, E. D., Smith, R., Yancy, C. W., & Faxon, D. P. (2006). A taxonomy for disease management: A scientific statement from the American Heart Association Disease Management Taxonomy Writing Group. *Circulation*, 114(13), 1432–1445. <https://doi.org/10.1161/CIRCULATIONAHA.106.177322>
- Kumar, N., Narayan Das, N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient automated disease diagnosis using machine learning models. *Journal of Healthcare Engineering*, 2021. <https://doi.org/10.1155/2021/9983652>
- Laghmati, S., Tmiri, A., & Cherradi, B. (2019, October 1). Machine learning based system for prediction of breast cancer severity. *Proceedings - 2019 International Conference on Wireless Networks and Mobile Communications, WINCOM 2019*. <https://doi.org/10.1109/WINCOM47513.2019.8942575>
- Landier, J., Parker, D. M., Thu, A. M., Carrara, V. I., Lwin, K. M., Bonnington, C. A., Pukrittayakamee, S., Delmas, G., & Nosten, F. H. (2016). The role of early detection and treatment in malaria elimination. *Malaria Journal* 15(1), 1–8. <https://doi.org/10.1186/s12936-016-1399-y>
- Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine*, 129, 104151. <https://doi.org/10.1016/J.COMPBIOMED.2020.104151>

- Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., Liu, Y., Zhang, Q., Cao, X., Chen, Q., & Niu, B. (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), 935–946. <https://doi.org/10.1111/TBED.13424>
- Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M. A., Sameer, A., Maude, R. J., Huang, J. X., Jaeger, S., & Thoma, G. (2017). CNN-based image analysis for malaria diagnosis. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, 493–496. <https://doi.org/10.1109/BIBM.2016.7822567>
- Lozano, R., Fullman, N., Mumford, J. E., Knight, M., Barthelemy, C. M., Abbafati, C., Abbastabar, H., Abd-Allah, F., Abdollahi, M., Abedi, A., Abolhassani, H., Abosetugn, A. E., Abreu, L. G., Abrigo, M. R. M., Abu Haimed, A. K., Abushouk, A. I., Adabi, M., Adebayo, O. M., Adekanmbi, V., & Murray, C. J. L. (2020). Measuring universal health coverage based on an index of effective coverage of health services in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), 1250–1284. [https://doi.org/10.1016/S0140-6736\(20\)30750-9](https://doi.org/10.1016/S0140-6736(20)30750-9)
- Ma, V., & Karki, M. V. (2020). Skin cancer detection using machine learning techniques. *Proceedings of CONECCT 2020 - 6th IEEE International Conference on Electronics, Computing and Communication Technologies*. <https://doi.org/10.1109/CONECCT50063.2020.9198489>
- Madhu, G. (2020). Computer Vision and Machine Learning Approach for Malaria Diagnosis in Thin Blood Smears from Microscopic Blood Images. *Springer, Singapore*. (191 - 209) https://doi.org/10.1007/978-981-15-3689-2_8
- Marealle, A. I., & Kirutu, P. (2018). Self-medication with anti-malarial medicines among high school students in Dar es salaam, Tanzania. *International Journal of Pharmacy and Pharmaceutical Sciences*, 10(8), 101. <https://doi.org/10.22159/ijpps.2018v10i8.17058>
- Masud, M., Alhumyani, H., Alshamrani, S. S., Cheikhrouhou, O., Ibrahim, S., Muhammad, G., Hossain, M. S., & Shorfuzzaman, M. (2020). Leveraging deep

- learning techniques for malaria parasite detection using mobile application. *Wireless Communications and Mobile Computing*. <https://doi.org/10.1155/2020/8895429>
- Matowo, N. S., Munhenga, G., Tanner, M., Coetzee, M., Feringa, W. F., Ngowo, H. S., Koekemoer, L. L., & Okumu, F. O. (2017). Fine-scale spatial and temporal heterogeneities in insecticide resistance profiles of the malaria vector, *Anopheles arabiensis* in rural south-eastern Tanzania. *Wellcome Open Res*, 2, 96. <https://doi.org/10.12688/wellcomeopenres.12617.1>
- Mboera, L. E. G., Makundi, E. A., & Kitua, A. Y. (2007). Uncertainty in malaria control in Tanzania: Crossroads and challenges for future interventions. <https://www.ncbi.nlm.nih.gov/books/NBK1714/>
- Menard, D., & Dondorp, A. (2017). Antimalarial drug resistance: A threat to malaria elimination. *Cold Spring Harbour Perspectives in Medicine*, 7(7), 1–24. <https://doi.org/10.1101/cshperspect.a025619>
- Metta, E., Haisma, H., Kessy, F., Hutter, I., & Bailey, A. (2014). “We have become doctors for ourselves”: Motives for malaria self-care among adults in south-eastern Tanzania. *Malaria Journal*, 13(1), 249. <https://doi.org/10.1186/1475-2875-13-249>
- Michael, D., & Mkunde, S. P. (2017). The malaria testing and treatment landscape in mainland Tanzania, 2016. *Malaria Journal* 2017 16:1, 16(1), 1–15. <https://doi.org/10.1186/S12936-017-1819-7>
- Mishra, V., Singh, Y., & Kumar Rath, S. (2019). Breast cancer detection from thermograms using feature extraction and machine learning techniques. *IEEE 5th International Conference for Convergence in Technology, I2CT 2019*. <https://doi.org/10.1109/I2CT45611.2019.9033713>
- Mlacha, Y. P., Wang, D., Chaki, P. P., Gavana, T., Zhou, Z., Michael, M. G., Khatib, R., Chila, G., Msuya, H. M., Chaki, E., Makungu, C., Lin, K., Tambo, E., Rumisha, S. F., Mkude, S., Mahende, M. K., Chacky, F., Vounatsou, P., Tanner, M., ... Zhou, X. N. (2020). Effectiveness of the innovative 1,7-malaria reactive community-based testing and response (1, 7-mRCTR) approach on malaria burden reduction in South-

eastern Tanzania. *Malaria Journal*, 19(1), 1–12. <https://doi.org/10.1186/S12936-020-03363-W/TABLES/4>

