**A Machine Learning Approach to Syndromic Management of Sexually Transmitted Diseases in Zimbabwe using** **Artificial Neural Networks**

**By**

Benson Misi

(C18135140L)

A Research Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

**Master of Science in Data Analytics**

Graduate Business School

Chinhoyi University of Technology

Zimbabwe

Sunburst chart

Description automatically generated

**Supervisor: Mr. W Kanyongo**

**July 2022**

# APPROVAL FORM

The undersigned certify that they have read and recommended to the Graduate Business School, Chinhoyi University of Technology, for acceptance a dissertation entitled, “A Machine Learning Approach to Syndromic Management of Sexually Transmitted Diseases in Zimbabwe using Artificial Neural Networks”, submitted by Benson Misi, in partial fulfilment of the requirements for the Master of Science Degree in Data Analytics.

Name of Supervisor: Mr. W Kanyongo

Signature: …………………………

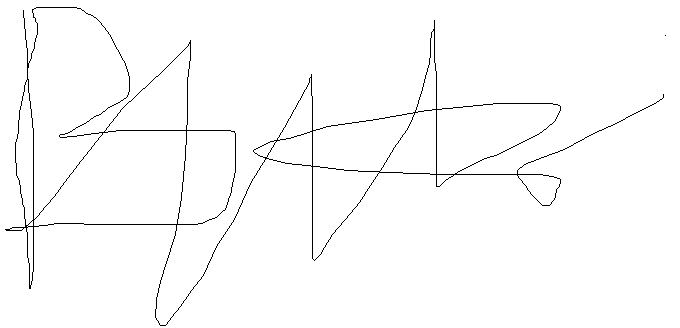
Date: July 2022

# DECLARATION

I, Benson Misi, declare that this MSc study is my own effort and is a true reflection of research executed by me. This research in full or part thereof has not been submitted for examination for any degree at any other university/institution. No part of this dissertation may be reproduced, stored in any retrieval system, or transmitted in any form, or by any means (e.g. electronic, mechanical, photocopying, recording or otherwise) without the prior express permission of me the author, or Chinhoyi University of Technology on my behalf. I, Benson Misi, grant Chinhoyi University of Technology permission to reproduce this dissertation in whole or in part, in any manner or format, which Chinhoyi University of Technology may deem fit.

Name of Student: Benson Misi (C18135140L)

Signature:



Date: 26 July 2022

**CHINHOYI UNIVERSITY OF TECHNOLOGY**

**GRADUATE BUSINESS SCHOOL**

**SUPERVISION PROGRESS REPORT FORM**

**Name of student**: Benson Misi **Registration number**: C18135140L

**Name of supervisor**: Mr W. Kanyongo

**Topic: A Machine Learning Approach to Syndromic Management of Sexually Transmitted Diseases in Zimbabwe using Artificial Neural Networks**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Date of review** | **Area covered** | **Lecturer’s recommendation** | **Key deliverables for next review** | **Date of next review** | **Supervisor’s signature** |
| 17/02/2022 | Concept Note presentation | a) To rephrase my topic so that it clearly defines area of study  b) I should come up with smart objectives | a) Implementation of recommendations  b) Present draft chapter 1 | 07/03/2022 |  |
| 15/04/2022 | a) Discussion on concept note correction recommendations  b) Chapter One presentation | a) To define context or environment in which the study is conducted  b) To be consistent on area of focus between STD or STI  c)To remove unnecessary stories on background study  d) To have an objective developed around ascertaining the syndromic actor that have statistical significance in developing a ML model  e) To add more depth on Significance of the Study  d) To add more depth on Limitations of the Study | a) Implementation of recommendations  b) Present draft chapter 2 | 09/05/2022 |  |
| 09/05/2022 | a) Discussion on chapter 1 correction recommendations  b) Chapter 2 draft presentation | a)To mention the theories around your study and the theoretical framework you are basing your study on.  And empirical literature  b) To add additional literature review |  | 01/06/2022 |  |
| 01/06/2022 | a) Discussion on chapter 2 correction recommendations  b) Chapter 3 draft presentation | a) To maintain a quantitative research philosophy  b) To further elaborate on my understanding on the variables and data  c) To elaborate on what I will be doing to achieve business understanding | a) Implementation of recommendations  b) Present draft chapter 4 | 25/06/2022 |  |
| 25/06/2022 | a) Discussion on chapter 3 correction recommendations  b) Chapter 4 draft presentation | a) To make all letters capitals for all your chapter headings | a) Implementation of recommendations  b) Present draft chapter 5 | 11/07/2022 |  |
| 11/07/2022 | a) Discussion on chapter 4 correction recommendations  b) Chapter 5 draft presentation | a) To organize my summary of findings and conclusions by themes of my objectives. | a) Implementation of recommendations  b) Finalize document for submission |  |  |

**Overall comments by the supervisor**: ………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**Student’s signature**: …………………………… **Date**…………………………..

**Supervisor’s signature**: ……………………….. **Date**…………………………..

# ABSTRACT

The aim of this study was to evaluate the use of Machine Learning algorithms in the syndromic case management to increase the prevention of Sexually Transmitted Diseases in Zimbabwe by making clinical decisions based on a patient's symptoms and signs. The primary objective was to develop a Machine Learning model on syndromic management of Sexually Transmitted Diseases in Zimbabwe. The study utilised a case study approach and used the cross-industry standard process for data mining (CRISP-DM) methodology. The data for the study were collected at a government clinic over a five-year period from 2015 to 2020. The findings showed that the types organism causing the infection, sex, where symptoms appear, patient symptoms, time when symptoms showed, availability of symptoms, treatment, tests, age, time for symptoms to show and availability of test results were the statistically significant features for determine the type of sexually transmitted diseases. Four different Machine Learning algorithms, Logistic Regression, Decision Trees, Random Forest, Naïve Bayes and Artificial Neural Networks were tested to find the best performing algorithm. Artificial Neural Networks performed the best with over 92% accuracy which went up to 97% after some Hyperparameter tuning. The findings showed that syndromic management can be improved by using Artificial Neural Networks. The researcher recommends that the ministry of Health must invest in the training of personnel as a way to start adopting the use of ML technologies and that academia can take a leading role in the research and implementation of the programme. The research also recommended that future studies look into: 1) Identification of other factors which may be localised in the determination of STDs, 2) Time series analysis of syndromic management in Zimbabwe using ANN and hybrid sliding window models and 3) identification of repeat infections of STDs using Artificial Neural Networks

TABLE OF CONTENTS

[APPROVAL FORM i](#_Toc109744274)

[DECLARATION ii](#_Toc109744275)

[ABSTRACT vi](#_Toc109744276)

[LIST OF FIGURES x](#_Toc109744277)

[LIST OF TABLES xi](#_Toc109744278)

[LIST OF ABBREVATIONS xii](#_Toc109744279)

[CHAPTER ONE 1](#_Toc109744280)

[INTRODUCTION AND BACKGROUND 1](#_Toc109744281)

[1.1 Background to the study 1](#_Toc109744282)

[1.2 Statement of the Problem 5](#_Toc109744283)

[1.3 Research Objectives 5](#_Toc109744284)

[1.3.1 Primary Research Objective 5](#_Toc109744285)

[1.3.2 Secondary objectives 5](#_Toc109744286)

[1.4 Research Questions 6](#_Toc109744287)

[1.5 Hypothesis of the Study 6](#_Toc109744288)

[1.6 Assumptions of the Study 6](#_Toc109744289)

[1.7 Significance of the Study 6](#_Toc109744290)

[1.7.1 Health Sector 7](#_Toc109744291)

[1.7.2 Universal Body of Knowledge 7](#_Toc109744292)

[1.7.3 Development Partners 7](#_Toc109744293)

[1.8 Structure of the Study 7](#_Toc109744294)

[1.9 Chapter Summary 8](#_Toc109744295)

[CHAPTER TWO 9](#_Toc109744296)

[LITERATURE REVIEW 9](#_Toc109744297)

[2.1 Introduction 9](#_Toc109744298)

[2.2 Artificial Intelligence, Machine Learning and Deep Learning 9](#_Toc109744299)

[2.3 Theoretical Framework 11](#_Toc109744300)

[2.3.1 The Machine Learning Theory 11](#_Toc109744301)

[2.3.2 Computational Learning Theory 12](#_Toc109744302)

[2.4 Control of Sexually Transmitted Diseases 12](#_Toc109744303)

[2.4.1 The Rationale for Syndromic Management of STIs 13](#_Toc109744304)

[2.4.2 Understanding the Syndromic Approach to Management of STIs 13](#_Toc109744305)

[2.4.3 The Efficiency of Syndromic Management of STIs 14](#_Toc109744306)

[2.5 Machine Learning 15](#_Toc109744307)

[2.5.1 Supervised Learning 16](#_Toc109744308)

[2.5.2 Unsupervised Learning 16](#_Toc109744309)

[2.5.3 Reinforcement Learning 16](#_Toc109744310)

[2.6 Machine Learning in the Health Sector 17](#_Toc109744311)

[2.6.1 Artificial Neural Networks in Health Management 18](#_Toc109744312)

[2.6.2 The Impact of Machine Learning on Detecting STIs 20](#_Toc109744313)

[2.7 Factors That Hinder the Use of Machine Learning On STI Screening 21](#_Toc109744314)

[2.8 Summary 22](#_Toc109744315)

[CHAPTER 3: RESEARCH METHODOLOGY 23](#_Toc109744316)

[3.1. Introduction 23](#_Toc109744317)

[3.2. Research philosophy 23](#_Toc109744318)

[3.2. Research Design 24](#_Toc109744319)

[3.2.1 Case Study Research Design 24](#_Toc109744320)

[3.3. Research Approach 24](#_Toc109744321)

[3.4. Population and Sampling 24](#_Toc109744322)

[3.4.1 Population 25](#_Toc109744323)

[3.4.2 Sampling 25](#_Toc109744324)

[3.5.1 Business Understanding 26](#_Toc109744325)

[3.5.2 Data Understanding 27](#_Toc109744326)

[3.5.1 Data Preparation 29](#_Toc109744327)

[3.5.2 Modelling of the Algorithm 32](#_Toc109744328)

[3.5.3 Selection and Evaluation of the Algorithm 32](#_Toc109744329)

[3.6 Summary 33](#_Toc109744330)

[CHAPTER 4: FINDINGS AND DISCUSSIONS 35](#_Toc109744331)

[4.1 Introduction 35](#_Toc109744332)

[4.2 Exploratory Data Analysis 35](#_Toc109744333)

[4.2.1 Distribution by gender 35](#_Toc109744334)

[4.2.2 Age 36](#_Toc109744335)

[4.2.3 The distribution of the STIs 37](#_Toc109744336)

[4.3 To determine factors that have a statistical significance to syndromic management. 38](#_Toc109744337)

[4.4 Developing a Machine Learning model to accurately screen and diagnose patients based on their symptoms. 39](#_Toc109744338)

[4.4.1 Modelling 39](#_Toc109744339)

[4.4.2 Model Preparation 40](#_Toc109744340)

[4.5 To assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms using Artificial Neural Networks. 42](#_Toc109744341)

[4.5.3 Hyperparameter Tuning 44](#_Toc109744342)

[4.6 Discussion of the findings 45](#_Toc109744343)

[4.6.1 Objective 1: To determine factors that have a statistical significance to syndromic management. 45](#_Toc109744344)

[4.6.2 Objective 2: To develop a Machine Learning model to accurately screen and diagnose patients based on their symptoms. 46](#_Toc109744345)

[4.6.3 To assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms. 46](#_Toc109744346)

[4.6.4 Hypotheses testing 47](#_Toc109744347)

[4.6 Summary 47](#_Toc109744348)

[Chapter 5: Summary, Conclusions and Implications 48](#_Toc109744349)

[5.1 Summary of research findings 48](#_Toc109744350)

[5.1.1Factors that have a statistical significance to syndromic management 48](#_Toc109744351)

[5.1.2 Developing a Machine Learning model that accurately screens, and diagnose patients based on their symptoms 49](#_Toc109744352)

[5.1.3 Assessing the Effectiveness of the Machine Learning Model that Categorises and Diagnoses Patients Based on their Symptoms 49](#_Toc109744353)

[5.2 Conclusions of the research 49](#_Toc109744354)

[5.3 Implications of the research 49](#_Toc109744355)

[5.3.1 Theoretical Implications 50](#_Toc109744356)

[5.3.2 Policy Implications 50](#_Toc109744357)

[5.3.3 Practical Implications 50](#_Toc109744358)

[5.4 Limitations of the study 50](#_Toc109744359)

[5.5 Areas for further study 51](#_Toc109744360)

[References 52](#_Toc109744361)

Appendix………………………………………………………………………………………………………………………………………….59

# LIST OF FIGURES

[Figure 2.1: Artificial Intelligence, Machine Learning and Deep Learning 10](#_Toc109738736)

[Figure 2.2: Machine Learning Framework 17](#_Toc109738737)

[Table 3.1. Feature Description 27](#_Toc109738738)

[Figure 4.1. A screenshot of the dataset 28](#_Toc109738739)

[Figure 4.2. The feature data types according to pandas 30](#_Toc109738740)

[Figure 3.3 the columns before each array is split 31](#_Toc109738741)

[Figure 3.4 Dataset after splitting 32](#_Toc109738742)

[Figure 4.1 Dataset distribution by gender 35](#_Toc109738743)

[Figure 4.2 Age distribution 36](#_Toc109738744)

[Figure 4.3 Distribution of STIs 37](#_Toc109738745)

[Figure 4.4: The number of days against the infections 38](#_Toc109738746)

[Figure 4.5: The final dataset 40](#_Toc109738747)

[Figure 4.6: Data Splitting 40](#_Toc109738748)

[Figure 4.6: Modelling 41](#_Toc109738749)

[Figure 4.7: Model being run 41](#_Toc109738750)

[Figure 4.9: Model Results 42](#_Toc109738751)

[Figure 4.10: Accuracy rate training vs testing 43](#_Toc109738752)

[Figure 4.11: Model Loss 44](#_Toc109738753)

[Figure 4.12: Model statistics after Hyperparameter tuning 45](#_Toc109738754)

# LIST OF TABLES

[Figure 1.1: Level of STDs infection in Africa 4](#_Toc109738802)

[Table 3.2 Machine Learning Algorithms tests 33](#_Toc109738803)

[Table 4.1 Statistical significance test for categorical features 38](#_Toc109738804)

[Table 4.2 ANOVA Tests for the numeric features vs the categorical dependent feature 39](#_Toc109738805)

[Table 4.2 Model Hyperparameters 40](#_Toc109738806)

# LIST OF ABBREVATIONS

STI – Sexually Transmitted Infection.

