**MACHINE LEARNING APPROACH IN THE PREDICTION OF MALARIA AND TUBERCULOSIS**

**SUPERVISOR**

**DR. MRS. ALIYU**

**CHAPTER ONE**

* 1. **INTRODUCTION**

Malaria and tuberculosis are deadly disease killing millions of people every year. Different countries of the world, governmental and non-governmental organizations including World Health Organization have taken it as a challenge to address the issue of deaths associated with tuberculosis and malaria. Extraction of data using data mining techniques and machine algorithms with data set containing necessary information of both malaria and tuberculosis would help develop a model that would aid in the prediction of malaria and tuberculosis. Decision tree and Naives bayes algorithm of machine learning would be adopted for the prediction respectively.

**1.2 BACKGROUND OF STUDY**

The main microorganism in tuberculosis is Mycobacterium tuberculosis (Bird, 2010). In this infectious disease, the microorganisms frequently penetrate into the lungs (through breathing) and are spread to the whole body through the blood circulatory system, lymphatic system or direct extension to other organs. Tuberculosis bacteria are spread into the air when contaminated person spit, talk, sneeze or cough. A person needs only to inhale a small number of these bacteria to become infected. There may be high risk of becoming infected if the intensity of exposure to the bacteria is high and for a long time, (Chiang, 2013). Typical outward indications of pulmonary tuberculosis include persistent cough, weight reduction, occasional fever, coughing blood and night sweats(Phillips, 2010). Tuberculosis evolves in the human body in two phases. The first phase occurs when someone who is subjected to micro-organisms from a contagious case of tuberculosis becomes infected (tuberculosis infection), and the second is when the infected person grows the illness (tuberculosis). Tuberculosis is a significant cause for illness and death worldwide, especially in Asia and Africa. It is an immense problem in most low income countries, and it is the single most frequent cause of death in children in particular. Based on a government health ministry statistics report in 2006 Varieties of machine learning based models have been proposed for performing clinical diagnostics. For instance, a standard instrumentation for building multivariate diagnosis models is discrimination analysis, specifically for operating on linear systems (Jacobsen el al, 2007). On the other hand, Malaria is a mosquito borne infectious diseases caused by a eukaryotic protist of the genus plasmodium. It is wide spread in tropical and subtropical regions, including parts of the American, Asia and African. Five species of the plasmodium parasite can infect humans, the most serious form of the disease is caused by plasmodium falciparum. Malaria caused by plasmodium vivax, plasmodium ovale and plasmodium malariae causes milder disease in humans that is not generally fatal. A fifth species, plasmodium Knowlesi, is a zoonosis that causes malaria in macaques but can also affect humans. Malaria deaths are frequently the result of delay in the diagnosis and treatment of the infection. Fever and septic shock are commonly misdiagnosed as severe malaria in Africa leading to failure to treat other life-threatening illness. Each year there are approximately 350-500 million cases of malaria, killing between one and three million people, the majority of whom are young children in Sub–Sahara African. Ninety percent of malaria –related deaths occurs in Sub-Sahara African. Half of the world’s population is at risk of malaria, and an estimated 243 million cases led to an estimated 863,000 deaths in 2008. Once a person develops malaria, the only means of reducing suffering and preventing death is by diagnosing and treating the disease.

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. (Friedman, 2010) Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task. It is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

In other words, it is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. (Christopher, 2010) Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. (Ju, 2019) The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., analyzing the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data mining uses machine-learning and statistical models to uncover clandestine or hidden patterns in a large volume of data.

The manual extraction of patterns from medical data has occurred for centuries. Early methods of identifying patterns in medical data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns (Kantardzic, 2013) in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

Hence, In this work, we would be using datamining techniques and machine learning methods and algorithms with the dataset containing important metrics for both malaria and tuberculosis in the prediction of such diseases.

**1.3 PROBLEM STATEMENT**

The high death rate of malaria and tuberculosis is fast increasing, which has demanded for adverse attention, common issue related to this is the delay in diagnosis of the disease in the human body. With data mining techniques, we would be able to predict if a patient has malaria or tuberculosis.

**1.4 AIM AND OBJECTIVES**

The aim of this work is to develop a model that would be able to predict malaria and tuberculosis diseases. To achieve the aim of this work, the following objectives would be considered.

1. To train the machine with malaria datasets.
2. To train the machine with tuberculosis dataset.
3. To test the machine with test data for both malaria and tuberculosis train datasets.
4. To determine the accuracy of the prediction.

**1.5 RESEARCH METHODOLOGY**

To achieve the aim of this work, we would consider the following methods.

1. The first task is to review related works and read about books that talks about malaria and tuberculosis so has to get the required metrics to be used in preparing the datasets.
2. The collated dataset would be preprocessed and cleans to remove unwanted metrics and empty data in the datasets.
3. Splitting of the datasets into train and test datasets into 70:30 ratio.
4. Train the machine with the allocated train datasets.

**1.6 SIGNIFICANT OF RESEARCH**

After the completion of this work, It is expected to have a model that would be predict tuberculosis and malaria for the dataset supplied.

**1.7 SCOPE OF STUDY**

This study focuses on just supervised machine learning algorithms in relation to data mining. And the dataset to be collected would be patient history of Adekunle Ajasin University Medical Clinic patient records.

**1.8 DEFINITION OF TERMS**

**Malaria:** An infective disease caused by sporozoan parasites that are transmitted through the bite of an infected Anopheles mosquito; marked by paroxysms of chills and fever

**Tuberculosis:** Tuberculosis (TB) is a potentially serious infectious disease that mainly affects your lungs. The bacteria that cause tuberculosis are spread from one person to another through tiny droplets released into the air via coughs and sneezes.

**Model:** A machine learning model can be a mathematical representation of a real-world process. The learning algorithm finds patterns in the training data.

**Prediction:** the action of predicting something.

**Machine Learning:** Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

**Data mining:** Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.