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542–81554. <https://doi.org/10.1109/ACCESS.2019.2923707>

Mohammed, A. (2020). What is Cross-Validation? Testing your machine learning models. *Towards Data Science*. <https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75>

Morang'a, C. M., Amenga-Etego, L., Bah, S. Y., Appiah, V., Amuzu, D. S. Y., Amoako, N., Abugri, J., Oduro, A. R., Cunningham, A. J., Awandare, G. A., & Otto, T. D. (2020). Machine learning approaches classify clinical malaria outcomes based on haematological parameters. *BMC Medicine* 2020 18:1, 18(1), 1–16. <https://doi.org/10.1186/S12916-020-01823-3>

Ibarra, M., Villuendas Rey, Y., Lytras, M. D., Yáñez-Márquez, C., & Salgado-Ramírez, J.C. (2021). Classification of Diseases Using Machine Learning Algorithms: A Comparative Study. *Mathematics* 2021, 9(15), 1817. <https://doi.org/10.3390/MATH9151817>

Mouatcho, J. C., Goldring, J. P. D., & JC Mouatcho, J. D. G. (2013). *Malaria rapid diagnostic tests: challenges and prospects*. 62, 1491–1505. <https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.052506-0>

Mpapalika¹, J. J., & Matowo², N. (2020). The application of Artificial Intelligence in the diagnosis and treatment of malaria in Tanzania. *Journal of Infectious Diseases Diagnosis*, <https://www.mdpi>.

Manisha, S.S., Ferme, E., & Camara, J. (2020). Machine learning for brain stroke: A review. *Journal of Stroke and Cerebrovascular Diseases: The Official Journal of National Stroke Association*, 29(10). <https://doi.org/10.1016/J.JSTROKECEREBROVASDIS.2020.105162>

- Muthumbi, A., Chaware, A., Kim, K., Zhou, K. C., Konda, P. C., Chen, R., Judkewitz, B., Erdmann, A., Kappes, B., & Horstmeyer, R. (2019). Learned sensing: Jointly optimized microscope hardware for accurate image classification. *Biomedical Optics Express*, *10*(12), 6351. <https://doi.org/10.1364/boe.10.006351>
- Mwai, L., Ochong, E., Abdirahman, A., Kiara, S. M., Ward, S., Kokwaro, G., Sasi, P., Marsh, K., Borrmann, S., MacKinnon, M., & Nzila, A. (2009). Chloroquine resistance before and after its withdrawal in Kenya. *Malaria Journal*, *8*(1), 106. <https://doi.org/10.1186/1475-2875-8-106>
- Mwanga, E. P., Mapua, S. A., Siria, D. J., Ngowo, H. S., Nangacha, F., Mgando, J., Baldini, F., González Jiménez, M., Ferguson, H. M., Wynne, K., Selvaraj, P., Babayan, S. A., & Okumu, F. O. (2019). Using mid-infrared spectroscopy and supervised machine-learning to identify vertebrate blood meals in the malaria vector, *Anopheles arabinosus*. *Malaria Journal*, *18*(1). <https://doi.org/10.1186/s12936-019-2822-y>
- Mwanga, E. P., Minja, E. G., Mrimi, E., Jiménez, M. G., Swai, J. K., Abbasi, S., Ngowo, H. S., Siria, D. J., Mapua, S., Stica, C., Maia, M. F., Olotu, A., Sikulu-Lord, M. T., Baldini, F., Ferguson, H. M., Wynne, K., Selvaraj, P., Babayan, S. A., & Okumu, F. O. (2019). Detection of malaria parasites in dried human blood spots using mid-infrared spectroscopy and logistic regression analysis. *Malaria Journal*, *18*(1). <https://doi.org/10.1186/S12936-019-2982-9>
- Mwita, S., Meja, O., Katabalo, D., & Richard, C. (2019). Magnitude and factors associated with anti-malarial self-medication practice among residents of Kasulu Town Council, Kigoma-Tanzania. *African Health Sciences*, *19*(3), 2457–2461. <https://doi.org/10.4314/ahs.v19i3.20>
- Nadjm, B., Amos, B., Mtove, G., Ostermann, J., Chonya, S., Wangai, H., Kimera, J., Msuya, W., Mtei, F., Dekker, D., Malahiyo, R., Olomi, R., Crump, J. A., Whitty, C. J. M., & Reyburn, H. (2010). WHO guidelines for antimicrobial treatment in children admitted to hospital in an area of intense *Plasmodium falciparum* transmission: Prospective study. *BMJ (Online)*, *340*(7751), 848. <https://doi.org/10.1136/BMJ.C1350>

- Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, 2022. <https://doi.org/10.1155/2022/7351061>
- Nandal, N., Goel, L., & TANWAR, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000Research* 2022 11:1126, 11, 1126. <https://doi.org/10.12688/f1000research.123776.1>
- Navdeep Singh Gill. (2022). Machine learning Model Validation Testing | A Quick Guide. <https://www.xenonstack.com/insights/what-is-model-validation-testing>
- Ndomondo, M., Mbwasi, R., Shirima, R., Heltzer, N., & Clark, M. (2005). Accredited Drug Dispensing Outlets: Improving Access to Quality Drugs and Services in Rural and Periurban Areas with Few or No Pharmacies.
- Ngasala, B., & Bushukatale, S. (2019). Evaluation of malaria microscopy diagnostic performance at private health facilities in Tanzania. *Malaria Journal* 2019 18:1, 18(1), 1–7. <https://doi.org/10.1186/S12936-019-2998-1>
- Ngasala, B., Mubi, M., Warsame, M., Petzold, M. G., Masele, A. Y., Gustafsson, L. L., Tomson, G., Premji, Z., & Bjorkman, A. (2008). Impact of training in clinical and microscopy diagnosis of childhood malaria on antimalarial drug prescription and health outcome at primary health care level in Tanzania: A randomized controlled trial. *Malaria Journal*, 7, 199. <https://doi.org/10.1186/1475-2875-7-199>
- Nithya, A., Appathurai, A., Venkatadri, N., Ramji, D. R., & Anna Palagan, C. (2020). Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. *Measurement*, 149, 106952. <https://doi.org/10.1016/J.MEASUREMENT.2019.106952>
- Nkumama, I. N., O'Meara, W. P., & Osier, F. H. A. (2017). Changes in Malaria Epidemiology in Africa and New Challenges for Elimination. *Trends in Parasitology* 33(2), 128–140. <https://doi.org/10.1016/j.pt.2016.11.006>
- Nsagha, D. S., Njunda, A. L., Kamga, H. L. F., Nsagha, S. M., Assob, J. C. N., Wiysonge, C. S., Tabah, E. N., & Njamnshi, A. K. (2011). Knowledge and practices relating to