STD – Sexually Transmitted Disease

ANN – Artificial Neural Netwwork

ML – Machine Learning

AI – Artificial Intellegence

# CHAPTER ONE

# INTRODUCTION AND BACKGROUND

## 1.1 Background to the study

Sexually transmitted diseases (STDs) are of important public health concern because they mostly affect young adults, carry stigma, facilitate transmission, and acquisition of HIV infection, and have sequelae that impose a significant socioeconomic cost (Levy, Gunta & Edemekong, 2019). Pelvic inflammatory disease (PID), infertility, ectopic pregnancy, persistent pelvic pain, cervical cancer, and urethral stricture are all complications of failing to diagnose and treat infections. Miscarriages, stillbirths, neonatal mortality, mental retardation, neonatal conjunctivitis, and pneumonia are all examples of the effects on of STDs among people in the society (Barrow *et al.,* 2020). According to the World Bank (2020), STDs, excluding HIV, are the second leading cause of healthy life loss in women between the ages of 15 and 44, behind maternal mortality and morbidity. Despite this burden, many impoverished countries, like Zimbabwe, have given STDs a low priority.

Many countries, particularly developing countries, lack adequate STD control programs (Nguyen, 2019). Fortunately, information from recent research indicating that controlling STDs can significantly reduce the incidence of HIV, Syphilis, Gonorrhoea, and other STDs has boosted interest in STDs (Widman *et al.,* 2018; Morales *et al.,* 2018; Henderson *et al.,* 2020). The primary goal of STD control is to prevent STD transmission, progression, and consequences. Detecting and curing disease by providing enough diagnostic and treatment facilities, as well as reducing infection consequences by providing early and effective treatment for both symptomatic and asymptomatic patients and their contacts, are examples of strategies to achieve these goals. One method for secondary STD prevention is syndromic case management. This study aims to develop a Machine Learning Model that uses a Recurrent Neural Network (RNN) to detect STDs in order to quickly identify the infections making it easier in providing the appropriate treatment to the patients. Primary method involves issues such as counselling, education, increasing access to medical facilities and counselling services but due to cultural and behavioural changes among target audience this method is subject to stigma hence less effective (Beksinska, Wong & Smit, 2020).

STDs have been there since the beginning of time, and outbreaks are even mentioned in the Bible (Pfennig, 2019). The attitude of people about STDs have altered over time as more information about the condition has been available. In ancient times, STDs were seen as an individual retribution for living a sinful lifestyle or as a result of poor cleanliness and hygiene. In medieval times, the link between sexual behaviour and disease was recognized, but the range of clinical symptoms were considered as varieties of one disease, depending on the stage of the disease and the overall health of the diseased individual (Moore, Rosenthal & Mitchell, 2020). In the late 15th and 16th centuries, a supposedly new plague was carried to Europe and quickly spread by soldiers. Misinterpretations of inaccurate investigations on the assumed identity of syphilis and gonorrhoea led to nosologic misunderstandings in the 17th and early 19th centuries. Thanks to major discoveries in microbiology and chemistry, the STDs, which had tormented millions of ordinary and renowned individuals of all socioeconomic levels for centuries and been linked to several scandals, were eventually destroyed in the late 19th and early 20th centuries. Furthermore, STDs have progressed from a one-time occurrence to a widespread public health concern.

STDs and STIs infection among the people in the world have remained very high, despite the efforts by many health organisations calls to combat the spreading of the diseases. A study conducted by the World Health Organisation (WHO, 2018) has showed that more than one million sexually transmitted diseases are acquired every day worldwide which is a significant figure that calls for immediate innervations to reduce the rate of infection. More so, in every year, WHO has found that about 376 million new infections occur worldwide, with 1 of 4 STIs which are chlamydia, gonorrhoea, syphilis and trichomoniasis. A major worry to the existence of STDs is drug resistance, especially for gonorrhoea, which is a big threat to reducing the impact of STIs worldwide (Curtis, Hoolaghan & Jewitt, 2018). Establishing ways to detect the diseases earlier and easily is important for it enables early cure for the diseases as well.

According to the Center for Disease Control and Prevention, annual STD cases in the United States have continued to rise, hitting an all-time high in 2019. Since then, the number of people infected with these STDs has risen, especially during the Covid-19 pandemic (Irwin & Shafer, 2021). Such alarming increase in the spread of STDs around the world calls for strengthening of preventive and curing measures in order to ensure the health and wellbeing of the citizens.

However, countries can take the changes brought by technology advancement in big data analytics where the world is swiftly shifting towards the use of machine learning in detection and reducing the spread of diseases in the health sector various such STDs (Jong & Stevens, 2021). Machine learning involves the use of advanced analytic techniques to big, diverse data sets that include structured, semi-structured, and unstructured data from a variety of sources with sizes ranging from terabytes to zettabytes (Galetsi, Katsaliaki & Kumar, 2020). Machine learning, according to Saranya and Asha (2019), is concerned with how to create computer programs that automatically improve with experience. Hospitals and healthcare institutions have adopted machine learning to in management of diseases through extracting clinically useful information from data, plan medication, manage patients, and screen for STDs (Kumar & Singh, 2018). Nevertheless, the use of machine learning in the health sector has relatively been low in developing countries.

Early detection for STD identification and treatment necessitates sophisticated and expensive devices, which are unfortunately lacking in most underdeveloped nations like Zimbabwe (Biswas, 2021). Furthermore, test results are rarely available in a fair length of time, necessitating many visits for regular check-ups or screening for STDs, delaying treatment and extending the period of infectivity, increasing the danger of undesirable, often inconceivable, problems. The syndromic method to STD management is centred on identifying a consistent group of symptoms and syndromes to classify the precise disease or infection ahead of time, so that additional investigations can be conducted based on this first criterion (Almugti *et al.,* 2022). The current study develops a machine learning model that aids in the detection of various STDs based on symptoms, which is a fully automated computer task that delivers life-saving information in the quickest time feasible.

In addition, STDs have been responsible for severe adverse reproductive outcomes in Africa and the continent remains the worst affected in the world (WHO, 2016). The more, Sub-Saharan Africa (SSA) ranks first in STD yearly incidence compared to other world regions. The World Health Organization has estimated that every year in Africa there are 3.5 million cases of syphilis, 15 million cases of chlamydial disease, 16 million cases of gonorrhoea, and 30 million cases of trichomoniasis. These statistics reveal on how serious countries ought to manage the spread of the STDs as strategies to achievement of sustainable development goal of health and wellbeing of citizens. Figure 1.1 below shows the top 10 countries in SSA with high STD infections.

###### Figure 1.: Level of STDs infection in Africa

***Source: WHO (2019)***

Figure 1.1 above shows that STD infections are high in SSA, which necessitates the current study to develop better management tools in a bid to reduce the effects of the diseases.

Zimbabwe is also among countries that are experiencing high levels of STD infections as indicated in Figure 1.1. With the aid of a number of development partners, including the World Health Organization, the government has since made steps to manage and control the spread of STDs (Lowe *et al.,* 2019). Treatment for STDs is provided in both the formal and informal public and private sectors in the country. All basic health care and reproductive health facilities, including primary care clinics, mother and child health care centres, family planning clinics, and prenatal and postnatal clinics, are available in the formal public sector. STD care is also accessible at district, provincial, and central hospitals' outpatient departments, as well as classified STD clinics in Harare and Bulawayo. All registered doctors in the private sector are able to treat people with STDs. In addition, people with STDs may seek treatment in the informal sector. Antibiotics and other pharmaceuticals are also available in the marketplaces, and drug peddlers. Despite having such facilities in management of STDs in the country, the country still faces challenges with management of STDs which the current study seeks to improve by developing a machine learning model that improves on STDs management.

STDs constitute important primary health issues in Zimbabwe which face inadequacy of resources required in early detection and investigative procedures for their diagnosis and treatment. Syndromic approach to management of STDs which the current study adopts through machine learning is based on the identification of a consistent group of symptoms and syndromes to classify the exact disease or infection beforehand, so that further investigations are sought for based on these initial criteria. Zimbabwe is in an early stage regarding the use of Informatics and Data Science in health care for sustainable health system but is also under international obligation to adapt it (Furusa & Coleman, 2018). The current study presents Data Analytics in developing a machine learning model that identifies STDs using Deep Learning methods.

## 1.2 Statement of the Problem

STD control aims at reducing the spreading and development of sexually transmitted diseases through the use primary and secondary preventative methods. Secondary prevention method involves detecting and curing diseases which involves prescriptions and laboratory test. However, in Zimbabwe, there is often an absence of laboratory support and medical facilities are often not available or too expensive to properly examine patients, and people often have to go to the cities in order to access a physician and laboratory facilities (Chawurura *et al.,* 2019). The aim of this study is to fill this gap by coming up with a machine learning model that utilises the syndromic case management approach to increase the prevention of STDs by making clinical decisions based on a patient’s symptoms and signs.

## 1.3 Research Objectives

The study seeks to achieve the following research objectives.

### 1.3.1 Primary Research Objective

The primary objective of this study is to develop a machine learning model on syndromic management of STDs in Zimbabwe.

### 1.3.2 Secondary objectives

1. To determine factors that have a statistical significance to the Machine Learning syndromic management model.
2. To develop a Machine Learning model that accurately screens, and diagnose patients based on their symptoms in Zimbabwe.
3. To assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms in Zimbabwe.

## 1.4 Research Questions

The research questions of the study are:

1. Which of the factors are statistically significant to the Machine Learning syndromic management model?
2. Which Machine Learning algorithm can be used to create a model for syndromic management in Zimbabwe?
3. How effective is the proposed machine learning model that categorises and diagnose patients based on their symptoms using Artificial Neural Networks in Zimbabwe?

## 1.5 Hypothesis of the Study

: The use of Recurrent Neural Networks improves the accuracy of STD diagnosis in Zimbabwe.

: The use of Recurrent Neural Networks does not improve the accuracy of STD diagnosis in Zimbabwe.

## 1.6 Assumptions of the Study

The data set used in the study provides the accurate information that improves the accuracy of the model.

## 1.7 Significance of the Study

This study is important because it seeks to improve on syndromic management of STDs in Zimbabwe by adding machine learning model. The application of machine learning models in syndromic management of STDs is still limited in Zimbabwe as big data analytics is finding its space in the country recently. Thus, the current study makes application of big data analytics in the Zimbabwean heath sector which is a gap the current study seeks to address. The control of the spread and effects of STDs in Zimbabwe at large benefits the citizens, and the country Zimbabwe will be able to achieve sustainable development goal of health and wellbeing of the citizens. However, the findings of this study are beneficial to a number of stakeholders including:

### 1.7.1 Health Sector

The Zimbabwean health sector is set to benefit from the findings of this study on machine learning approach to categorisation and diagnosis of STDs using Artificial Neural Networks. This will help in easier identification of STDs so that appropriate cure is given to the patients in the earliest time possible. By doing so, active cases of STDs in Zimbabwe will decrease.

### 1.7.2 Universal Body of Knowledge

Several studies (Chirenje *et al*., 2018; Machiha *et al.,* 2018; Lowe *et al.,* 2019) have been conducted to examine the management of STDs in Zimbabwe. Different approaches have been used though syndromic management has been mostly used. However, there still exist a gap in the Zimbabwean literature pertaining the use of machine learning models in diagnosis of STDs in the country. The current study serves as empirical literature in the country which uses machine learning model in management of STDs, which is a gap in literature that the study seeks to address. Machine learning models have become very popular in the health sector around the world as they are effective in diagnoses of STDs (Elder *et al.,* 2021). Development of a highly accurate machine learning model of diagnoses of STDs in Zimbabwe will help in improving the health status of the Zimbabweans.

### 1.7.3 Development Partners

There are a number of development partners like WHO, UNICEF, and other non-governmental organisations that supports the Zimbabwean health sector which can also benefit from the findings of the study. The findings highlight on the accuracy of diagnosis of patients of STDs where funders have to focus their aid in reducing the spread and effects of STDs in the country. Therefore, it becomes easier for these development partners to support with the required resources such as medicines which will improve the health of the citizens. Development partners like WHO is also provided with data on the efforts made in the country on prevention of STDs, and where possible, funding can be availed to increase data collection so as to improve on the accuracy of the machine models developed.

## 1.8 Structure of the Study

The focus of this research is on a recurrent neural network-based machine learning approach to categorise and diagnose STDs in Zimbabwe.

Chapter One looked at the background to the study, on STDs around the world, Africa, and Zimbabwe. The first chapter also presents the research problem, research objects and research questions. In addition, the first chapter discusses the significance of carrying out the study to different stakeholders, delimitations, and limitations to the study.

Chapter Two focuses on both theoretical and empirical literature review which is also guided by the research objectives in developing an appropriate machine learning model used for control and management of STDs in Zimbabwe.

Chapter Three of the study looks at research methodology which covers research philosophy, research strategy, research design, and the target population for the study, the sample size, data collection procedure, and sampling methods.

Chapter Four deals with presentation, analysis and interpretation of the results closely looking at the research objectives.

Chapter Five provides the summary of main findings, conclusions, policy recommendations, and suggested areas for further research.

## 1.9 Chapter Summary

The current Chapter introduced the study by discussing the need to use machine learning models in management of STDs in the country. Also, the Chapter discussed the background to STDs around the World, Africa, and Zimbabwe, in conjunction with the use of Machine Learning in detecting STDs. The discussion of the background to the study led to the establishment of the statement of the problem, where the objectives of the study would be achieved using ANN. The next chapter reviews both theoretical and empirical literature.

# CHAPTER TWO

# LITERATURE REVIEW

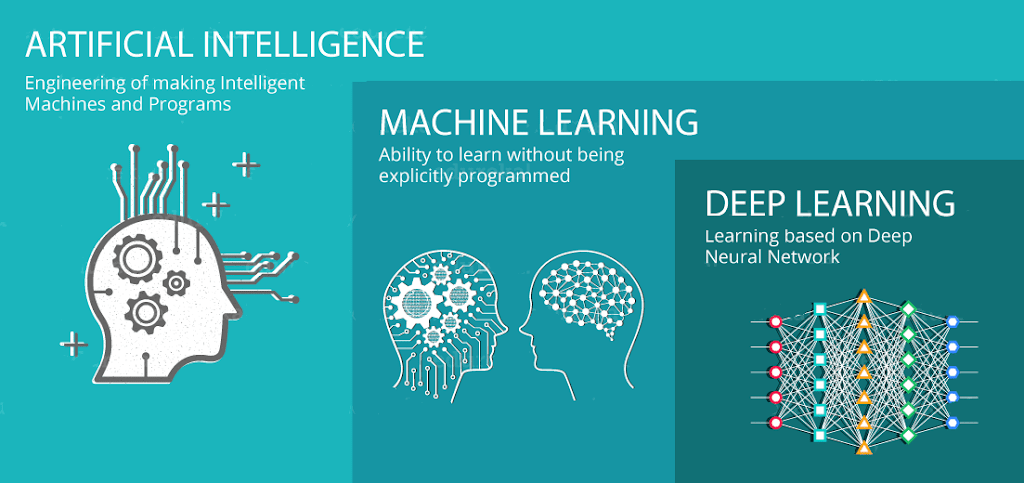
## 2.1 Introduction

The first chapter set out the background to the study and the pertinent research questions and objectives. This chapter will now prove a synthesis of the related literature. The chapter will first of all provide definition of key terms that will be used throughout the thesis before providing the theoretical framework that underpinned the research then it will provide a brief overview of Sexually Transmitted Infections before identifying the various interventions that have been used to contain STIs. Thereafter, the chapter will then provide the a synthesis on the use of Artificial Intelligence (including Deep Learning and Machine Learning technologies) and their use in the health sector before zoning in on their use in syndromic management.

