OPEN ACCESS
Computer Science & IT Research Journal
P-ISSN: 2709-0043, E-ISSN: 2709-0051
Volume 3, Issue 1, P.1-22, January 2022
DOI: 10.51594/csitrj.v3i1.290
Fair East Publishers
Journal Homepage: www.fepbl.com/index.php/csitrj

A MODEL FOR PREDICTION OF DRUG RESISTANT TUBERCULOSIS USING DATA MINING TECHNIQUE

Abdullahi, Halliru[1], Gregory Msksha Wajiga[2], Yusuf Musa Malgwi[3] and Abba Hamman Maidabara[4].
[1]No:5B Sabara Avenue U.D Bord OFF Kawo Bus Stop Nasara, Kano State
[2,3]Department of Computer Science,
Modibbo Adama University, Yola, P.M.B. 2076 Yola, Adamawa State, Nigeria.
[4] No: 45, Old G.R.A, Polo Ground Maidugu Street, Maiduguri, Borno State, Nigeria

*Corresponding Author: Abdullahi, Halliru
Corresponding Author Email: h.abbahamman@gmail.com
Article Received: 25-12-21 **Accepted:** 10-01-22 **Published:** 17-01-22

ABSTRACT

The rate of mortality in the recent time because of tuberculosis disease is so alarming. Drug-Resistant Tuberculosis is a communicable disease very dangerous that attack lungs, many victims were not identified due to weak health systems facilities, poor doctor-patient relationship, and inefficient mechanisms for predicting of the disease. Data mining can be applied on medical data to foresee novel, useful and potential knowledge that can save a life, reduce treatment cost, increases diagnostic and prediction accuracy as well as delay taking during prediction which reduce the treatment cost of a patience. Several data mining technique such as classification, clustering, regression, and association rule were used to enhance the prediction of tuberculosis. In this project I used Naïve Bayes Classifier to design a model for predicting tuberculosis. I considered the following parameters; Gender, Chills, Fever, Night sweat, Fatigue, Cough with Blood, Weight loss, and Loss of Appetite for classification phase 1. While Gender Chest Pain, Sputum, Contact DR, Weight Loss, In-adequate treatment for classification phase 2 as the clinical symptom. The Naïve Bayes Classifier has the advantage of

attribute independency, it is easy in construction, can classify categorical data, and can work on high dimensional data effectively. The model designed using Naïve Bayes Classifier is divided o into classification phase 1 and classification phase 2 and implemented using Phython 3.2 Programing Language. The result shows that Naïve Bayes Classfier was suitable in predicting drug resistant tuberculosis with performance accuracy of 82%, 98% and area under curve (AUC) is 88%.

Keywords: Model Prediction, Tuberculosis. Drug, Resistant, Data Mining.

INTRODUCTION

Background of the Study

Tuberculosis (TB) is a serious problem and transmit through bacteria known as Mycobacterium Tuberculosis (Zakhmi, and Arora, 2006). "It is one of the communicable disease that lead to high global mortality (Evora, seixas and kritski, 2017). Tuberculosis (TB) is a contagious disease caused by the bacterium Mycobacterium tuberculosis (M.tb) Every year since 1997, the World Health Organization (WHO) has published a Global Tuberculosis (TB) report, which provides an up-to-date assessment of the Global TB situation, and summarizes progress and efforts in prevention, diagnosis, and treatment of the disease, at country, regional and global levels. The 2020 Global TB report was released on 14 October 2020 and was compiled in the context of global TB control strategies and United Nations (UN) targets set in the political declaration at the September 2018 UN General Assembly high-level meeting on TB held in New York. The data accounted for over 99% of the world's population and reported data from 198 countries. The 2020 Global TB Report has two additional features of note: it complements and elaborates on the United Nations (UN) Secretary-General's 2020 progress report on TB, which was requested in the political declaration at the high-level meeting on TB. The report al so includes preliminary assessments of how the unprecedented corona virus disease -2019 (COVID-19) pandemic may affect TB health services, treatment and prevention efforts. However, it is important to note that global TB control efforts were not on track even before the advent of the COVID-19 pandemic, and the numerical gap between the estimated number of people with TB globally and the numbers reported to public health authorities remains wide. If all the missing people with TB, including those from the private sector, could be identified, the shortfall in reaching all the targets might be greater; thus, the failure to reach the targets cannot be assumed to be because of the gap between reported and estimated numbers of people with TB. In 2019, TB remained the most common cause of death from a single infectious pathogen. Globally, an estimated 10.0 million people developed TB disease in 2019, and there were an estimated 1.2 million TB deaths among HIV-negative people and an additional 208, 000 deaths among people living with HIV. Adults accounted for 88% and children, aged <15 years, for 12% of all people with TB. Most people who developed TB in 2019 were in the WHO regions of South-East Asia (44%), Africa (25%), and the Western Pacific (18%), with smaller percentages in the Eastern Mediterranean (8.2%), the Americas (2.9%) and Europe (2.5%). Eight countries accounted for two thirds of the global total: India (26%), Indonesia (8.5%), China (8.4%), the Philippines (6.0%), Pakistan