malaria in a semi-urban area of Cameroon: Choices and sources of antimalarials, self-treatment and resistance. *Pan African Medical Journal*, 9. <https://doi.org/10.4314/pamj.v9i1.71180>

Nzobo, B. J., Ngasala, B. E., & Kihamia, C. M. (2015). Prevalence of asymptomatic malaria infection and use of different malaria control measures among primary school children in Morogoro Municipality, Tanzania. *Malaria Journal*, 14(1), 1–7. <https://doi.org/10.1186/s12936-015-1009-4>

Oguntimilehin, A., Adetunmbi, A. O., & Abiola, O. B. (2015). A review of predictive models on diagnosis and treatment of malaria fever.

Okiring, J., Epstein, A., Namuganga, J. F., Kanya, E. v., Nabende, I., Nassali, M., Sserwanga, A., Gonahasa, S., Muwema, M., Kiwuwa, S. M., Staedke, S. G., Kanya, M. R., Nankabirwa, J. I., Briggs, J., Jagannathan, P., & Dorsey, G. (2022). Gender difference in the incidence of malaria diagnosed at public health facilities in Uganda. *Malaria Journal*, 21(1), 1–12. <https://doi.org/10.1186/S12936-022-040464/tables/6>

Pan, W. D., Dong, Y., & Wu, D. (2018). Classification of malaria-infected cells using deep convolutional neural networks. *Machine Learning - Advanced Techniques and Emerging Applications*. <https://doi.org/10.5772/intechopen.72426>

Patil, P., Yaligar, N., & Meena, S. (2018). Comparison of performance of classifiers - SVM, RF and ANN in potato blight disease detection using leaf images. *IEEE International Conference on Computational Intelligence and Computing Research*, <https://doi.org/10.1109/ICCIC.2017.8524301>

Patouillard, E., Griffin, J., Bhatt, S., Ghani, A., & Cibulskis, R. (2017). Global investment targets for malaria control and elimination between 2016 and 2030. *BMJ Glob Health*, 2(2). <https://doi.org/10.1136/bmjgh-2016-000176>

Paul, S., Ranjan, P., Kumar, S., & Kumar, A. (2022). Disease predictor using random forest classifier. *International Conference for Advancement in Technology*, <https://doi.org/10.1109/ICONAT53423.2022.9726023>

Pillay, E., Khodaiji, S., Bezuidenhout, B. C., Litshie, M., & Coetzer, T. L. (2019). Evaluation of automated malaria diagnosis using the Sysmex XN-30 analyser in a

- clinical setting. *Malaria Journal*, 18(1), 15. <https://doi.org/10.1186/s12936-019-2655-8>
- Polan, D. F., Brady, S. L., Kaufman, R. A., Garg, A., Sharma, B., Khan, R., Pal, M., & Parija, S. (2021). Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*, 1817(1), 012009. <https://doi.org/10.1088/17426596/1817/1/012009>
- Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. *Translational Research*, 194, 36–55. Mosby Inc. <https://doi.org/10.1016/j.trsl.2017.12.004>
- Population Council, & WHO. (2008). Periodic presumptive treatment for sexually transmitted infections experience from the field and recommendations for research. <http://apps.who.int/iris/bitstream/handle/10665/43950/9789241597050eng.pdf;jsessionid=80580A9B3A9CD8A9000EE6FDBE93F873?sequenc=1>
- Prajna, KB., Shreshta, K., & Shetty, S. (2021). A comprehensive study of malaria detection using machine learning. www.ijcrt.org
- Priyadarshini, R., Dash, N., & Mishra, R. (2014). A Novel approach to predict diabetes mellitus using modified Extreme learning machine. *International Conference on Electronics and Communication Systems*, <https://doi.org/10.1109/ECS.2014.6892740>
- Puspitasari, I. W., Rinawan, F. R., Purnama, W. G., Susiarno, H., & Susanti, A. I. (2022). Development of a Chatbot for Pregnant Women on a Posyandu Application in Indonesia: From qualitative approach to decision tree method. *Informatics*, 9(4). <https://doi.org/10.3390/INFORMATICS9040088>
- Quaresima, V., Agbenyega, T., Oppong, B., Awunyo, J. A. D. A., Adomah, P. A., Enty, E., Donato, F., & Castelli, F. (2021). Are malaria risk factors based on gender? A mixed-methods survey in an urban setting in Ghana. *Tropical Medicine and Infectious Disease*, 6(3). <https://doi.org/10.3390/TROPICALMED6030161>
- Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., & Thoma, G. R. (2018). Pre-trained convolutional neural networks as

feature extractors toward improved malaria parasite detection in thin blood smear images. *Peer Journal*, 2018(4). <https://doi.org/10.7717/peerj.4568>

Rajaraman, S., Jaeger, S., & Antani, S. K. (2019). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *Peer Journal*, 7, e6977. <https://doi.org/10.7717/peerj.6977>

Rao, A. R., & Renuka, B. S. (2020). A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis. *IEEE International Conference for Innovation in Technology*, <https://doi.org/10.1109/INOCON50539.2020.9298252>

Ravalji, R., Shah, N., & Nai, M. (2020). Malaria disease detection using machine learning.