This study was based on syndromic management of STIs using a Deep Learning approach in categorising STIs using Artificial Neural Networks and this Chapter presents the literature review of the study guided by research objectives which are 1) to develop a machine learning model that accurately screens, and diagnose patients based on their symptoms using Artificial Neural Networks, 2) evaluating the effectiveness of the Machine Learning model created in (1), and 3) identify the shortcomings of using Machine Learning model in syndromic management.

## 2.2 Artificial Intelligence, Machine Learning and Deep Learning

Most researchers may use the same terms in slightly different terms which may confuse readers (Kivunja, 2018). Thus, this section will provide an explanation of such terms. The first major definition will of Artificial Intelligence (AI). Artificial intelligence is the imitation of human intelligence by technology, particularly computer systems. Expert systems, natural language processing, speech recognition, and machine vision are examples of AI applications. Figure 2.1 shows the differences between Artificial Intelligence and the other fields.



##### Figure 2.1: Artificial Intelligence, Machine Learning and Deep Learning

*Source: Medium.com*

At its most fundamental, Machine Learning is the discipline of processing data, learning from it, and then making a decision or forecast about an event or phenomenon in the real world (Shailaja, Seetharamulu & Jabbar, 2018). Instead of hard-coding software routines with a specific sequence of commands to achieve a certain activity, the machine is can be said to be “trained” using enormous volumes of data and algorithms that allow it to learn how to perform the task (El Naqa & Murphy, 2015). The machine is then tested against a different set of data in order to evaluate it accuracy or effectiveness in carrying out the said tasks.

Deep Learning is a subset of Machine Learning. Shailaja, Seetharamulu and Jabbar (2018) opine that the key differences are in how each algorithm learns and how much data each type of algorithm requires. Deep learning automates most of the feature extraction process, removing part of the need for manual human intervention. Through a collection of algorithms, neural networks, specifically, artificial neural networks, mimic the human brain (Bartlett, Montanari & Rakhlin, 2021). A neural network is made up of four primary components: inputs, weights, a bias or threshold, and an output.

Sexually transmitted infections (STIs) are sometimes referred to as sexually transmitted diseases (STDs) (Barrow, Ahmed, Bolan & Workowski, 2020). They are infections contracted by sexual intercourse with another individual. According to the World Health Organization (n.d.), there are over 20 different forms of STDs/STIs. Sexual contact, including vaginal, anal, and oral sex, is the most common way STIs spread. During pregnancy, childbirth, and breastfeeding, several STIs can be passed from mother to kid (WHO, n.d; Barrow et al., 2020).

The phrase syndromic management will be defined. The detection of consistent, easily recognised signs and symptoms of STIs in order to guide treatment without the need of laboratory tests is referred to as syndromic management (WHO, n.d). Syndromic management is straightforward, ensures same-day treatment, and avoids costly or inaccessible diagnostic tests for patients who present with symptoms (Garrett, Osman, Maharaj, Naicker, Gibbs, Norman & Mindel, 2018). This method, which frequently depends on clinical algorithms, enables health personnel to diagnose a specific infection based on observable syndromes (Garret et al., 2018).

## 2.3 Theoretical Framework

A theoretical framework explains the important elements in a study, including the identification of interconnections amongst them, and analyses pertinent theories based on the review of the literature (Kivunja, 2018; Osanloo & Grant, 2016). According to Osanloo and Grant (2016) a robust theoretical framework directs the study, helping with the analysis, explanation, and provides the ability for a researcher to generalise findings convincingly. For this study, the major aim was to develop a Machine Learning model to aid in the control of STIs in Zimbabwe which may be adopted by health institutions for use. Therefore, the major concepts or elements for the research include, training algorithms to make inferences from past data for future interventions. Thus, the researcher used a combination of two theories as the basis for the study’s theoretical framework. The theories that were adopted are the Machine Learning Theory posited by Arthur Samuel in 1959 (Shi & Iyengar, 2020) and the Computational Learning Theory, which was put across by Leslie Valiant in 1994 (Wang & Hill, 2018).

### 2.3.1 The Machine Learning Theory

Arthur Samuel established the Machine Learning Theory in 1959 with the goal of finding answers to issues that cannot be handled solely by numerical means (Shi & Iyengar, 2020). The theory is concerned with the creation of programs capable of learning rules from data, adapting to changes, and improving performance over time. Machine Learning has grown critical as computers are expected to handle increasingly complicated issues and become more interwoven into people’s daily lives, in addition to being one of the original dreams of Computer Science (Murphy, 2022). The goal of Machine Learning Theory, also known as Computer Learning Theory, is to comprehend the basic principles of learning as a computational process. This topic aims to understand what capabilities and information are required to learn different types of tasks successfully at a mathematical level, as well as the underlying algorithmic principles involved in teaching computers to learn from data and improve performance with feedback. The objectives of this theory are to aid in the development of improved automated learning systems as well as to comprehend key concerns in the learning process (Boehmke & Greenwell, 2019). The theory is adopted in the current study and is still relevant as it is a building foundation to development of models that solve complex problems as in the case of managing STIs.

### 2.3.2 Computational Learning Theory

Leslie Valiant created the computational theory in 1994 in order to research machine learning algorithm design and analysis (Wang & Hill, 2018). Computational learning theory, often known as learning theory, is a subfield of machine learning that focuses on the design and study of machine learning algorithms in computer science. It aims to quantify learning issues by employing methods from theoretical computer science which involves determining how complex tasks are performed. Theoretical results in machine learning, according to Newth and McDonald (2021), are primarily concerned with supervised learning, a sort of inductive learning. An algorithm is provided with samples that are labelled in a useful way in supervised learning. The purpose of the supervised learning method is to optimize a performance metric, such as the number of errors made on new samples. Therefore, the theory is relevant to the current study as it guides in model building through a learning process given a set of data.

## 2.4 Control of Sexually Transmitted Diseases

Sexually transmitted illnesses are of important public health relevance because they mostly affect young individuals, carry stigma, facilitate transmission and acquisition of HIV infection, and have sequelae that impose a significant socioeconomic cost. Pelvic inflammatory disease (PID), infertility, ectopic pregnancy, persistent pelvic pain, cervical cancer, and urethral stricture are all complications of failing to diagnose and treat infections. Miscarriages, stillbirths, neonatal mortality, mental retardation, neonatal conjunctivitis, and pneumonia are all examples of the effects on foetuses and infants. According to the World Bank (2019), STDs, excluding HIV, are the second leading cause of healthy life loss in women aged 15–44 years, behind maternal mortality and morbidity. Despite this burden, many developing countries have given STDs a low priority. The majority of countries do not have an efficient STD control program (Gupta & Sharma, 2019).

Fortunately, information from recent research indicating that controlling STDs could significantly reduce the prevalence of HIV has boosted interest in STDs (Pevernagie, 2021; Bradshaw *et al.,* 2018). The primary goals of STD control are to prevent STD transmission, progression, and consequences. Detecting and treating disease by providing enough diagnostic and treatment facilities, as well as reducing infection consequences by providing early and effective treatment for both symptomatic and asymptomatic patients and their contacts, are examples of strategies to achieve these goals. One method for secondary STD prevention is syndromic case management. This study proposes an efficient method of carrying out syndromic case management, which is the use of a machine learning model.

### 2.4.1 The Rationale for Syndromic Management of STIs

The signs and symptoms of different STDs are not always consistent, making clinical diagnosis problematic. Only around 70% of single genital ulcer disease infections are detected accurately clinically (Garrett *et al.,* 2018). Only 40% of chancroid and 24% of syphilis infections were appropriately detected clinically (Maduna *et al*., 2019). Numerous clinicians regard laboratory-confirmed aetiological diagnosis to be scientific because it is commonly utilized in the management of many diseases. This method, which includes microscopy, culturing, and serology, is costly and may cause delays in diagnosis.

Many resource-poor countries lack laboratory support, or laboratory support is only available in urban areas, serving a small portion of the population. Patients may have to travel significant distances from rural health clinics to city-based specialists or laboratories. Many STD patients have several infections. The syndromic approach arose from the constraints of clinical diagnosis without tests and laboratory-based aetiological diagnosis.

### 2.4.2 Understanding the Syndromic Approach to Management of STIs

A syndrome is a collection of symptoms and signs that describe a medical condition. Extending this definition, syndromic management refers to a case management method in which clinical algorithms such as decision trees are utilized to address common presenting signs and symptoms (such as urethral discharge or genital ulcer). The symptoms chosen are fairly constant and straightforward to identify. The protocol provides treatment for the syndrome's most common biological causes. In the absence of laboratory support and in acknowledgment of the limitations of clinical diagnosis, treatment for genital ulcer disease (GUD) is given concurrently for the two most prevalent causes, chancroid and syphilis.

Vaginal discharge (VD), male urethral discharge, lower stomach pain, scrotal enlargement, and ophthalmia neonatorum are examples of STDs treated syndromically. Patient management includes education and counselling for future infection prevention, condom promotion, compliance promotion, and partner management (WHO, 2018). Recognition of common and consistent combinations of STD signs and symptoms, knowledge of the most common causative organisms for the various syndromes, knowledge of the socio-behavioural characteristics of people with STDs, knowledge of health-seeking behaviour of STD patients, co-operation of people with STDs' partners, local antibiotic susceptibility patterns, and data on drug availability are all requirements for effective STD management using the syndromic method (Garrett *et al.,* 2018).

### 2.4.3 The Efficiency of Syndromic Management of STIs

A study conducted by Masatu *et al.* (2022) showed successful use of syndromic case management among asymptotic women in Mwanza, Tanzania. In that study, areas that were randomly assigned to get improved STD case care saw a 42% decrease in HIV cases. Further examination of the data revealed that the intervention reduced the prevalence of active syphilis and symptomatic male urethral discharge by 30% to 50% (Chaponda et al., 2021).

Success stories of the use of syndromic management of STDs include a study by Maina et al. (2021) in Nairobi, Kenya which found a 92.9% diagnostic accuracy of syndromic management.  Khatun (2021) found a 62.3% decrease in vaginal discharge due to the use of syndromic management of STDs in Bangladesh. In Rwanda, Wall *et al*. (2021) investigated the sensitivity of three ways to diagnosing and treating genital ulcers. The approaches included a concomitant algorithm in which patients with genital ulcers were treated for both chancroid and syphilis without laboratory tests, a hierarchical algorithm in which treatment was determined based on the results of laboratory tests, and a clinical diagnosis-only algorithm in which treatment was indicated without laboratory tests. These therapies successfully controlled 99 percent, 82.1 percent, and 38.3 percent of chancroid and syphilis patients, respectively. Karnad *et al.* (2018) also supported that a simple syndromic approach should be used for case management in situations where no laboratory support is available and chancroid or syphilis are the most common causes of genital ulcers.

The syndromic diagnosis of sexually transmitted diseases is still widely recognized as the most practical, feasible, and cost-effective diagnostic tool in resource-limited settings. Cheng, Paintsil and Ghebremichael (2020) assessed the diagnostic accuracy of syndromic versus laboratory testing of STIs among 794 men randomly selected from the Moshi district of Tanzania. The syndromic approach had a 82.8% probability of correctly identifying STIs in study participants. Though the syndromic case management is still useful in the control of STDs, some studies recommend that the approach has to be updated and improved upon, in order to ensure its effectiveness (Verwijs *et al*., 2019). Thus, the current study seeks to improve on syndromic approach to control and monitoring of STDs in the country by developing a machine learning model that is able to categorise STDs based on their symptoms so that there is effective diagnosis of the diseases.

## 2.5 Machine Learning

Machine learning has come a long way in the last two decades, from a laboratory curiosity to a real technology with widespread commercial use (Jordan & Mitchell, 2015). More interestingly, machine learning has emerged as the preferred method for producing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications in artificial intelligence (AI) (Mahesh, 2020). Many AI system developers now acknowledge that, for many purposes, training a system by providing it examples of desired input-output behaviour can be significantly easier than programming it manually by anticipating the required response for all potential inputs. Machine learning has had a widespread impact in computer science and a variety of businesses that deal with data-intensive issues, such as consumer services, defect diagnostics in complex systems, and logistics chain control. According to El Naqa and Murphy (2015), machine-learning approaches have been created to evaluate high-throughput experimental data in unique ways, resulting in a similarly broad spectrum of effects throughout empirical sciences, from biology to cosmology to social science.

Ali (2020) conducted a study on a comparative study of cancer detection models using deep learning and noted that machine learning is part of artificial intelligence which is described as a computer which works like the human mind and gather data with a logical construct. Maharjan (2020) denotes that machine learning can be categorized into unsupervised learning, supervised learning, and reinforcement learning.

### 2.5.1 Supervised Learning

Supervised learning is a sort of machine learning in which machines are trained using well-labeled training data and then predict the output based on that data (Murphy, 2022). The labelled data indicates that some of the input data has already been tagged with the appropriate output. Models are trained using a labelled dataset in supervised learning, where the model learns about each category of input. The model is tested using test data (a subset of the training set) when the training phase is completed, and it then predicts the output. There are steps taken in supervised machine learning which are: determining the type of training dataset, collecting the labelled training data, splitting the training dataset into training dataset, test dataset, and validation dataset, determining the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output, determining the suitable algorithm for the model, such as support vector machine, decision tree, executing the algorithm on the training dataset and evaluating the accuracy of the model by providing the test set. Supervised learning is made up of regression and classification. The current study adopts classification particularly logistic regression in categorising, detecting and control of STDs in Zimbabwe.

### 2.5.2 Unsupervised Learning

Unsupervised learning is the polar opposite of supervised learning, as the name implies. One does not provide the machine any training data in this situation. Without any labelled data, the machine must draw judgments and it is a little more difficult to put into practice than guided learning. Unsupervised learning is usually used for data clustering and anomaly detection (Hutter, Kotthoff & Vanschoren, 2019).

### 2.5.3 Reinforcement Learning

Reinforcement learning differs from other types of machine learning in several ways (supervised and unsupervised). The relationship between data and machine is also distinct from that of other machine learning kinds. The machine learns from its mistakes via reinforcement learning (Mahesh, 2020). One creates a specialized environment in which the machine can carry out a set of tasks and it will now learn through trial and error. Figure 2.1 presents a machine learning framework.

|  |
| --- |
|  |

##### Figure 2.2: Machine Learning Framework

*Source: Burkov (2019).*

## 2.6 Machine Learning in the Health Sector

While the application of machine learning and artificial intelligence in medicine dates back to the early days of the profession (Ahmad Eckert & Teredesai, 2018), there has been a push in recent years to recognize the need for healthcare solutions that are powered by machine learning. As a result, academics believe it is only a matter of time until machine learning becomes commonplace in healthcare (Shailaja, Seetharamulu & Jabbar, 2018). Despite the fact that machine learning (ML) has been recognized as valuable in healthcare, there are still significant barriers to its widespread implementation.