(5.7%), Nigeria (4.4%), Bangladesh (3.6%) and South Africa (3.6%). While progress is being made, it has been very slow, and it is anticipated that the world will not end TB as a global public health threat by 2035 as envisioned in the End TB Strategy. For example, while the target was to reduce TB incidence by 20% between 2015 and 2020, the 2020 global TB report indicates that there was a reduction of only 9% in TB incidence in this period, with an annual reduction of only about 2%. Similarly, mortality reduction targets, set at 35%, were not reached, with only a 14% change in death rates achieved between 2015 and 2020. The major constraints are related to inadequacies in the identification of people-including children-with TB, for all TB and drug resistant TB in particular, prevention of TB and financing of the TB response including of essential TB research. In this paper we highlight the status of TB care and prevention as presented in the 2020 WHO Global TB report, outline the persisting constraints, summarize the efforts that are being made to address these constraints and offer suggestions on how these efforts can be ramped up. The gap in finding people with Tuberculosis out of the 10 million people estimated to have developed TB in 2019, 7.1 million (71%) were identified and reported to national TB programs around the world, leaving a gap of 2.9 million people (29%). These missing people with TB include those who were diagnosed with TB, but were not reported to public health authorities, (including those not reported from the private sector) and those who were not diagnosed and therefore not treated. This pool also includes patients with drug - susceptible or drug-resistant TB, using current definitions. However, a greater proportion of people with drug resistant TB are missing: of the estimated 500,000 people with rifampicin resistant/multidrug-resistant TB (RR/MDR-TB), only 206, 030 (41%) were identified, as a result of inadequate testing for drug susceptibility especially among new people with TB. In 2019, only 57% of people identified to have pulmonary TB were bacteriologically confirmed, and of those who were bacteriologically confirmed, only 61% were tested for RR/MDR-TB, comprising 59% of people not previously treated for TB and 81% of those who had previously been treated for the disease. Whilst 206, 030 people with MDR/RR-TB were detected and notified in 2019, (a 10% increase from 186, 883 in 2018), only 177,099 people were enrolled in treatment, comprising only 38% of the estimated number of people who developed RR/MDR-TB in 2019. Another important sub-group of the missing people with TB are children under 15 years of age. Tuberculosis diagnosis and treatment gaps are wider among children than adults. There needs to be a sustained, concerted effort and universal focus on identifying and treating missing people with TB. WHO reports that a 50% drop in the number of people with TB detected could result in up to 400,000 additional TB deaths in a year. Governments of high TB pendemic countries need to ensure there are rapid TB diagnostic services available in every health facility, so all people with TB can be reached. With COVID-19 causing disruption of health services, many countries have been reported to be using GeneXpert machines for COVID-19 testing, and others have reassigned TB programme staff to COVID-19, causing further J. Chakaya, M. Khan, F. Ntoumi et al. International Journal of Infectious Diseases xxx (xxxx) xxx–xxx G Model IJID-5190; No. of Pages 6 2shortages of the already meagre TB diagnostic and treatment resources. Innovative plans are needed to maintain TB diagnostic services in the wake of the COVID-19 pandemic