Reyburn, H., Mbakilwa, H., Mwangi, R., Mwerinde, O., Olomi, R., Drakeley, C., & Whitty, C. J. M. (2007). Rapid diagnostic tests compared with malaria microscopy for guiding outpatient treatment of febrile illness in Tanzania: Randomised trial. *British Medical Journal*, 334(7590), 403–406. <https://doi.org/10.1136/BMJ.39073.496829.AE>

Ritthipravat, P. (2009). Artificial neural networks in cancer recurrence prediction. *Proceedings - International Conference on Computer Engineering and Technology*, 2, 103–107. <https://doi.org/10.1109/ICCET.2009.84>

Rumisha, S. F., Shayo, E. H., & Mboera, L. E. G. G. (2019). Spatio-temporal prevalence of malaria and anaemia in relation to agro-ecosystems in Mvomero district, Tanzania. *Malaria Journal*, 18(1), 228. <https://doi.org/10.1186/s12936-019-2859-y>

Russell, T. L., Govella, N. J., Azizi, S., Drakeley, C. J., Kachur, S. P., & Killeen, G. F. (2011). Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. *Malar J*, 10, 80. <https://doi.org/10.1186/1475-2875-10-80>

Sanchez, L., Vidal, M., Jairoce, C., Aguilar, R., Ubillos, I., Cuamba, I., Nhabomba, A. J., Williams, N. A., Díez-Padrisa, N., Cavanagh, D., Angov, E., Coppel, R. L., Gaur, D., Beeson, J. G., Dutta, S., Aide, P., Campo, J. J., Moncunill, G., & Dobaño, C. (2020). Antibody responses to the RTS,S/AS01E vaccine and Plasmodium

falciparum antigens after a booster dose within the phase 3 trial in Mozambique. *Vaccines*, 5(1). <https://doi.org/10.1038/S41541-020-0192-7>

Sapkota, A. R., Coker, M. E., Rosenberg Goldstein, R. E., Atkinson, N. L., Sweet, S. J., Sopeju, P. O., Ojo, M. T., Otivhia, E., Ayepola, O. O., Olajuyigbe, O. O., Shireman, L., Pottinger, P. S., & Ojo, K. K. (2010). Self-medication with antibiotics for the treatment of menstrual symptoms in southwest Nigeria: A cross-sectional study. *BMC Public Health*, 10. <https://doi.org/10.1186/1471-2458-10-610>

Saranya, G., & Pravin, A. (2020). A comprehensive study on disease risk predictions in machine learning Related papers A comprehensive study on disease risk predictions in machine learning. *International Journal of Electrical and Computer Engineering*, 10(4), 4217–4225. <https://doi.org/10.11591/ijece.v10i4.pp4217-4225>

Sarkar, D., Bali, R., Sharma, T., Sarkar, D., Bali, R., & Sharma, T. (2018). Machine learning basics. *Practical Machine Learning with Python*, 3–65. Apress. https://doi.org/10.1007/978-1-4842-3207-1_1

Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020). Comparative study of machine learning algorithms for breast cancer prediction. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 796–801. <https://doi.org/10.1109/ICSSIT48917.2020.9214267>

Serpen, A. A. (2016). Diagnosis rule extraction from patient data for chronic kidney disease using machine learning. *International Journal of Biomed and Clinical Engineering*, 5(2), 64–72. <https://doi.org/10.4018/ijbce.2016070105>

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, 52, 456–462. <https://doi.org/10.1016/J.BSPC.2017.01.012>

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine learning in healthcare: A review. *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology*, 910–914. <https://doi.org/10.1109/ICECA.2018.8474918>

- Shekalaghe, S., Cancino, M., Mavere, C., Juma, O., Mohammed, A., Abdulla, S., & Ferro, S. (2013). Clinical performance of an automated reader in interpreting malaria rapid diagnostic tests in Tanzania. *Malaria Journal* 2013 12:1, 12(1), 1–9. <https://doi.org/10.1186/1475-2875-12-141>
- Shretta, R., Liu, J., Cotter, C., Cohen, J., Dolenz, C., Makomva, K., Newby, G., Ménard, D., Phillips, A., Tatarsky, A., Gosling, R., & Feachem, R. (2017). Malaria elimination and eradication. *Disease Control Priorities, Third Edition (6): Major Infectious Diseases*, 315–346. https://doi.org/10.1596/978-1-4648-0524-0_CH12
- Sidey, J. A. M., & Sidey, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1), 64. <https://doi.org/10.1186/s12874-019-0681-4>
- SMO. (2020). Malaria in Tanzania: Statistics & Facts | Severe Malaria Observatory. *Severe Malaria Observatory*. <https://www.severemalaria.org/countries/tanzania>
- Snow, R., Guerra, C., Noor, A., Myint, H., Hay, S. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, 434(7030), 214–217. <https://doi.org/10.1038/nature03342>
- Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, 6. <https://doi.org/10.1177/2055207620914777>
- Sriporn, K., Tsai, C. F., Tsai, C. E., & Wang, P. (2020). Analyzing malaria disease using effective deep learning approach. *Diagnostics*, 10(10). <https://doi.org/10.3390/diagnostics10100744>
- Sundram, S., & Pereira, A. (2007). Comorbid Disorders and Stress. *Encyclopaedia of Stress*, 542–548. <https://doi.org/10.1016/B978-012373947-6.00443-8>
- Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., Kappel, G., & Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLOS ONE*, 12(11), e0188532. <https://doi.org/10.1371/JOURNAL.PONE.0188532>