One major stumbling block is the opacity, or black box character, of many machine learning algorithms. There is some hesitancy in deploying such models in crucial use cases, such as clinical decision making, because the cost of model misclassification is potentially substantial (Bhardwaj, Nambiar & Dutta, 2017). Predicting patient risk of sepsis (a potentially life-threatening response to infection), predicting a patient's likelihood of readmission to the hospital, and predicting the need for end-of-life care are just a few examples of possible high-stakes applications of machine learning algorithms in healthcare. The end user can then probe, analyse, troubleshoot, and even improve the machine learning system using interpretable ML. In such cases, interpretable machine learning models have a lot of potential and demand. End users, such as clinicians, can employ interpretable machine learning models to evaluate the model before taking action. Interpretable machine learning systems give consumers reasons to accept or reject forecasts and recommendations by explaining the reasoning behind them, which is especially important in the health industry where saving lives is vital (Chen *et al*., 2021).

## 2.6.1 Artificial Neural Networks in Health Management

The use of machine learning in the heath sector has gained much attention in the recent decade both in developed economies and though with a lag in developing countries. This increased use of machine learning to improve health outcomes through effective diagnosis has been accompanied with success, even though improvements have to be made. Widman *et al.* (2018) conducted research on the use of technology based innovations in reducing STIs among youths in the USA and found that technology based interventions were more effective in increasing sexual health knowledge compared to control programs. In support of the effectiveness of machine learning models, Bao *et al*. (2019) in their study puts that machine learning even outperforms simple regressions in predicting STDs diagnosis where the top 10 predictors in the study collectively explained 62.7% to 73.6% of variations in prediction. Hence, machine learning falls under technology based interventions which the current study seeks to use in categorising, predicting and diagnosis of STDs in Zimbabwe. The intention is to improve diagnosis of these STDs in the country for improved health and wellbeing of the citizens.

Through real-time data collecting and processing, technology systems have also helped to improve the quality of healthcare services. Queralta *et al.* (2019) suggested a system architecture with integrated artificial intelligence for health monitoring activities that incorporates Edge and Fog computing, low power wide area network (LPWAN) technologies, Internet of Things (IoT), and deep learning algorithms. The paper used a use case of fall detection using recurrent neural networks to demonstrate the feasibility and usefulness of this design. In fall detection, the system used inertial data as input and obtained an average precision of over 90% and an average recall of over 95%. Such high predictive prowess of machine learning model was also found by Chang *et al.* (2018) in predicting drug effectiveness from cancer genomic signature which was 98%. This indicates high ability of recurrent neural network models in monitoring the health system. The current research therefore uses RNN in developing a machine learning model that monitors STDs in the country.

Elder *et al*. (2021) used a Bayesian Additive Regression Trees algorithm with a cross-validated area under the receiver operating curve of 0.75, which is the best-performing algorithm of any single approach. The study examined structured EHR data on patients 15 and older who received an incident STI diagnosis in eastern Massachusetts between 2008 and 2015. Using more than 180 different EHR factors, machine learning methods were employed to model the chance of getting 1 or 2 additional STI diagnoses within 365 or 730 days after the initial diagnosis. The sensitivity and positive predictive value (PPV) of a predictive probability threshold with a sensitivity of 91.5 percent had a matching PPV of 3.9 percent, according to receiver operating curves for this method. With a PPV of 29.5 percent, a higher threshold had a sensitivity of 11.7 percent. The present study makes use of a larger data set, which is critical for enhancing the model's accuracy in predicting STDs in health care facilities.Machine Learning Models still prove to be of paramount importance in the health sector as they have correctly predicted the occurrence of diseases like cervical cancer among women. Asadi, Salehnasab and Ajori (2020) used supervised algorithms of machine learning for the prediction of cervical cancer in Iran. The accuracy, sensitivity, specificity, and AUC of the Quest and C&R trees were 95.5%, 90.5%, 100%, and 95.20%, respectively, while RBF 95.5%, 90%, 100%, and 91.5%, SVM 93.3%, 90.5%, 95.8%, and 95.8%, and MLP 90.9%, 90%, 91.7%, and 91.5%. The study showed the influence of machine learning models in the health sector that health institutions have to adopts to improve the quality of their services.

In the United States, Mackey *et al*. (2020) employed unsupervised machine learning to detect self-reporting of symptoms, testing access, and recovery from Covid-19. Despite the fact that the study did not focus on STDs, it is nevertheless significant because it detects Covid-19 based on symptoms, which is similar to the current study's goal of detecting STDs based on symptoms. The results from the study demonstrated that machine learning had a positive impact on forecasting the occurrence of health illnesses based on symptoms. Similarly, Ni *et al.* (2020) also applied a deep machine learning in characterizing Covid-19 pneumonia in chest CT images and the model achieved sensitivity of 95% and the algorithm showed excellent performance in detecting COVID-19 pneumonia on chest CT images compared with resident radiologists. The current work implements an RNN-based supervised model.

Healthcare firms are using ML techniques like ANNs to enhance care delivery while lowering costs. ANN are increasingly being utilised to influence managerial decisions in health care. Artificial intelligence is at the center of new technologies that have the potential to deliver affordable and appropriate health care in real-time. With the rapid adoption of artificial intelligence to make increasingly complex decisions across various industries, a plethora of solutions capable of addressing these health care management challenges exist; however, there is a lack of guidance on selecting appropriate methods tailored to the health care industry. Health care institutions are taking advantage of analysing large sets of routinely collected Big Data information in order to improve service and reduce costs.

## 2.6.2 The Impact of Machine Learning on Detecting STIs

Xu *et al.* (2022), conducted a study on the use of machine learning based risk prediction tool for HIV and STI acquisition and found that using machine learning techniques improves risk prediction reliability in predicting the acquisition of STIs over the next 12 months. Petersen et. (2019) applied machine learning in diagnosis of gonorrhoea and the accuracy rate of prediction was 92.9% with 100% precision. These findings that machine has positive impact on prediction and is highly effective in the health sector particularly for the management of STDs.

Machine learning algorithms are becoming increasingly popular in HIV and STDs research, as they can manage a larger number of covariates in a large dataset, handle complex interactions between predictors and outcomes, and achieve high accuracy (Asadi *et al.,* 2020). Machine learning algorithms may be used to unearth useful predictors in social media and Internet search data to construct prediction models of HIV diagnosis and incidence, as well as the number of syphilis cases, according to recent studies (Bao *et al.,* 2021; Medland *et al.,* 2021). Machine learning approaches have also been used to predict HIV progression and ART resistance, and various research studies have constructed and validated STD diagnosis prediction models in clinical settings using machine learning approaches since 2019. This has demonstrated the utility of machine learning models in the diagnosis of STDs, and the current project aims to construct a model that can aid in the classification and diagnosis of STDs in Zimbabwe.

Machine learning models have also proved to correctly identify patients at high risk of STD (HIV) acquisition. Marcus *et al.* (2019) conducted research on the use of electronic health record data and machine learning in identifying candidates for HIV pre-exposure and found that machine learning models were able to efficiently predict high risk of HIV infection. The findings were supported by a study by Drew *et al.* (2020) where machine learning was successful in diagnosis of bacterial vaginosis in maternal hospitals. The current study also uses a machine learning model in diagnosis of STDs in Zimbabwe.

## 2.7 Factors That Hinder the Use of Machine Learning On STI Screening

There are several factors that hinders the adoption and effective use of machine learning in the health sector in most developing countries which includes lack of experience model developers, high sensitiveness of the sector, among other factors (Salazar *et al.,* 2021). Managing of human lives using models that predicts the probability of occurrence or existence of a disease requires extra care for it affects human life in some cases leading to deaths of patients (Ferdous, Debnath, & Chakraborty, 2020). This usually happens in cases where a model gives a wrong prediction that would result in giving of wrong prescription of drugs to patients.

In addition, lack of required skills to develop models with high accuracy rates hinders the use of machine learning models in the health sector. Developing machine learning models in the health sector requires expects that are able to build efficient models (Marcus *et al.,* 2020). Most developing countries lack the expertise to do so, though the teaching of data science and data analytics is being embraced at a higher pace in the past decade. According to Goodman *et al.* (2020), machine learning models that are highly accurate need big data sets that are used to train the models of which data issues are challenging in most developing countries, Zimbabwe as an example. Therefore, lack of skills coupled with lack of data has been a challenge in use of machine learning models in the health sector which has affected the effectiveness of efforts to improve health outcomes (Bhardwaj *et al*., 2017).

Another issue observed was a loss of data control (Salazar et al., 2021). Once data are viewed or shared, they will leave the original data holder's (data controller) information system and become outside of their control unless particular data stewardship and processing safeguards are in place. Individuals who contribute their data and consent to its re-use and dissemination are treated similarly. Individuals and data proprietors then lose control over how their data is reused. They must rely only on law enforcement and redress to object to or reject such usage. The risks of losing control multiply when data is shared downstream across numerous tiers, especially when these tiers are located in different jurisdictions. Therefore, with such a challenge most health institutions would rather not share their data and keep it “safe”, which means that the initiative will not work.

## 

## 2.8 Summary

This chapter discussed syndromic management of STIs and definitions of machine learning and the different types of ML in existence. The chapter also discussed about theoretical and empirical literature based on research objects pertaining development of ML model using ANN, that predicts STIs and testing on the efficiency of the model. The empirical literature also covered on the factors that hinders the use machine learning in the screening of STIs in Zimbabwe. The next chapter looks at the methodology used in the study.

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.1. Introduction

This chapter describes in detail the methodologies employed in the study. Research methodology is the specific procedures or techniques used to identify, select, process, and analyse information about a topic. In a research paper, the methodology section allows the reader to critically evaluate a study's overall validity and reliability. To meet the objectives of this research, a number of research techniques were carefully selected and used with the goal of coming up with the pertinent findings. This chapter will explain the research philosophy, and the data analysis procedure. The chapter will provide insights on how the data was prepared, cleaned and transformed into data that was used in creating the Machine Learning model.

## 

## 3.2. Research philosophy

A research philosophy is a set of beliefs about how data about a phenomenon should be collected, analysed, and used (Saunders et al., 2018). In research, various research philosophies are employed and according to Saunders et al. (2018), the major ones are positivism and interpretivism. Positivist research philosophy claims that the social world can be understood in an objective way. In this research philosophy, the scientist is an objective analyst and, on the basis of it, dissociates himself from personal values and works independently (Žukauskas, Vveinhardt, & Andriukaitienė, 2018). For this research the scholar selected the positivism as the research paradigm.

According to Alharahsheh and Pius (2020) and Park, Konge and Artino (2020) it is the belief agreed upon by positivists that reality is stable and can be observed and described from an objective viewpoint.

This research is suited well to be carried out under the positivist paradigm since the positivism paradigm believe that one does not need to interfere with the phenomena being studied. They contend that phenomena should be isolated and that observations should be repeatable. Study results are typically observable and quantifiable in these types of studies. Positivism refers to the philosophy of the natural scientist and includes dealing with an observable social truth to generalise (Saunders & Lewis, 2017).

## 3.2. Research Design

A research design is the general strategic approach chosen by a researcher to incorporate the various parts of the research in a logical format, making sure that the research problem is best answered. The research design acts as a guideline for data collection, measurement, and analysis.

### 3.2.1 Case Study Research Design

The case study research design was chosen by the researcher from among the various research designs available. The phenomenon being studied is referred to as the “case”. A case study is a research method used to develop an in-depth, multifaceted understanding of a complex issue in its real-world context. A case study can be defined in a variety of ways, with the central tenet being the need to thoroughly investigate an event or phenomenon in its natural context. This study lends itself well to the case study methodology because it seeks to investigate the causes and effects of a real-life situation, STI diagnosis. Case studies, according to Yin (2018), can be used to explain, describe, or explore events or phenomena in their everyday contexts, providing an understanding and explanation of causal links and pathways resulting from a new policy initiative or service development. The case study method works well for gathering information on more explanatory 'how, what, and why' questions (Yin, 2018).

## 

## 3.3. Research Approach

For its strong points that complement the study being conducted, the quantitative research approach was selected in this research. The research strategy was utilized since it facilitates in the interpretation and understanding of social phenomena related to the case. According to Ferdous et al. (2020) Machine Learning approaches are essentially quantitative methods and models of analysing qualitative data systematically. Machine Learning algorithms work best with large data sets, when they have thousands, or millions, of sources in which to identify patterns.

## 3.4. Population and Sampling

A population is the entire group that a researcher wants to draw conclusions about. A sample is the specific group from where the data for the study will be gathered from. The size of the sample is always less than the total size of the population.

### 3.4.1 Population

A research population is generally a large collection of individuals or objects that is the main focus of a scientific query (Etikan, and Bala, 2017). It is for the benefit of the population that studies are carried out. A research population is also known as a well-defined collection of individuals or objects known to have similar characteristics. All individuals or objects within a certain population usually have a common, binding characteristic or trait (Etikan, and Bala, 2017). The population for study comprised of recordings of symptoms of all STI related diagnosis at a clinic from Harare’s high density clinic from 1 January 2015 up to 27 July 2020. The total population of the research is represented by the size of the dataset which is 15256 rows or records, which correspond to the entries made.

### 3.4.2 Sampling

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual. In Machine Learning and Big Data Analytics sampling is especially effective when working with data sets that are too huge to fully analyze, such as in big data analytics applications or surveys. Most of the times, Big Data Analytics involves working with massive amounts of data on computationally limited machines. Selecting and evaluating a sample size is much more convenient and cost-effective than polling the entire data set or population. However, various Big Data Analytics Tools like Hadoop have made it easier to handle the voluminous data and analyse it (Gunawan et al., 2020). For Machine Learning projects, this can mean dealing with all the available data at once instead of using the traditional sampling methods.

This research utilised Google Colab a cloud-based Software-as-a-Service (SaaS) Machine Learning platform which can be used for research, collaboration and training of models without the need for a powerful machine (Gunawan et al., 2020). Therefore, there was no sampling done for this research. The entire population was used in all the steps of the study.

**3.5. Data Analysis Procedure**

This section will present the CRISP-DM methodology used in the study. The researcher used the CRISP-DM methodology for Machine Learning and Data Mining modelling as part of the research strategy. CRISP-DM stands for cross-industry process for data mining, a technique used in the creation of Machine Learning models, according to Huber, Wiemer, Schneider, and Ihlenfeldt (2019). The CRISP-DM process is made up of the following steps: business comprehension, data comprehension, data pre-processing, model construction, evaluation, and implementation (Hotz, 2022).

##### Figure 3.1 CRISP-DM process (source: https://www.datascience-pm.com/crisp-dm-2)

### 3.5.1 Business Understanding

From the CRISP-DM methodology the first aspect is to understand the business requirements. This includes determining the business objectives, assessing the current situation, determining the goals of the machine learning model and the creation of a project plan. The business understanding was carried out by first analysing the statement of the problem and the need to find alternative STI case management strategies that are not only efficient but also cost-effective.