(Zumla et al., 2020). Innovations to adapt TB diagnostic platforms to screen for both TB and COVID-19, rolling out additional machines, and investing in development of low-cost rapid diagnostic tests for both infections are important and urgently needed. The methods used by the WHO to estimate the global burden of TB includes the use of data from national prevalence surveys; notification data adjusted by a standard factor to account for under -reporting, over - and under diagnosis; inventory studies that measure under-reporting; and expert opinion on TB detection gaps. Prevalence surveys, more commonly than not, have shown that country estimates of prevalence have been lower than what was observed, consequently, the incidence estimates upon which the global estimates are based may be lower than what actually exists. While the methods used to estimate the burden of TB are continuously being improved, imperfections remain which may explain the relatively wide uncertainty intervals especially with country level estimates. Differences in methodological approaches for the estimation of the burden of TB have in the past led to significant differences in the estimates of the global burden of TB provided by the WHO and the Institute for Health Metrics and Evaluation that undertakes the Global Burden of Disease project (Garcıa-Basteiro et al., 2018).

These methodological challenges in the estimates of the global burden of TB may have significant consequences at the programmatic level in countries. National TB programs may set unrealistically high targets or appear to be setting unambitious targets if national targets are based on trends of TB notifications as has been proposed by some experts (Trébucq and Schwoebel, 2016). We urge the global TB community under the leadership of the WHO Global Task Force on TB Impact measurement (WHO, 2019) to continue to address this challenge and to harmonize approaches for the estimation of the burden of TB so that disease burden estimates become more robust with a more precise estimate of the size of the TB incidence -detection gap. It is also equally important to develop robust approaches and tools for the estimation of the burden of TB at sub-national level where efforts to identify people with TB are focused as has been proposed by some countries e.g., Indonesia (Parwati et al., 2020). Several interventions, many of which have been implemented to a variable degree in various settings are in place to narrow the TB incidence-detection gap and excellent reviews have been published on this subject (Reid et al., 2019).

Some of the major steps that need to be taken to address the challenge of missing people with TB at the national and sub-national level in countries include the identification of TB at risk populations to be targeted to actively find people with TB, the use of highly sensitive methods for TB screening and specific TB diagnostic tests (e.g. Xpert MTB/RIF), linkage to care and treatment of all people identified to have active TB, measures to ensure retention in care and adherence to treatment, engagement with communities, development of strong partnerships with the non-state health sector and ensuring that TB is an integral component of evolving universal health care programs. Cascade analysis at all levels of the national TB program is an effective way of identifying bottlenecks in the pathway to care, treatment and cure of TB (Subbaraman et al., 2019). The misalignment of TB services with the places where people seek health care may be a significant impediment to detection of people with TB. Alignment of health services would

be useful, not only for TB but for more efficient provision of health care services in general. Ensuring that a large proportion of the targeted population is reached with TB screening and diagnostic test services that is sustained over time has been documented to reduce the incidence of TB in diverse settings (Corbett et al., 2010; Kaplan et al., 1972; Marks et al., 2019). Such approaches need to be adopted and scaled up in each country to ramp up TB detection efforts and to narrow the TB incidence- notification gap. The challenge of sub-optimal TB treatment outcomes The 2020 WHO global TB report contains data on the treatment outcomes of people treated for drug susceptible TB in 2018 and those initiated on treatment for drug resistant TB in 2017. Globally 85%, 76% and 57% of people with new and relapse TB, HIV-associated TB and RR/MDR-TB, respectively, were treated success-fully.