- Tangirala, S. (2020). Evaluating the impact of GINI Index and Information Gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), 612–619. <https://doi.org/10.14569/IJACSA.2020.0110277>
- Tangpukdee, N., Duangdee, C., Wilairatana, P., & Srivicha, K. (2009). Malaria diagnosis: A brief review. *Korean Journal of Parasitology*, 47, 93–103. <https://doi.org/10.3347/kjp.2009.47.2.93>
- Tekale, S., Shingavi, P., & Wandhekar, S. (2018). Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advance Research Computer and Communication Engineering*, 7(10), 92–96. <https://doi.org/10.17148/ijarce.2018.71021>
- Thawer, S. G., Chacky, F., Runge, M., Reaves, E., Mandike, R., Lazaro, S., Mkude, S., Rumisha, S. F., Kumalija, C., Lengeler, C., Mohamed, A., Pothin, E., Snow, R. W., & Molteni, F. (2020). Sub-national stratification of malaria risk in mainland Tanzania: A simplified assembly of survey and routine data. *Malaria Journal*, 19(1), 1–12. <https://doi.org/10.1186/S12936-020-03250-4/FIGURES/2>
- Triantafyllidis, A. K., & Tsanas, A. (2019). Applications of machine learning in real-life digital health interventions: Review of the literature. *Journal of Medical Internet Research* 21(4), 12286. <https://doi.org/10.2196/12286>
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1), 1–16. <https://doi.org/10.1186/S12911-019-1004-8/FIGURES/12>
- Ullah, R., Khan, S., Ali, H., Chaudhary, I. I., Bilal, M., & Ahmad, I. (2019). A comparative study of machine learning classifiers for risk prediction of asthma disease. *Photodiagnosis and Photodynamic Therapy*, 28, 292–296. <https://doi.org/10.1016/J.PDPDT.2019.10.011>

- UM, C. (2016). Malaria treatment in children based on presumptive diagnosis: a make or mar? *Paediatric infectious diseases: Open access*, 01(02). <https://doi.org/10.21767/2573-0282.100006>
- USAID. (2018). *Tanzania - PMI*. U.S President's Malaria Initiative. <https://www.pmi.gov/where-we-work/Tanzania/>
- Van Driel, N. (2020). Automating malaria diagnosis: a machine learning approach: Erythrocyte segmentation and parasite identification in thin blood smear microscopy images using convolutional neural networks.
- Velliangiri, S., Alagumuthukrishnan, S., & Thankumar Joseph, S. I. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, 165, 104–111. <https://doi.org/10.1016/J.PROCS.2020.01.079>
- Walse, R. S., Kurundkar, G. D., Khamitkar, S. D., Muley, A. A., Bhalchandra, P. U., & Lokhande, S. N. (2021). Effective Use of Naïve Bayes, Decision Tree, and Random Forest Techniques for Analysis of Chronic Kidney Disease. *Smart Innovation, Systems and Technologies*, 195, 237–245. https://doi.org/10.1007/978-981-15-7078-0_22/COVER
- Wang, D., Chaki, P., Mlacha, Y., Gavana, T., Michael, M. G., Khatibu, R., Feng, J., Zhou, Z. bin, Lin, K. M., Xia, S., Yan, H., Ishengoma, D., Rumisha, S., Mkude, S., Mandike, R., Chacky, F., Dismasi, C., Abdulla, S., Masanja, H., & Zhou, X. N. (2019). Application of community-based and integrated strategy to reduce malaria disease burden in southern Tanzania: The study protocol of China-UK-Tanzania pilot project on malaria control. *Infectious Diseases of Poverty*, 8(1), 1–6. <https://doi.org/10.1186/s40249-018-0507-3>
- Wang, H., & Zheng, H. (2013). Model Validation, Machine Learning. *Encyclopaedia of Systems Biology* (pp. 1406–1407). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_233
- White, N. J. (2005). Intermittent Presumptive Treatment for Malaria. *PLOS Medicine*, 2(1), 3. <https://doi.org/10.1371/JOURNAL.PMED.0020003>

- WHO. (2015). Global technical strategy for malaria 2016-2030. *WHO Global Malaria Programme*.
- WHO. (2018). World malaria report 2018. Geneva. *World Health Organization*.
[https://doi.org/ISBN 978 92 4 1564403](https://doi.org/ISBN%20978%2092%204%201564403)
- WHO. (2019). World malaria report 2019. *World Health Organization*.
<https://www.who.int/publications/i/item/9789241565721>
- WHO. (2020). World malaria report 2020. *World Health Organization*.
<https://www.who.int/publications/i/item/9789240015791>
- WHO. (2021). Introduction - WHO *Guidelines for malaria* - NCBI Bookshelf.
<https://www.ncbi.nlm.nih.gov/books/NBK568497/>
- WHO AFRICA. (2018). *WHO recognizes national efforts towards Malaria elimination | WHO | Regional Office for Africa*. <https://www.afro.who.int/news/who-recognizes-national-efforts-towards-malaria-elimination>
- WHO. (2022). *Malaria Report*. <https://www.who.int/news-room/fact-sheets/detail/malaria>
- WHO-Guidelines. (2015). For the treatment of malaria guidelines. www.who.int
- Winskill, P., Rowland, M., Mtove, G., Malima, R. C., & Kirby, M. J. (2011). Malaria risk factors in north-east Tanzania. *Malaria Journal* 2011 10:1, 10(1), 1–7.
<https://doi.org/10.1186/1475-2875-10-98>
- Yadav, S. S., Kadam, V. J., Jadhav, S. M., Jagtap, S., & Pathak, P. R. (2021). Machine learning based malaria prediction using clinical findings. *2021 International Conference on Emerging Smart Computing and Informatics*, 216–222.
<https://doi.org/10.1109/ESCI50559.2021.9396850>
- Yang, F., Yu, H., Silamut, K., Maude, R. J., Jaeger, S., & Antani, S. (2019). Smartphone-supported malaria diagnosis based on deep learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11861 LNCS, 73–80. https://doi.org/10.1007/978-3-030-32692-0_9