The researcher therefore, spent some time at the clinical facility understanding the various ways that were used to diagnose STIs, the way the data were captured and used and also understanding the cost in terms of time and money it takes for the tests to happen. With the business understanding the researcher then also truly understood the need for a ML model that utilises the syndromic case management approach to increase the prevention of STDs by making clinical decisions based on a patient's symptoms and signs.

### 3.5.2 Data Understanding

The second phase under the CRISP-DM methodology is known as the Data Understanding phase. The activities that are come under this stage are designed to let the researcher get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. The data is from a clinic in one of Harare’s high density suburbs for the period January 2015 to end of July 2020. The primary data were contained in a number of volumes of which had initially been capture by hand. These were later retroactively captured using Microsoft Excel after funding had been given by a certain non-governmental organisation (NGO) for data capturers.

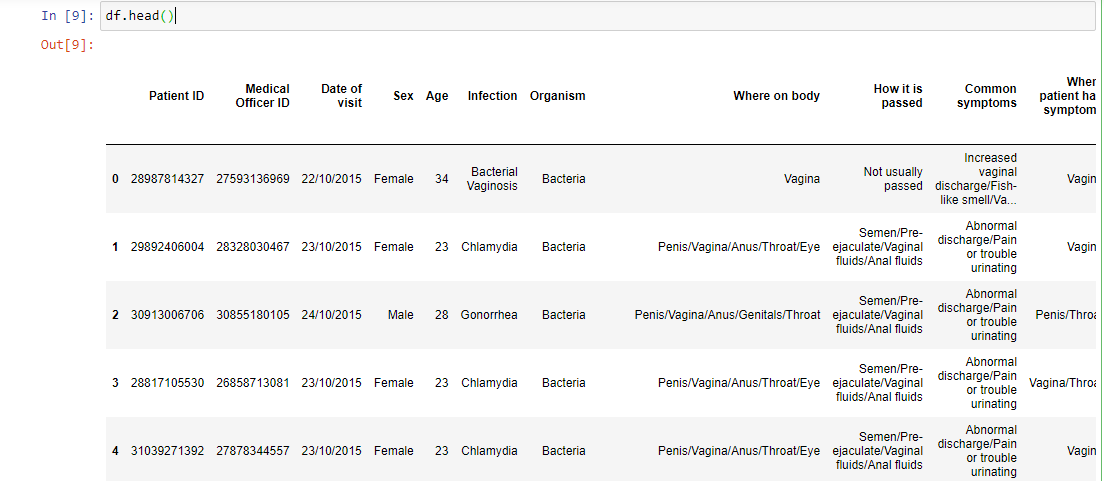
The data were captured in two Excel spreadsheets. The first spreadsheet had the patient details like age, sex, patient identity number (generated at the clinic), days the patient visited, the symptoms, the diagnosis and the actual treatment given to the patient. The second spreadsheet had the diseases, treatment and symptoms which were coded. The two spreadsheets were amalgamated to form one dataset. This resulted in a dataset with 21 variables. Some columns were removed as they were deemed redundant. These were, the general disease symptoms, the way each STI can be passed on, the general treatment, the organism that causes the infection and the general body area where the symptoms appear. The resultant dataset had 17 columns or features as shown in Table 3.1.

##### Table 3.1. Feature Description

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Feature Name** | **Description** | **Possible values** | **Data Type** |
| 1 | Patient ID | The ID that a patient was given | - | String (Python uuid) |
| 2 | Date of visit | Date of visit | - | Datetime |
| 3 | Sex | Gender of patient | 2 | String |
| 4 | Age | Age of patient when disease was first diagnosed | - | Integer |
| 5 | Medical Officer ID | The id of the medical officer who attended every visit | - | String (Python uuid) |
| 6 | Infection | The disease as diagnosed | 22 | String (categorical) |
| 7 | Organism | Cause of disease | 4 | String (categorical) |
| 8 | Where on body patient had symptoms | Body part affected | 8 | String (categorical) |
| 9 | How it is passed | Way of transmission | 5 | String (categorical) |
| 10 | Patient symptoms | Array of symptoms for each disease | 70 | String (Array of strings) |
| 11 | Minimum time for symptoms to show | The earliest time for symptoms to show | - | Integer |
| 12 | Minimum time for symptoms to show | The maximum expected time for symptoms to show | - | Integer |
| 13 | Actual time | The actual time each patient had symptoms show up | - | Integer |
| 14 | Tests | Tests carried out to ascertain the disease for each patient for each disease | 5 | String |
| 15 | Availability of tests | When tests results were made available | - | Integer |
| 16 | Symptoms readily available | Whether or not symptoms were easily seen | 2 | Boolean |
| 17 | Treatment | Treatment given a patient for the disease | 8 | String |

#### 3.5.2.1 Data description

The dataset was comprised of 17 as shown in Table 3.1. The features which are will be described here.



##### Figure 4.1. A screenshot of the dataset

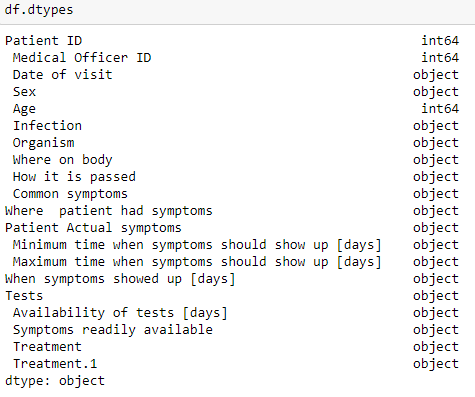
The *Patient\_ID* is the randomly created identification number that is given to a patient the first time they use the facility. This was done in order to hide the actual patient details from the data. The *date\_of\_visit* was the time each patient visited the facility. However, for the model each visit was taken to be separate from each other. The sex represented the gender of the patient which had two possible values “Male” or “Female”. The age feature captured the age of the patient when the disease was first diagnosed. The Medical Officer ID represented a randomly generated number to identify the medical officer who attended a patient every visit. The infection feature shows the disease as diagnosed and had 22 possible values, each representing an STI as diagnosed. This is the dependent variable in this dataset. The organism is the cause of disease and had 4 possible value including “bacteria” and “virus”. The other variable like where on body patient had symptoms and body part affected, indicated the body part which was affected and the body part which showed the symptoms. The variable ow it is passed indicated the way of transmission. The patient symptoms comprises of an array of symptoms for each disease many combinations and values for each disease. Minimum time for symptoms to show indicates the earliest time for symptoms to show and the maximum time for symptoms to show indicates the maximum expected time for symptoms to show, whereas the actual time shows the actual time each patient had symptoms show up. The variable tests show the type of tests that were carried out to ascertain the disease for each patient. The availability of tests variable shows the time the results were made available. The symptoms readily available indicates whether or not symptoms were easily seen by the patient and the treatment variable shows the type of treatment given a patient for the disease.

### 3.5.1 Data Preparation

The data preparation phase encompasses all activities that result in the final dataset being constructed from the initial raw data. Data preparation tasks are likely to be repeated multiple times and in no particular order. Tasks include record and attribute selection, data cleaning, attribute construction, and data transformation for modelling tools. Exploratory Data Analysis (EDA), Data Cleaning or Transformation, and Feature Selection were the three distinct steps in this phase. The first section is where the data is understood. The first phase is discussed in Section 3.5.1.1, and the second section is where the actual data cleaning takes place is discussed in Section 3.5.1.2. Feature selection is the phase where the features that are needed to create the model are created. The process is described in 3.5.1.3.

#### 3.5.1.1 Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data in order to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Some processes were carried out to determine the data that was going to be used in the research. First it was important to determine the data types of the features or variables. Determination of data types was done using the *dtypes()* function of the *pandas* library. The dataset has 17 variables or features, 11 are string data type are categorical features, five are numerical and the other one is a date. The variables and explanations are shown in Table 3.1.



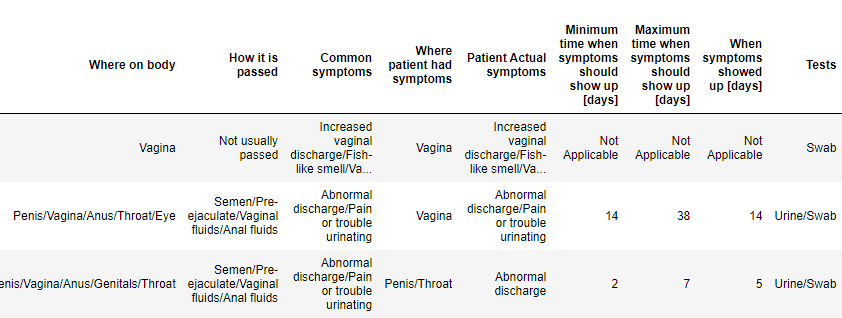
##### Figure 4.2. The feature data types according to pandas

The final aspect of the EDA was to check for missing data. Missing data can provide challenges in statistical analysis. Thus, the need to identify and deal with the missing data points. Using the *pandas* library testing for missing data is simplified. A simple piece of code was used to identify the number of missing data points.

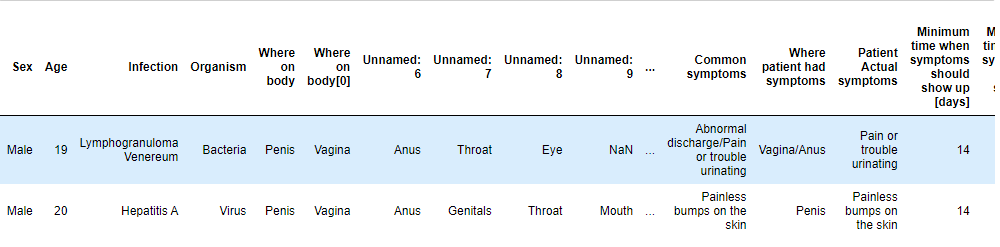
#### 3.5.1.2 Data Cleaning and Transformation

After the Exploratory Data Analysis, the next phase was used to now clean the data. The first was to see how missing data could be dealt with. The *Infection* variable did not have any missing data. Only a few rows had missing data under the *Organism*, *Tests*, and *Treatment* columns. The missing was inferred since most of the data was available for each missing category. A total of 492 entries which had been entered wrongly were corrected. A total of 347 entries in the gender column had “M” instead of “Female” and 145 had “M” instead of “Male”. These were all change to “Female” and “Male” respectively as this is how the data was being entered. In addition, seven entries in the gender column had *“l”* (small letter L) and five had *“O”* (Capital letter O) in them instead of zero (0). These were taken to mean 0 as female and 1 as male and were corrected thus. The date were entered in full, for example “2 March 2017”. The dates were reformatted to ISO 8601-1:2019 standards to avoid ambiguities. Thus the date mentioned before became “2017-03-02”. Thirteen dates had been entered with the year as 2000. This error was corrected and the assumption was that the year was meant to be 2020.

After that the symptoms columns and other columns that initially contained arrays were split so that each symptom would then be a separate. The initial dataset is shown before being split in Figure 3.3 and after the split in Figure 3.4.



##### Figure 3.3 the columns before each array is split



##### Figure 3.4 Dataset after splitting

#### 3.5.1.3 Feature Selection

The third step after EDA and data cleaning the dataset was yet again analysed to find information that would be useful for modelling. According to Das and Das (2017), feature selection removes redundant and discriminatory features (variables) from a data set. The first feature that was deemed unusable in the model was the customer identification. The *Patient ID* and the *Medical Officer ID* variables were dropped as they will not have any bearing on the modelling. Another reason that feature was dropped was because even though the identification has been hashed, there is a danger that the Machine Learning model may eventually learn the identity of the patient from other pieces of information that are collected. Thus, the feature was dropped because of these possible challenges. The *date\_of\_visit* feature was also dropped as the model could (theoretically in future) also use the visit patterns to identify a patient.

### 3.5.2 Modelling of the Algorithm

The next phase in the CRISP-DM methodology is the modelling stage. In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. For this study, Sci-Kit Learn, a library in Python that is used for scientific and machine learning modelling. It greatly simplifies the process of creating a Machine Learning model. This is the library that was used to split into two sets. The first one was used as the training set and the other one was used as the test data set. After the split, the training dataset had 12,204 rows and the test dataset had 1,553 rows with the validation dataset having 1,552 rows.

### 3.5.3 Selection and Evaluation of the Algorithm

To come up with the best model for the loan application prediction the researcher created models using five Machine Learning algorithms and tested each of them against the dataset. The models that were created were Logistic Regression, Decision Trees, Random Forest, Naïve Bayes and Artificial Neural Networks (ANN). These algorithms were identified as the best classification and prediction algorithms according to Uddin et al. (2019).

The metrics of each model were recorded, and the results are presented in Table 3.2. According to (Huang, Chai & Cho, 2020) the main metrics used to evaluate a classification model are accuracy, precision, and F1 score and recall. Accuracy is the measurement used to determine which model best identifies the relationships and patterns between variables in a dataset. Precision is the ratio between the True Positives all the Positives in the model, while recall is used to measure the model’s ability to correctly identify all the True Positives. The F1 score measures the model’s accuracy on a dataset in a binary classification model. These are the metrics that were measured in the test runs that were done to come up with the best ML algorithm to use in the study.

###### Table 3.2 Machine Learning Algorithms tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Logistic Regression | 92.65 | 85.43 | 90.32 | 87.69 |
| Decision Trees | 96.13 | 85.78 | 92.66 | 88.90 |
| Random Forest | 94.32 | 86.01 | 94.22 | 89.34 |
| Naïve Bayes | 89.67 | 83.12 | 89.90 | 88.38 |
| Artificial Neural Networks | 97.64 | 86.96 | 95.24 | 90.91 |

According to the tests that were run as shown in Table 3.2, Artificial Neural Networks (ANN) had the best performance of the algorithms. The tests indicate that using ANN would give an accuracy rate of 98.52 which is higher than the rest of the algorithms that were run. Artificial Neural Networks outperformed the others on all the metrics that were measured. Thus, this was the Machine Learning algorithm that was adopted for use in this study.

## 3.6 Summary

This chapter discussed the research methodology adopted in the study, including the research design, sampling techniques, data collection procedure and instruments. The selection of the model that was used for the study is explained in detail. The features that were used to create the Machine Learning model are discussed in depth and the justification for each variable is given. In addition, data analysis and presentation procedures were indicated. The next part of this thesis is to present the findings and analyse them in relation to the research questions and the research objectives. This will be done in the next section which is Chapter 4 of the research.

# CHAPTER 4: FINDINGS AND DISCUSSIONS

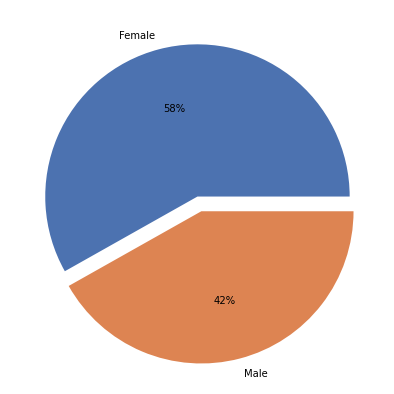
## 4.1 Introduction

This chapter will provide the findings that came out of the data. These findings will then be used to interpret the data. A discussion of the findings will be done which will provide insights into whether syndromic management of STDs in Zimbabwe can be done using Machine Learning algorithms. The findings will be presented in various types of graphs and tables for ease of interpretation by the reader.

## 4.2 Exploratory Data Analysis

The chapter will start by proving the results of the Exploratory Data Analysis. Various EDA techniques were run so that the researcher could get an understanding of the data he was working with.

### 4.2.1 Distribution by gender

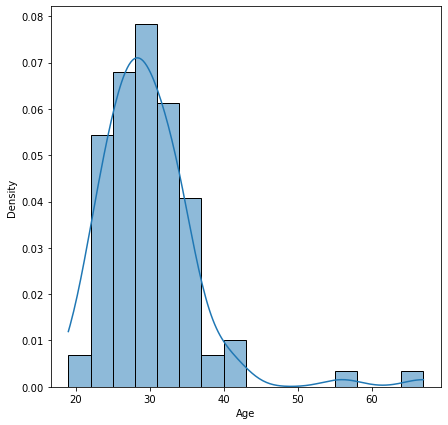


##### Figure 4.1 Dataset distribution by gender

The first EDA step was to make a comparison of the gender feature in the dataset. The dataset was composed of 6,429 males, making up 42% of the dataset, and 8,879 females who made up 58% of the dataset.

### 4.2.2 Age

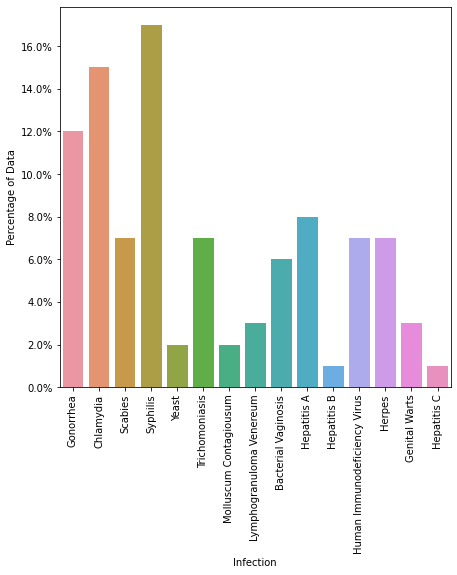
To get a clear picture of the age distribution for the dataset the ages were shown in a histogram in Figure 4.2.



##### Figure 4.2 Age distribution

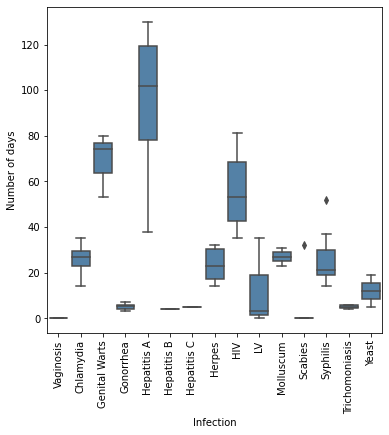
The ages range from 19-67 years. According to Figure 4.2 the ages are concentrated in the 25-35 year categories.

### 4.2.3 The distribution of the STIs



##### Figure 4.3 Distribution of STIs

The dataset showed that syphilis is had the highest number of cases with over 16% followed by Chlamydia with over 15% then followed by Gonorrhoea with 12% of the cases in the dataset.



##### Figure 4.4: The number of days against the infections

Figure 4.4 shows the time symptoms for each infection to show.

## 4.3 To determine factors that have a statistical significance to syndromic management.

The first objective for the research was to determine which factors have a statistical significance on the proposed model. Therefore, this research shall start by looking at that objective which seeks to establish factors that have a statistical significance on the Machine Learning models.

The research question that will provide the answer to this objective is, *“Which features of the patients’ data are relevant to syndromic management?”* To answer this question the features were all tested against the dependent variable. The categorical features were tested first and those which were shown to statistically significant were selected. The tests were done using Chi-square contingency from the *scipy* package. This shows the association and independence between two categorical features or variables.

###### Table 4.1 Statistical significance test for categorical features

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Statistical Value** | **p-Value** | **Degree of freedom** |
| Organism | 17.32120 | 0.004878 | 5 |
| Sex | 8.75 | 0.012588 | 2 |
| Where patient had symptoms | 14.57460 | 0.00282 | 12 |
| Patient Actual symptoms | 13.19217 | 0.00418 | 10 |
| When symptoms showed up | 15.57460 | 0.00364 | 4 |
| Symptoms readily available | 21.06753 | 2.7536e-05 | 2 |
| Treatment | 12.85451 | 0.00364 | 4 |
| Tests | 17.23456 | 0.0032415 | 3 |

According to the findings as illustrated in Table the features *organism, sex, where symptoms appear, patient symptoms, time when symptoms showed, availability of symptoms, treatment* and *tests* were all statistically significant. From the tests the findings also showed that the variable *treatment* also had a strong association (correlation) with the other features and therefore it was dropped from the final dataset as it was redundant.

The correlation tests for the numerical features were also done. As illustrated in Figure 4.2. The tests were carried out using ANOVA.

###### Table 4.2 ANOVA Tests for the numeric features vs the categorical dependent feature

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | sum\_sq | F-Statistic | p-value |
| Age | 1775.258343 | 4.292584 | 0.000012 |
| Minimum time symptoms show | 3515.741497 | 58.768772 | 0.514132 |
| Maximum time symptoms show | 1.131666 | 63.714662 | 0.270337 |
| When symptoms showed up | 67018.215286 | 28.862753 | 0.0424427 |
| Availability of test results | 2695.029412 | 33.639292 | 4.924427e-26 |

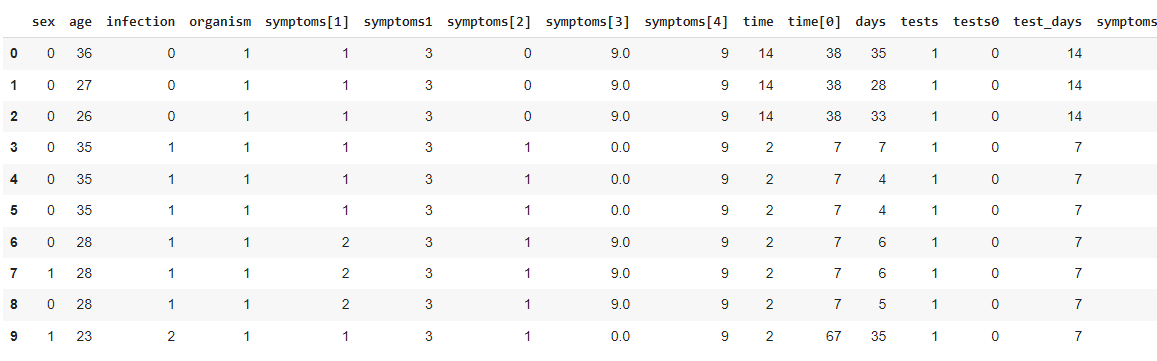
The findings show that the variables Age (p-value 0.000012), time for symptoms to show (p-value 0.0424427) and availability of test results (p-value 4.924427e-26) as being statistically significant.

## 4.4 Developing a Machine Learning model to accurately screen and diagnose patients based on their symptoms.

The second objective was to develop a machine learning model that accurately screens, and diagnose patients based on their symptoms using Artificial Neural Networks in Zimbabwe. Model development was done according to the steps given under CRISP-DM as explained in Section 3.5.2 in the methodology chapter.

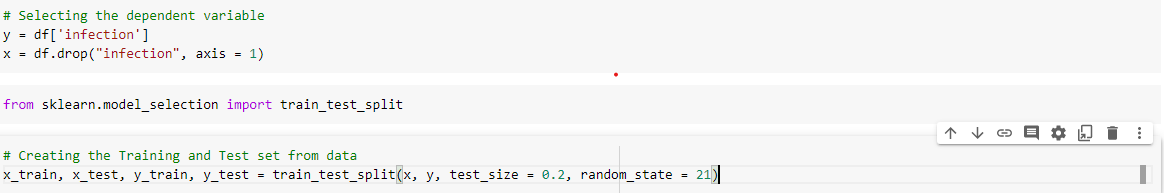
### 4.4.1 Modelling

First the features identified in 4.3 as being statistically significant were the ones that were taken as the final dataset. These are *organism, sex, where symptoms appear, patient symptoms, time when symptoms showed, availability of symptoms, treatment, tests, age*, *time for symptoms to show* and *availability of test results*. After that the multipart columns like symptoms, were split so that each symptom or part would be in its own column. The categorical features were then converted to numerical using the *get\_dummies()* function from the *pandas* library. A snapshot of the dataset is shown in Figure 4.5.



##### Figure 4.5: The final dataset

After the features were selected the data was split into the dependent and independent variables and then into training and testing data. The testing data was also split into testing and validation. All this was done using the SKLearn module of Python. Figure 4.6 shows the process.



##### Figure 4.6: Data Splitting

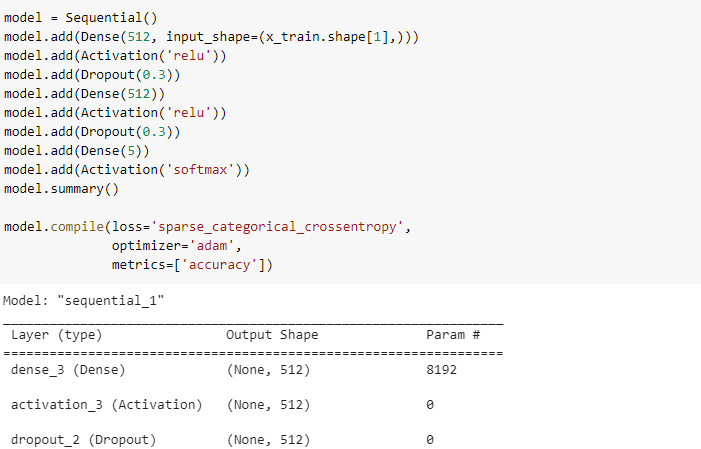
### 4.4.2 Model Preparation

The ANN model was created using the Keras Deep Learning Library. This has some parameters that are used to determine the model behaviour. For this model the following parameters were used:

###### Table 4.2 Model Hyperparameters

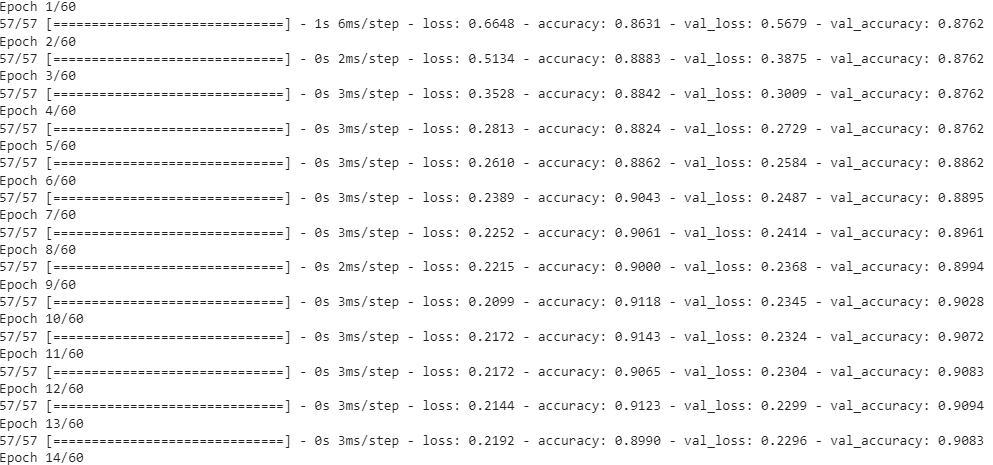
|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Description** |
| Epochs | 60 |  |
| Optimiser | Adam |  |
| Metrics | Accuracy | The metric to be evaluated. (Accuracy or Loss) |
| Input shape | X\_train.shape[1] | The size or number of features to be trained |
| Activation | Relu/sigmoid |  |
| Dropout rate | 0.5 |  |
| Dense | 512 |  |

Table 4.2 shows the Hyperparameters that were used to create this model. The modelling is shown in Figure 4.



##### Figure 4.6: Modelling

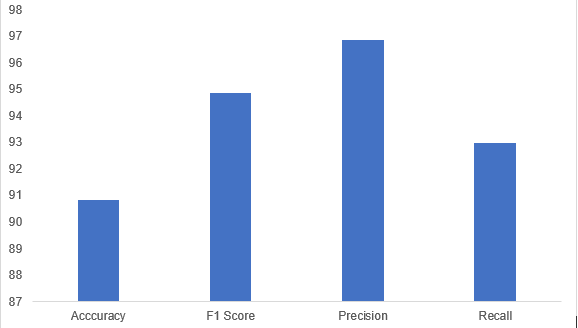
The model was created and the tested. The model was run using 60 epochs.



##### Figure 4.7: Model being run

After running the model the statistics were gathered and are shown in Figure 4.8

According to the model the accuracy rate attained was 90.83%, Precision 96.85, Recall 92.98% and f1 score 94.87%. The figures are shown in Figure 4.9.



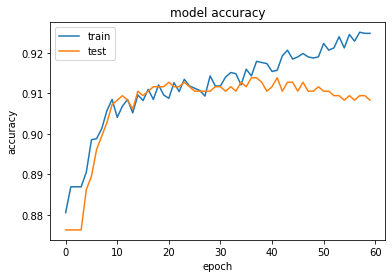
##### Figure 4.9: Model Results

## 4.5 To assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms using Artificial Neural Networks.

The third objective was to ascertain the effectiveness of the model in predicting the STIs based on the various features that were determined in 4.3. The effectiveness of the model is assessed using its accuracy, F1 Score, Precision and Recall rate. However, the major metrics used to asses a classification model are accuracy and the model loss.

**4.5.1 Accuracy of the model**

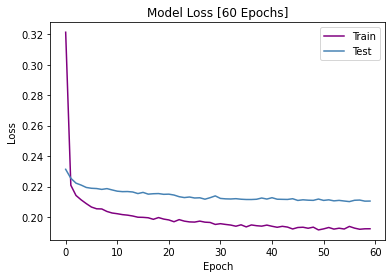
According to the model outcome the accuracy was 90.83%. This is a good rate. However, given that this is a health based model a higher accuracy would be better. Figure 4.10 shows how the accuracy rate grew with the epochs. It shows the accuracy rate for the training data and the test data.



##### Figure 4.10: Accuracy rate training vs testing

The accuracy is seen as being more accurate in training than in testing. In testing the model peaks at just over 91%, whereas in the training dataset it keeps increasing and ends over 92%. The final accuracy rate for the training dataset is under 89%.

**4.5.2 Value Loss of the model**



##### Figure 4.11: Model Loss

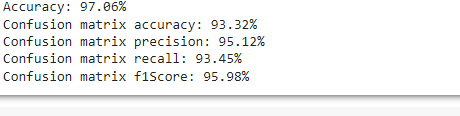
The model loss is shown in Figure 4.11. The illustration shows that over 60 epochs using training data the loss reduced from 32% down to below 20%. On the other the testing data was steady coming down from 24% to just under 22%.