- Yapa, H. M., & Bärnighausen, T. (2018). Implementation science in resource-poor countries and communities. *Implementation Science* 2018 13:1, 13(1), 1–13. <https://doi.org/10.1186/S13012-018-0847-1>
- Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., Nsobya, S., Staedke, S. G., Donnelly, M. J., Wabwire-Mangen, F., Talisuna, A., Dorsey, G., Kanya, M. R., & Rosenthal, P. J. (2012). Malaria in Uganda: Challenges to control on the long road to elimination: I. Epidemiology and current control efforts. *Acta Tropical*, 121(3), 184–195. <https://doi.org/10.1016/j.actatropica.2011.03.004>
- Yeung, S., & White, N. J. (2005). How do patients use antimalarial drugs? A review of the evidence. In *Tropical Medicine and International Health* 10(2), 121–138. <https://doi.org/10.1111/j.1365-3156.2004.01364.x>
- Yuan, X., Chen, S., Sun, C., & Yuwen, L. (2022). A novel early diagnostic framework for chronic diseases with class imbalance. *Scientific Reports* 2022 12:1, 12(1), 1–16. <https://doi.org/10.1038/s41598-022-12574-x>
- Zimmerman, P. A., & Howes, R. E. (2015). Malaria diagnosis for malaria elimination. *Current Opinion in Infectious Diseases*, 28(5), 446–454. <https://doi.org/10.1097/QCO.000000000000191>

APPENDICES

Appendix 1: Study ethical clearance from NIMR



THE UNITED REPUBLIC
OF TANZANIA



National Institute for Medical Research
3 Barack Obama Drive
P.O. Box 9653
11101 Dar es Salaam
Tel: 255 22 2121400
Fax: 255 22 2121360
Email: nimrethics@gmail.com

Ministry of Health, Community
Development, Gender, Elderly
& Children
University of Dodoma, College of
Business Studies and Law
Building No. 11
P.O. Box 743
40478 Dodoma

NIMR/HQ/R.8c/Vol. I/1352

25th October, 2019

Martina Mariki
Information Communication Science and Engineering
Nelson Mandela African Institution of Science and Technology
C/o Dr. Joseph Mwangoka
NM-AIST
P O Box 447
Arusha


RE: APPROVAL FOR EXTENSION OF ETHICAL CLEARANCE

This letter is to confirm that your application for extension on the already approved proposal: Machine learning model to improve malaria diagnosis in a resource poor country like Tanzania (Mariki M. et al) whose local investigator is Dr. Joseph Mwangoka of Nelson Mandela African Institution of Science and Technology has been approved.


The extension approval is based on the progress report dated 19th October 2019 on the project, Ref. NIMR/HQ/R.8a/Vol. IX/2486, dated 12th May 2017. Extension approval is valid until 11th May 2020.

The Principal Investigator must ensure that other conditions of approval remain as per ethical clearance letter. The PI should ensure that progress and final reports are submitted in a timely manner.

Name: Prof. Yunus Daud Mgya


Signature
CHAIRPERSON
MEDICAL RESEARCH
COORDINATING COMMITTEE

Name: Prof. Muhammad Bakari Kambi


Signature
CHIEF MEDICAL OFFICER
MINISTRY OF HEALTH, COMMUNITY
DEVELOPMENT, GENDER, ELDERLY
& CHILDREN

Appendix 2: Malaria Patients Records Collection Form

Study title: MACHINE LEARNING MODEL TO IMPROVE MALARIA DISGNOSIS IN A RESOURCE POOR COUNTRIES LIKE TANZANIA

DATA COLLECTION TOOL

S/ N	Age	Sex	Date Of Visit	Residence Area	No of Visits For Malaria Cases in 6 months	Symptoms Observed	Test Taken	Results obtained	Drugs Prescribed

Appendix 3: Questionnaire Used for Patient’s Survey

THE NELSON MANDELA
AFRICAN INSTITUTION OF SCIENCE AND TECHNOLOGY
(NM-AIST)

School of Computational and Communication Sciences and Engineering

Direct Line: +255 272555070
Mobile Phone: +255272970005
Fax: +255 272555071
E-mail: deancocse@nm-aist.ac.tz



Tengeru
P.O. Box 447
Arusha, TANZANIA
Website: www.nm-aist.ac.tz

Study title: MACHINE LEARNING MODEL TO IMPROVE MALARIA DISGNOSIS IN A RESOURCE POOR COUNTRIES LIKE TANZANIA

Instructions

Complete this questionnaire by checking the right answers or filling the right answer in the space provided

Demographics Information

1. Area of residence
 - Village or Ward):
 - District :
2. Gender
 - Male⁰
 - Female¹
3. Age (Years)
 - 5- 18
 - 19 – 45
 - 46+
4. Occupation
5. Education Level
 - Primary School Level
 - Secondary School Level
 - Higher Learning (Collage& University) Level

- Professional Education
- Non- Educated

MALARIA DIAGNOSIS

1. Have you been diagnosed with malaria in the past three months?
 - Yes¹
 - No⁰
2. Did you any malaria related symptoms in the past three months?
 - Yes¹
 - No⁰
3. If yes, how many times have you been diagnosed with malaria in the past three months?
 - 1 time
 - 2 times
 - 3 times
 - More than 3 times
4. Have been diagnosed with
5. Did you get any treatment for such diagnosis of malaria-related symptoms?
 - Yes¹
 - No⁰
6. If yes, what type of treatment did you get?
 - Traditional Treatment
 - Anti-Malaria Drugs § §
 - No treatment
7. If you were given anti-malaria drugs, what was the medication given?

8. If you use traditional treatment what treatment do you use?

9. Do you use any of the following malaria control initiatives? (*check the initiatives used*)
 - Insecticides Treated Nets
 - Insecticides Spray
 - Malaria Vaccine
 - No use of anti-malaria initiatives
10. If not, what are the reasons for not using any of the malaria control initiatives?