### 4.5.3 Hyperparameter Tuning

Hyperparameter tuning is done to boost or improve model accuracy. For this model, the Hyperparameters that were identified in Table 4.2 were tweaked in order to come up with the optimal solution. The Hyperparameters were tuned by hand.

The following Hyperparameters were changed: activation from relu/sigmoid to relu/softmax, dropout rate was changed from 0.5 to 0.3, epochs were reduced from 60 to 50 and batch\_size from 10 to 20.

The resulting model metrics are shown in Figure 4.12.



##### Figure 4.12: Model statistics after Hyperparameter tuning

After Hyperparameter tuning the model metrics improved. The overall accuracy went up to 97.06%, whereas the recall rate improved to 93.45%, the F1 score improved to 95.98%.

## 4.6 Discussion of the findings

This section provides a discussion of the findings. The findings are discussed under each of the objectives. The primary objective of the study was to develop a Machine Learning model which can be used in syndromic management of STDs in Zimbabwe.

### 4.6.1 Objective 1: To determine factors that have a statistical significance to syndromic management.

The first sub-objective of the research was to determine the features that are statistically significant to syndromic management. Therefore, this research shall start by looking at that objective which seeks to establish factors that have a statistical significance on the Machine Learning models. The research question that will provide the answer to this objective is, *“Which features of the patients’ data are relevant to syndromic management?”*

These tests were run using ANOVA, and Chi Square. The testing of statistical significance was also done by some researchers in related studies (Gesesew, 2017; Gross, 2015). According to the findings, organism, sex, where symptoms appear, patient symptoms, time when symptoms showed, availability of symptoms, treatment, tests, age, time for symptoms to show and availability of test results are the major features that contribute to the prediction of a disease. These findings are in sync with literature which has identified most of them as being pertinent (Guy et al., 2015; Falasinnu, 2015; Graham et al., 2016; Enomoto, Noor & Widner, 2017; Vithalani, & Herreros-Villanueva, 2018). However, a number of researchers also include income and place or suburb of residence as other factors that are related to STI prevalence (Vithalani, & Herreros-Villanueva, 2018). The findings also showed that the different types of STIs are dependent on the gender of the patient. The other feature determining where symptoms appeared was partially dependent on gender.

According to the findings more females than males report or visit public medical facilities for assistance. This is shown in different studies where it is suggested that male patients may opt for private clinic or private treatment of STIs due to various stigmas (Huang et al., 2015). The finding shows that even though numerically females are more the diffence is not statistically significant when sex is taken alone. It becomes significant and important when other factors like symptoms and organism causing the infection are considered. This is due to the biological differences in male and female beings (Reidy et al., 2016; Reisner et al., 2016).

According to literature age is one of the factors with a high correlation to STIs (Graham et al., 2016; Guy et al., 2015). According to the findings the age group that is most vulnerable to STIs is the 15-40 age group. This is in line with what various literature says as a number of studies indicate that the 15-35 year age group is where the most cases of STIs are reported (Falasinnu, 2015; Enomoto, Noor & Widner, 2017; Vithalani, & Herreros-Villanueva, 2018).

## 4.6.2 Objective 2: To develop a Machine Learning model to accurately screen and diagnose patients based on their symptoms.

The second sub-objective was to develop a Machine Learning model to accurately screen and diagnose patients. A total of five Machine Learning algorithms were tested against the dataset. Initial testing showed that all the ML models had an accuracy rate just above 90%. This is not far off from the 85% identified by Paintsil and Ghebremichael (2020) using manual methods. However, after selecting the best performing model which used Artificial Neural Networks, and then varying the different parameters (Hyperparameter tuning), the accuracy improved to over 97%.

The accuracy performed better than what some researchers had. Bao et al. (2021) had a number of models of which the best performing reached 88.56%. However, their research only focused on one gender which are males. The second objective of the study was therefore achieved.

### 4.6.3 To assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms.

The third and final sub-objective was to assess the effectiveness of the machine learning model that categorises and diagnose patients based on their symptoms. The model was created and trained against the data and then tested against data which it had not been exposed to.

The model statistics of 97% accuracy, recall rate of 93.45% and F1 score of 95.98% were higher than studies which had been done in similar studies Bao et al. (2021) (88.56%), Elder et al. (2021) (91.25%). The statistics show that the model can be used in syndromic management and according to the tests it is effective.

### 4.6.4 Hypotheses testing

After creating the Machine Learning model, the next step was to test the hypotheses that were formulated at the start of the research. These are:

: The use of Recurrent Neural Networks improves the accuracy of STD diagnosis in Zimbabwe.

: The use of Recurrent Neural Networks does not improve the accuracy of STD diagnosis in Zimbabwe.

According to the findings as indicated by the objectives, 1) A Machine Learning model was created busing Artificial Neural Networks and 2) the ML model was evaluated and seen to effective with a high accuracy. Therefore, the study fails to reject the null hypothesis and concludes that the use of Artificial Neural Networks improves diagnosis of STIs and improves syndromic case management.

## 4.6 Summary

This chapter provided an analysis of the findings. The findings and analysis were presented in different tables and graphs for easier interpretation. The findings established that the Machine Learning models can be created to accurately diagnose STIs and be used for syndromic case management. The next chapter will provide a summary of the research together with the recommendations that came out of the findings.

# Chapter 5: Summary, Conclusions and Implications

## 5.1 Summary of research findings

The aim of this study is to fill this gap by coming up with a machine learning model that utilises the syndromic case management approach to increase the prevention of STDs by making clinical decisions based on a patient's symptoms and signs. The study seeks to achieve the following research objectives.

The primary objective of this study was to develop a Machine Learning model which could be used for syndromic management of Sexually Transmitted Diseases in Zimbabwe. In order to be able to develop the model the researcher came up with some sub-objectives which were:

1. To determine factors that have a statistical significance to syndromic management.
2. To develop a machine learning model that accurately screens, and diagnose patients based on their symptoms in Zimbabwe.
3. To assess the effectiveness of the Machine Learning model that categorises and diagnose patients based on their symptoms in Zimbabwe.

The researcher then formulated a hypothesis for the research:

: The use of Recurrent Neural Networks improves the accuracy of STD diagnosis in Zimbabwe.

: The use of Recurrent Neural Networks does not improve the accuracy of STD diagnosis in Zimbabwe.

The researcher used the Computational Learning Theory as the theoretical framework to underpin the study. The study used a positivist research strategy and case study as the research design. The research used the Cross Industry Standard Process for Data Mining (CRISP-DM) to undertake the stages that are found in the whole data mining process. The population for study comprised of recordings of symptoms of all STI related diagnosis at a clinic from one of Harare’s high density clinic from 1 January 2015 up to 27 July 2020.

### 5.1.1Factors that have a statistical significance to syndromic management

According to the findings, organism, sex, where symptoms appear, patient symptoms, time when symptoms showed, availability of symptoms, treatment, tests, age, time for symptoms to show and availability of test results are the major features that contribute to the prediction of a Sexually Transmitted Diseases.

### 5.1.2 Developing a Machine Learning model that accurately screens, and diagnose patients based on their symptoms

These statistically significant features were then used to as the variables in the Machine Learning model. A total of five Machine Learning algorithms were tested against the dataset. Initial testing showed that all the ML models had an accuracy rate just above 89%. The best performing algorithm was Artificial Neural Networks which had a final accuracy rate of over 97% after various Hyperparameter tuning.

### 5.1.3 Assessing the Effectiveness of the Machine Learning Model that Categorises and Diagnoses Patients Based on their Symptoms

The Machine Learning model that was created had a high accuracy rate and maintained that on testing data. The findings show that the model could correctly predict the type of infection given the symptoms and the other statistically significant features. Therefore, the findings showed that the model was highly effective and can be used by the Ministry of Health to effectively diagnose STI infections.

## 5.2 Conclusions of the research

The research concludes that the use of Machine Learning models greatly enhances the diagnosis of STD’s in Zimbabwe. From the findings the research also concludes that the best model to be used in diagnosis of STDs is Artificial Neural Networks (ANN). The study also concludes that since all the objectives of the study were met, that is 1) A Machine Learning model was created using Artificial Neural Networks and 2) the model accurately predicted the infection and 3) the Machine Learning model was evaluated and seen to effective with a high accuracy, the study therefore, concludes that the null hypothesis cannot be rejected and the use of Artificial Neural Networks improves diagnosis of STDs and improves syndromic case management.

## 5.3 Implications of the research

The implications of the research on policy are indicated are identified as thus and explained in the following paragraphs.

### 5.3.1 Theoretical Implications

Regarding the theoretical contribution, this study adds to the body of knowledge on the use of the Computational Learning Theory. The study shows that the use of data to “teach” algorithms to make decisions and predictions is the future. The findings in the study can be used to buttress future studies and can also be used as a conceptual framework.

### 5.3.2 Policy Implications

The study showed that it is possible to implement the use of AI in diagnosis of STIs. Therefore, the Ministry of Health and Child Welfare should partner with donors and funders in order to equip clinics with equipment to capture STIs data. This would be the first step to creating an enabling environment for the use of AI and Big Data in syndromic management. The Ministry of Health should engage industry and academia and have their personnel trained in using Machine Learning in syndromic case management

### 5.3.3 Practical Implications

Regarding the practical implications, this study intends to stimulate a meaningful dialogue with administrative and healthcare workers about how AI might be used to improve work quality. It has the potential to help advance scientific studies in this area. Academia should be funded to enable them to carry out more research into the use of Machine Learning algorithms in the health case management.

## 5.4 Limitations of the study

The researcher came across a number of limitations that may have impacted the research. The first one was that for a significant part of the period under research the data were captured manually. This meant that the researcher lost some data since the cost of having the data captured retroactively was beyond the researcher’s capacity. Secondly access to such data was hard to come by. The researcher spend some time looking for a clinic with the data in electronic format until settling on the clinic used in the research. Those two challenges made it difficult to carry out the study given that the time frame for the research was only six months.

## 5.5 Areas for further study

Future research can be done in the following areas:

1. Identification of other factors which may be localised in the determination of STDs
2. Time series analysis of syndromic management in Zimbabwe using ANN and hybrid sliding window models
3. Identification of repeat infections of STDs using Artificial Neural Networks

# References

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, *11*(1), 189.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559-560).

Alfarraj, M., & AlRegib, G. (2019). Semi-supervised learning for acoustic impedance inversion. In *SEG Technical Program Expanded Abstracts 2019* (pp. 2298-2302). Society of Exploration Geophysicists.

Almugti, H. S., Al Hakeem, R. N., Alghamdi, A. M., Aldamen, S. A., Alfaifi, A. A., Algharbi, S., ... & Al Barakati, A. S. (2022). Assessment of Using the Syndromic Approach in Managing Patients With Sexually Transmitted Diseases Among the National Guard Primary Health Care Physicians, Jeddah City, Saudi Arabia. *Cureus*, *14*(1). e21502. doi:10.7759/cureus.21502

Asadi, F., Salehnasab, C., & Ajori, L. (2020). Supervised algorithms of machine learning for the prediction of cervical cancer. *Journal of biomedical physics & engineering*, *10*(4), 513-519.

Bao, Y., Medland, N. A., Fairley, C. K., Wu, J., Shang, X., Chow, E. P., ... & Zhang, L. (2021). Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection*, *82*(1), 48-59.

Barrow, R. Y., Ahmed, F., Bolan, G. A., & Workowski, K. A. (2020). Recommendations for providing quality sexually transmitted diseases clinical services, 2020. *MMWR Recommendations and Reports*, *68*(5), 10-24.

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta numerica*, *30*, 87-201.

Beksinska, M., Wong, R., & Smit, J. (2020). Male and female condoms: Their key role in pregnancy and STI/HIV prevention. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *66*, 55-67.

Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017, July). A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 236-241). IEEE.

Biswas, R. (2021). Outlining Big Data Analytics in Health Sector with Special Reference to Covid-19. *Wireless Personal Communications*, 1-12.

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with R*. Chapman and Hall/CRC.

Bradshaw, C. S., Horner, P. J., Jensen, J. S., & White, P. J. (2018). Syndromic management of STIs and the threat of untreatable Mycoplasma genitalium. *The Lancet Infectious Diseases*, *18*(3), 251-252.

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.

Chang, Y., Park, H., Yang, H.J., Lee, S., Lee, K.Y., Kim, T.S., Jung, J. and Shin, J.M., 2018. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, *8*(1), pp.1-11.

Chaponda, E.B., Bruce, J., Michelo, C., Chandramohan, D. and Chico, R.M., 2021. Assessment of syndromic management of curable sexually transmitted and reproductive tract infections among pregnant women: an observational cross-sectional study. *BMC pregnancy and childbirth*, *21*(1), pp.1-10.

Chawurura, T., Manhibi, R., Dijk, J. V., & Stam, G. V. (2019, May). eHealth in Zimbabwe, a case of techno-social development. In *International Conference on Social Implications of Computers in Developing Countries* (pp. 15-26). Springer, Cham.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, *4*, 123-144.

Cheng, Y., Paintsil, E. and Ghebremichael, M., 2020. Syndromic versus Laboratory Diagnosis of Sexually Transmitted Infections in Men in Moshi District of Tanzania. *AIDS Research and Treatment*, *2020*.

Chirenje, Z.M., Dhibi, N., Handsfield, H.H., Gonese, E., Barr, B.T., Gwanzura, L., Latif, A.S., Maseko, D.V., Kularatne, R.S., Tshimanga, M. and Kilmarx, P.H., 2018. The etiology of vaginal discharge syndrome in Zimbabwe: results from the Zimbabwe STI etiology study. *Sexually transmitted diseases*, *45*(6), p.422.

Curtis, H., Hoolaghan, T., & Jewitt, C. (2018). *Sexual health promotion in general practice*. CRC Press.

Drew, R.J., Murphy, T., Broderick, D., O'Gorman, J. and Eogan, M., 2020. An interpretation algorithm for molecular diagnosis of bacterial vaginosis in a maternity hospital using machine learning: proof-of-concept study. *Diagnostic microbiology and infectious disease*, *96*(2), p.114950.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham.

Elder, H. R., Gruber, S., Willis, S. J., Cocoros, N., Callahan, M., Flagg, E. W., ... & Hsu, K. K. (2021). Can machine learning help identify patients at risk for recurrent sexually transmitted infections? *Sexually Transmitted Diseases*, *48*(1), 56-62.

Enomoto, C., Noor, S., & Widner, B. (2017). Is social media to blame for the sharp rise in STDs?. *Social Sciences*, *6*(3), 78.

Falasinnu, T., Gilbert, M., Hottes, T. S., Gustafson, P., Ogilvie, G., & Shoveller, J. (2015). Predictors identifying those at increased risk for STDs: a theory-guided review of empirical literature and clinical guidelines. *International journal of STD & AIDS*, *26*(12), 839-851.

Ferdous, M., Debnath, J., & Chakraborty, N. R. (2020). Machine learning algorithms in healthcare: A literature survey. In *2020 11th International conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-6). IEEE.

Furusa, S. S., & Coleman, A. (2018). Factors influencing e-health implementation by medical doctors in public hospitals in Zimbabwe. *South African Journal of Information Management*, *20*(1), 1-9.

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques, and prospects. *International Journal of Information Management*, *50*, 206-216.

Garrett, N. J., Osman, F., Maharaj, B., Naicker, N., Gibbs, A., Norman, E., ... & Mindel, A. (2018). Beyond syndromic management: opportunities for diagnosis-based treatment of sexually transmitted infections in low-and middle-income countries. *PloS one*, *13*(4), e0196209.

Garrett, N.J., Osman, F., Maharaj, B., Naicker, N., Gibbs, A., Norman, E., Samsunder, N., Ngobese, H., Mitchev, N., Singh, R. and Abdool Karim, S.S., 2018. Beyond syndromic management: opportunities for diagnosis-based treatment of sexually transmitted infections in low-and middle-income countries. *PloS one*, *13*(4), p.e0196209.

Gesesew, H. A., Tesfay Gebremedhin, A., Demissie, T. D., Kerie, M. W., Sudhakar, M., & Mwanri, L. (2017). Significant association between perceived HIV related stigma and late presentation for HIV/AIDS care in low and middle-income countries: a systematic review and meta-analysis. *PLoS One*, *12*(3), e0173928.

Goodman, K., Zandi, D., Reis, A., & Vayena, E. (2020). Balancing risks and benefits of artificial intelligence in the health sector. *Bulletin of the World Health Organization*, *98*(4), 230.

Graham, S., Smith, L. W., Fairley, C. K., & Hocking, J. (2016). Prevalence of chlamydia, gonorrhoea, syphilis and trichomonas in Aboriginal and Torres Strait Islander Australians: a systematic review and meta-analysis. *Sexual health*, *13*(2), 99-113.

Gross, J. H. (2015). Testing What Matters (If You Must Test at All): A Context‐Driven Approach to Substantive and Statistical Significance. *American Journal of Political Science*, *59*(3), 775-788.

Gupta, V., & Sharma, V. K. (2019). Syndromic management of sexually transmitted infections: A critical appraisal and the road ahead. *National Medical Journal of India*, *32*(3), 147-152.

Guy, R., Ward, J., Wand, H., Rumbold, A., Garton, L., Hengel, B., ... & Kaldor, J. (2015). Coinfection with Chlamydia trachomatis, Neisseria gonorrhoeae and Trichomonas vaginalis: a cross-sectional analysis of positivity and risk factors in remote Australian Aboriginal communities. *Sexually transmitted infections*, *91*(3), 201-206.

Henderson, J. T., Senger, C. A., Henninger, M., Bean, S. I., Redmond, N., & O’Connor, E. A. (2020). Behavioural counselling interventions to prevent sexually transmitted infections: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA*, *324*(7), 682-699.

Hotz, N., 2022. *What is CRISP DM? - Data Science Process Alliance*. [online] Data Science Process Alliance. Available at: <https://www.datascience-pm.com/crisp-dm-2/> [Accessed 10 July 2022].

Huang, S., Tang, W., Zhu, Z., Lu, H., Tan, X., Zhang, B., ... & Chen, X. (2015). Higher prevalence of sexual transmitted diseases and correlates of genital warts among heterosexual males attending sexually transmitted infection clinics (MSCs) in Jiangmen, China: implication for the up-taking of STD related service. *PLoS One*, *10*(3), e0121814.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges* (p. 219). Springer Nature.

Irwin, C. E., & Shafer, M. A. (2021). Adolescent sexuality: Negative outcomes of a normative behaviour. *Adolescents at Risk*, 35-79.

Jong, E. C., & Stevens, D. L. (2021). *Netter's Infectious Diseases-E-Book*. Elsevier Health Sciences.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Karnad, D. R., Richards, G. A., Silva, G. S., & Amin, P. (2018). Tropical diseases in the ICU: A syndromic approach to diagnosis and treatment. *Journal of Critical Care*, *46*, 119-126.

Khatun, R., 2021. The Syndromic Management of Vaginal Discharge Using One-Day Combination Kit Therapy: A Randomized Controlled Trial. *Sch Int J Obstet Gynec*, *4*(11), pp.488-494.

Kumar, S., & Singh, M. (2018). Big data analytics for healthcare industry: impact, applications, and tools. *Big data mining and analytics*, *2*(1), 48-57.

Levy, S. B., Gunta, J., & Edemekong, P. (2019). Screening for sexually transmitted diseases. *Primary Care: Clinics in Office Practice*, *46*(1), 157-173.

Lowe, S., Mudzviti, T., Mandiriri, A., Shamu, T., Mudhokwani, P., Chimbetete, C. & Pascoe, M. (2019). Sexually transmitted infections, the silent partner in HIV-infected women in Zimbabwe. *Southern African journal of HIV medicine*, *20*(1), 1-6.

Machiha, A., Mugurungi, O., Tshimanga, M., Kilmarx, P.H., Nyakura, J., Shambira, G., Kupara, V., Lewis, D.A., Gonese, E., Barr, B.A.T. and Handsfield, H.H., 2018. The etiology of genital ulcer disease and coinfections with Chlamydia trachomatis and Neisseria gonorrhoeae in Zimbabwe: results from the Zimbabwe STI etiology study. *Sexually transmitted diseases*, *45*(1), p.61.

Mackey, T., Purushothaman, V., Li, J., Shah, N., Nali, M., Bardier, C., & Cuomo, R. (2020). Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study. *JMIR public health and surveillance*, *6*(2),19-35.

Maduna, L.D., Laumen, J.G., Radebe, O., Kock, M.M. and Peters, R.P., 2019. Failure of syndromic management due to drug-resistant Mycoplasma genitalium infection in South Africa: a case report. *International journal of STD & AIDS*, *30*(5), pp.519-521.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)[Internet]*, *9*, 381-386.

Maina, A.N., Mureithi, M.W., Ndemi, J.K. and Revathi, G., 2021. Diagnostic accuracy of the syndromic management of four STIs among individuals seeking treatment at a health centre in Nairobi, Kenya: a cross-sectional study. *The Pan African Medical Journal*, *40*.

Marcus, J. L., Hurley, L. B., Krakower, D. S., Alexeeff, S., Silverberg, M. J., & Volk, J. E. (2019). Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *The lancet HIV*, *6*(10), 688-695.

Masatu, E.S., Kajura, A., Mujuni, F., Chibwe, E., Nyawale, H.A., Rambau, P., Majigo, M., Mshana, S.E. and Mirambo, M.M., 2022. High prevalence of sexually transmitted infections among asymptomatic women opting for the intrauterine contraceptive device use in Mwanza, Tanzania: An urgent call for control interventions. *SAGE Open Medicine*, *10*, p.20503121221097536.

Medland, N., Fairley, C., Wu, J., Wu, X., Chow, E., Xu, X., ... & Zhang, L. (2021). P341 Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches.

Moore, S., Rosenthal, D., & Mitchell, A. (2020). *Youth, AIDS and sexually transmitted diseases*. Routledge.

Morales, A., Espada, J. P., Orgilés, M., Escribano, S., Johnson, B. T., & Lightfoot, M. (2018). Interventions to reduce risk for sexually transmitted infections in adolescents: A meta-analysis of trials, 2008-2016. *PloS one*, *13*(6), e0199421.

Mukayiranga, J., 2018. *Knowledge And Practice Regarding Syndromic Management Of Sexually Transmitted Disease Among Clinicians In Nyamasheke District, Rwanda* (Doctoral dissertation, Mount Kenya University).

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press.

Newth, D., & McDonald, S. (2021). A Computational Learning Approach for the Development of Karate Sequences. *Advances in Physical Education*, *11*(4), 503-512.

Nguyen, S. H., Dang, A. K., Vu, G. T., Nguyen, C. T., Le, T. H. T., Truong, N. T., ... & Ho, R. (2019). Lack of knowledge about sexually transmitted diseases (STDs): Implications for STDs prevention and care among dermatology patients in an urban city in Vietnam. *International journal of environmental research and public health*, *16*(6), 1080-1096.

Ni, Q., Sun, Z.Y., Qi, L., Chen, W., Yang, Y., Wang, L., Zhang, X., Yang, L., Fang, Y., Xing, Z. and Zhou, Z., 2020. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. *European radiology*, *30*(12), pp.6517-6527.

Osanloo, A., & Grant, C. (2016). Understanding, selecting, and integrating a theoretical framework in dissertation research: Creating the blueprint for your “house”. *Administrative issues journal: connecting education, practice, and research*, *4*(2), 7.

Petersen, J., Patel, S., Dalal, S. and Jhala, D., 2019. Laboratory-diagnosed gonorrhea and chlamydia infection rate in the veteran population: the Corporal Michael J. Crescenz Veteran Affairs Medical Center (CMCVAMC) Experience. *American Journal of Clinical Pathology*, *152*, pp.S136-S137.

Pevernagie, D. (2021). Future Treatment of Sleep Disorders: Syndromic Approach versus Management of Treatable Traits? *Sleep medicine clinics*, *16*(3), 465-473.

Pfennig, C. L. (2019). Sexually transmitted diseases in the emergency department. *Emergency Medicine Clinics*, *37*(2), 165-192.

Queralta, J. P., Gia, T. N., Tenhunen, H., & Westerlund, T. (2019, July). Edge-AI in LoRa-based health monitoring: Fall detection system with fog computing and LSTM recurrent neural networks. In *2019 42nd international conference on telecommunications and signal processing (TSP)* (pp. 601-604). IEEE.

Reidy, D. E., Brookmeyer, K. A., Gentile, B., Berke, D. S., & Zeichner, A. (2016). Gender role discrepancy stress, high-risk sexual behavior, and sexually transmitted disease. *Archives of sexual behavior*, *45*(2), 459-465.

Reisner, S. L., White Hughto, J. M., Pardee, D., & Sevelius, J. (2016). Syndemics and gender affirmation: HIV sexual risk in female-to-male trans masculine adults reporting sexual contact with cisgender males. *International journal of STD & AIDS*, *27*(11), 955-966.

Salazar, L. H. A., Leithardt, V. R., Parreira, W. D., da Rocha Fernandes, A. M., Barbosa, J. L. V., & Correia, S. D. (2021). Application of Machine Learning Techniques to Predict a Patient’s No-Show in the Healthcare Sector. *Future Internet*, *14*(1), 3-16.

Saranya, P., & Asha, P. (2019, November). Survey on big data analytics in health care. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 46-51). IEEE.

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 910-914). IEEE.

Shi, B., & Iyengar, S. S. (2020). *Mathematical theories of machine learning-Theory and applications*. Springer International Publishing.

Verwijs, M.C., Agaba, S.K., Sumanyi, J.C., Umulisa, M.M., Mwambarangwe, L., Musengamana, V., Uwineza, M., Cuylaerts, V., Crucitti, T., Jespers, V. and Van De Wijgert, J.H., 2019. Targeted point-of-care testing compared with syndromic management of urogenital infections in women (WISH): a cross-sectional screening and diagnostic accuracy study. *The Lancet Infectious Diseases*, *19*(6), pp.658-669.

Vithalani, J., & Herreros-Villanueva, M. (2018). HIV Epidemiology in Uganda: survey based on age, gender, number of sexual partners and frequency of testing. *African health sciences*, *18*(3), 523-530.

Wall, K.M., Nyombayire, J., Parker, R., Ingabire, R., Bizimana, J., Mukamuyango, J., Mazzei, A., Price, M.A., Unyuzimana, M.A., Tichacek, A. and Allen, S., 2021. Etiologies of genital inflammation and ulceration in symptomatic Rwandan men and women responding to radio promotions of free screening and treatment services. *Plos one*, *16*(4), pp.44-54.

Wang, C., & Hill, D. J. (2018). *Deterministic Learning Theory: For Identification, Recognition, and Control*. CRC Press.

Wang, R., Pizer, S. M., & Frahm, J. M. (2019). Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5555-5564).

Widman, L., Nesi, J., Kamke, K., Choukas-Bradley, S., & Stewart, J. L. (2018). Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. *Journal of Adolescent Health*, *62*(6), 651-660.

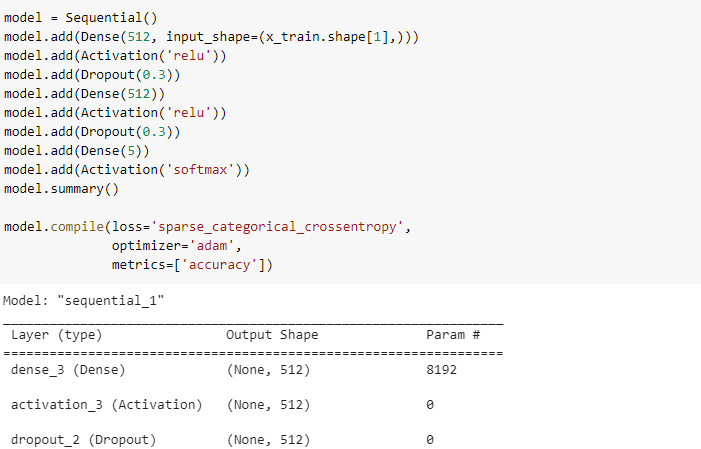
Žukauskas, P., Vveinhardt, J., & Andriukaitienė, R. (2018). Philosophy and Paradigm of Scientific Research. In P. Lukauskas, J. Vveinhardt, & R. Andriukaitien (Eds.), Management Culture and Corporate Social Responsibility. IntechOpen. <https://doi.org/10.5772/intechopen.70628>

# Appendix

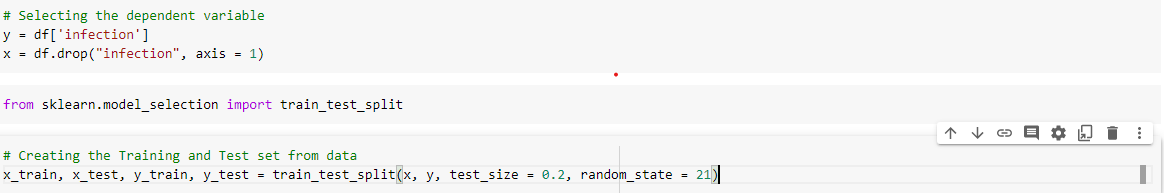
**Appendix A: Statistical significance test for categorical features**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Statistical Value** | **p-Value** | **Degree of freedom** |
| Organism | 17.32120 | 0.004878 | 5 |
| Sex | 8.75 | 0.012588 | 2 |
| Where patient had symptoms | 14.57460 | 0.00282 | 12 |
| Patient Actual symptoms | 13.19217 | 0.00418 | 10 |
| When symptoms showed up | 15.57460 | 0.00364 | 4 |
| Symptoms readily available | 21.06753 | 2.7536e-05 | 2 |
| Treatment | 12.85451 | 0.00364 | 4 |
| Tests | 17.23456 | 0.0032415 | 3 |

**Appendix B: shows Hyperparameters**



**Appendix C: Data splitting**



**Appendix D: Model results**