11. What are the common malaria symptoms you get when you feel like you have malaria?
 - High fever
 - shaking chills that can change from moderate to severe
 - Profuse sweating when the fever suddenly drops

- Fatigue (being fatigued)
- Headache
- Muscle aches/pain
- Abdominal discomfort /Diarrhea
- Nausea,
- vomiting
- Dizziness
- Delirium and confusion.
- Problem breathing
- Kidney failure
- Severe anaemia
- Yellow discolouration of the skin
- Low blood sugar
- Bloody Stool
- Seizure

12. Have you travelled out of your residential area for the past three months?

- Yes¹
- No⁰

13. If yes, where did you go?

14. For how long?

15. Did you get malaria again shortly after you had been treated?

- Yes¹
- No⁰

Appendix 4: Questionnaire Used for Medical Doctors Survey

This interview aims to gain knowledge on the perception of the medical practitioners on management of malaria specifically the study focuses on symptomatic and asymptomatic factors for malaria amongst patients in both low and high endemic areas in Tanzania.

Interview questions

1. What is your experience in the practice?
 - a. Less than 3 years
 - b. 3 years to 10 years
 - c. More than 10 years
2. Current working region (location)?

Response:

3. What Regions of Tanzania have you worked in the past?

Response:

.....

4. Have you treated malaria patients in you work experience?
 - YES
 - NO

5. If yes what is your general experience in the preliminary observation, diagnosis and treatment?

Response:

.....

6. What do you identify as the main malaria symptoms?

- a. Main symptoms (MS):

Response:

.....

.....

- b. Supporting Symptoms (SS):

Response:

.....

.....

c. Severe Symptoms (SVS)|:

Response:

.....
...

7. Duration of observing the symptoms vs temperature current readings?

Response:

.....
.....

8. Non-symptomatic – related factors

Hints:

a. Area of Residence

b. Age

c. Sex

d. History of Travelers to an endemic area for the patient or family member

.....

e. Non-use of malaria initiatives/ bed net use

.....

f. Pregnancy

g. History of family member suffering from malaria

.....

h. Yearly climate seasons/ weather conditions

.....

i. Distance from residential areas to the health facilities

.....

9. Do malaria symptoms vary from location to location?

▪ YES

▪ NO

10. If yes, what is the variation?

Response:.....

.....

11. In your own opinion, which of the symptoms, when observed, results in a possible positive malaria diagnosis?

Response:

.....
.....
.....

12. If a tool was created to simplify the work of health facilities workers, what are the most significant functions/ criteria it must have in diagnosing malaria?

Response:

.....
.....
.....
.....

Appendix 5: Python code that were employed for features selection

```
HIGH ENDEMIC AREA 01 AI x HIGH ENDEMIC AREA 01 M x KILIMANJARO 22 Aug 2021 x LOW ENDEMIC AREA 01 M x Python 3

[2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

import matplotlib.pyplot as plt
import seaborn as sns
#pd.to_numeric(df.values, errors='coerce')
from sklearn.ensemble import RandomForestClassifier

[3]: url = r"/Users/martinamariki/Google Drive/My Research/From Computer ile/attachments-2/MPR May.csv"
df = pd.read_csv(url, na_values=['#NAME?'])

...
0 1834
1 722
Name: MalariaDiagnosis, dtype: int64

[5]: # Import label encoder
from sklearn import preprocessing

# label_encoder object knows how to understand word labels.
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.
df['ResidenceArea'] = label_encoder.fit_transform(df['ResidenceArea'])

df['ResidenceArea'].unique()

[5]: array([2, 3, 0, 1])

from sklearn import preprocessing

# label_encoder object knows how to understand word labels.
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.
df['VisitDate'] = label_encoder.fit_transform(df['VisitDate'])

df['VisitDate'].unique()

[6]: array([ 3, 11, 10, 12, 1, 7, 6, 8, 9, 4, 0, 2, 5])

[7]: # Import label encoder
from sklearn import preprocessing

# label_encoder object knows how to understand word labels.
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.
df['Sex'] = label_encoder.fit_transform(df['Sex'])

df['Sex'].unique()

[7]: array([1, 0])

[8]: df.columns

[8]: Index(['ResidenceArea', 'VisitDate', 'Age', 'Sex', 'Fever', 'ShakingChills',
'Sweating', 'Fatigue', 'Headache', 'MusclePain', 'JointPain',
'GeneralBodyMalaise', 'ChestPain', 'AbdominalPain', 'Nausea',
'Vomiting', 'Coughing', 'Dizziness', 'Confusion', 'ProblemBreathing',
'Backache', 'Anemia', 'YellowSkin', 'BloodyStool', 'AppetiteLoss',
'Conversion', 'Dehydration', 'Pale', 'RunningNose', 'BlurredVision',
'PainInUrination', 'Palpation', 'Diarrhea', 'Restless', 'Flue',
'IfTestTaken', 'BloodSlide', 'MRDT', 'WidalTest', 'UrineAnalysisTest',
'MalariaDiagnosis', 'WidalDiagnosis', 'UTI'],
dtype='object')
```

```

[9]: feature = ['ResidenceArea', 'VisitDate', 'Age', 'Sex', 'Fever', 'ShakingChills',
              'Sweating', 'Fatigue', 'Headache', 'MusclePain', 'JointPain',
              'GeneralBodyMalaise', 'ChestPain', 'AbdominalPain', 'Nausea',
              'Vomiting', 'Coughing', 'Dizziness', 'Confusion', 'ProblemBreathing',
              'Backache', 'Anemia', 'YellowSkin', 'BloodyStool', 'AppetiteLoss',
              'Conversion', 'Dehydration', 'Pale', 'RunningNose', 'BlurredVision',
              'PainInUrination', 'Palpation', 'Diarrhea', 'Restless', 'Flue']

[10]: X = df[feature]
      y = df.MalariaDiagnosis

[11]: #Model - based feature Selection : Using **SelectFromModel
      from sklearn.feature_selection import SelectFromModel

      select = SelectFromModel(RandomForestClassifier(n_estimators=500, random_state=42), threshold="median")

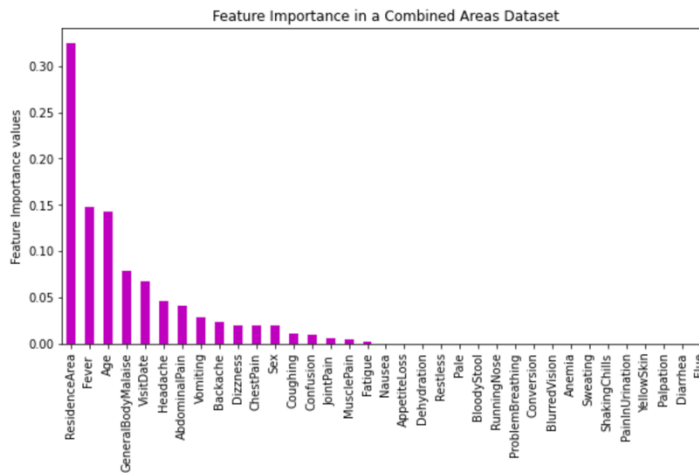
[12]: # Define a classifier
      rforest = RandomForestClassifier(max_depth=30, n_estimators=500, min_samples_leaf=50,
                                     min_samples_split=100, random_state=10)

[13]: # Fit the model
      rforest.fit(X, y)

[13]: RandomForestClassifier(max_depth=30, min_samples_leaf=50, min_samples_split=100,
                             n_estimators=500, random_state=10)

[14]: # Plot the important features
      imp_feat_rf = pd.Series(rforest.feature_importances_, index=X.columns).sort_values(ascending=False)
      imp_feat_rf.plot(kind='bar', title='Feature Importance in a Combined Areas Dataset', figsize=(10,6), color='m')
      plt.ylabel('Feature Importance values')
      plt.subplots_adjust(bottom=0.25)

```



Appendix 6: Python code that was employed for regional model development

```
Launcher x MODEL BUILDING MORO C x Python 3
Code v

[1]: # Packages / libraries
import os #provides functions for interacting with the operating system
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns

%matplotlib inline

# To install sklearn type "pip install numpy scipy scikit-learn" to the anaconda terminal

# To change scientific numbers to float
np.set_printoptions(formatter={'float_kind':'{:f}'.format})

# Increases the size of sns plots
sns.set(rc={'figure.figsize':(8,6)})

# Datetime lib
from pandas import to_datetime
import itertools
import warnings
import datetime
warnings.filterwarnings('ignore')

from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, r2_score

# pip install graphviz
# conda install python-graphviz

[2]: url = r"/Users/martinamariki/Google Drive/My Research/From Computer ile/attachments-2/MPR 2023 MORO FS.csv"
df = pd.read_csv(url, na_values=['#NAME?'])
```

```
[37]: # Your code goes here
X = new_df.drop('MalariaDiagnosis', axis=1).values# Input features (attributes)
y = new_df['MalariaDiagnosis'].values # Target vector
print('X shape: {}'.format(np.shape(X)))
print('y shape: {}'.format(np.shape(y)))

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.7, test_size=0.3, random_state=0)

X shape: (1527, 29)
y shape: (1527,)

[38]: dt = DecisionTreeClassifier( max_depth=4, random_state=1)
dt.fit(X_train, y_train)

[38]: DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
max_depth=4, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort='deprecated',
random_state=1, splitter='best')

[35]: import graphviz

dot_data = tree.export_graphviz(dt, out_file=None,
feature_names=new_df.drop('MalariaDiagnosis', axis=1).columns,
class_names=new_df['MalariaDiagnosis'].unique().astype(str),
filled=True, rounded=True,
special_characters=True)
graph = graphviz.Source(dot_data)
graph
```

Appendix 7: Research Outputs - Poster and Published Papers

FEATURE SELECTION APPROACH TO IMPROVE MALARIA DIAGNOSIS MODEL'S PERFORMANCE FOR HIGH AND LOW ENDEMIC AREAS OF TANZANIA

Martina Mariki, Neema Mduma, Elizabeth Mkoba





Background

Malaria remains a significant cause of death in sub-Saharan Africa. In disease management, self-medication and presumptive treatment are one of the major practiced methods. These methods can lead to drug resistance, treatment of other diseases and even death. Developing a Malaria diagnosis model using patients' symptoms and demographic features is a practical solution for those who can't access proper diagnosis or have self-medication behaviour. Important features were selected to improve model performance and reduce the processing time of the diagnostic model. Model-based feature selection method was used to obtain significant features in malaria diagnosis.

Methods

- Study sites – Morogoro and Kilimanjaro
- Dataset collection – Designed Form based on the summary of MoH patient's file
- Feature selection – Model-based feature selection method
- Features evaluation – Machine learning Model, Doctors perspective on the selected features



Results

The ranking of the features was different among the regional datasets. The trained models attained the highest performance accuracy with the selected important features. Non-symptomatic features are as important as symptomatic features. Seasonal malaria is significant in disease management.

	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Doctor 5	Doctor 6
Main Symptoms	Headache Backache Vomiting	Headache Vomiting High fever Body Pain Diarrhoea	Fever Headache Vomiting Nausea	Fever Body pain Headache Vomiting	Fever Headache Sweating Body pain	Headache Backache Vomiting
Severe Symptoms	Dizziness Anaemia Confusion	Excessive vomiting Yellow fever Paleness	High fever Vomiting conversion Loss of consciousness Anaemia Coca-Cola urine	Confusion fainting Kidney of failure	Confusion Loss of conscious	Dizziness Anaemia Confusion





Conclusion

The study important features identified matched the malaria diagnosis features criteria by WHO. The difference in the level of importance of the malaria diagnosis features for different regions signifies that each region is unique and their patients should not be observed the same

Presented at 1st ICTA – EMOS Conference 2022