These figures represent a sub-optimal performance. The poor treatment outcomes are driven mainly by lack of evaluation, poor linkage to treatment, death and high loss to follow-up. The underlying causes of poor treatment outcomes include un-identified or additional drug resistance (Cegielski et al., 2014), inadequate support provided to people with TB to ensure high level of adherence, weak recording and reporting systems and inadequate prevention and management of advanced HIV disease including the provision of anti-retroviral treatment. For children specifically, challenges include under-identification, inadequate recording and reporting, poor drug formulation options, poor caregiver availability/capacity for treatment, and persistent issues with stock out of the few drug options appropriate for this population. By 2019, only 30% of the 3.5 million five-year target for children treated for TB had been achieved, including only 8% of the 115,000 target for treatment of children with RR/MDR -TB. Lapses in data collection and proper and consistent disaggregation continue to negatively affect identification and treatment as well as programming and resource-allocation for children under 15 years and adolescents 10-19 years of age. Based on observational studies in Bangladesh and sub-Saharan Africa (Bulabula et al., 2019; Trebucq et al., 2018), which were confirmed by results of the standardised treatment regimen of anti-TB drugs for people with MDR-TB (STREAM) Stage 1 trial, WHO now recommends that the shorter 9-month MDR-TB regimens be preferred as it achieves treatment success in roughly 80% of participants (Nunn et al., 2019).

Furthermore, second-line injectable drugs, if possible, should be replaced by a fully oral, bedaquiline (BDQ)-containing regimen, one of the new (along with delamanid and pretomanid) potent MDR-TB drugs that are now being increasingly prescribed globally. Such regimens should be urgently scaled-up to improve outcomes, minimize side effects and improve adherence along with point-of-care molecular drug susceptibility testing of second line MDR-TB drugs (Bisimwa et al., 2020; Cox et al., 2018; Diacon et al., 2014; Gler et al., 2012). Furthermore, attempts have been made in the recent past to reduce the duration of treatment in drug susceptible TB focused on the use of fluoroquinolones, however, the large clinical trials that were undertaken for this purpose failed to demonstrate efficacy for relapse free cure (Grace et al., 2019).

The 2020 Global TB report includes a comprehensive description of new medicines and regimens that are at various stages of clinical development and the reader is urged to consult this

document. It is anticipated that medicines that are currently reserved for the treatment of drug-resistant TB will soon be available for the treatment of all TB including TB that is caused by pathogens currently considered to be drug-susceptible and which have the potential to significantly shorten treatment. This is exciting and we urge that efforts to develop new medicines and treatment regimens be ramped up in line with the ambition that came out of the first ever United Nations High Level meeting on TB. This will require additional funding and more robust partnerships between research funders and research groups. It is particularly important that the global south where TB is endemic be deeply engaged in all the processes required for the development of new medicines and regimens for TB. To ensure that TB treatment is taken as prescribed, directly observed of treatment (DOT), has been a pillar of TB care and prevention although its efficacy compared with self-administered treatment has been questioned (Volmink and Garner, 2007). However, direct observation of TB treatment ingestion may be disempowering to the person being treated in addition to imposing demands on that person and the health care system that can be difficult to cope with. An approach that is based on providing comprehensive and individualized support to people being treated for TB is more likely to be acceptable to people with TB and their families and has been associated with better treatment completion rates (Alipanah et al., 2018).

Furthermore, in the current digital world, new ways of supporting people on treatment, including interactive two-way mobile phone text message reminders and video assisted DOT have also been associated with good levels of treatment adherence while providing psychological support through remote counselling (Ngwatu et al., 2018). Addressing sub-optimal TB prevention recent estimates of the global burden of latently infected persons with Mycobacterium tuberculosis suggests that 23% (95% uncertainty interval 20-4%-26.4%) of the global population or about 1.7 billion people harbor this infection. This large pool of latently infected persons is the seedbed of future TB. It has been estimated that by 2030 and 2050, this pool of latently infected persons will generate 16.3 and 8.3 people with active TB per 100,000 population respectively (Houben and Dodd, 2016).

However, with current and hopefully future diagnostic and treatment tools most if not all these episodes of TB can be prevented. A major challenge for TB prevention currently is that there are no easily deployable tools that can identify the subset of latently infected persons who are likely to progress to active TB. While some bio-signatures offer this potential (Sumner et al., 2019) none has been developed to a state that can be used to support programs intended to deliver targeted TB preventive therapy. We urge that this area of research also receives due attention, especially funding, to help refine the targeting of TB preventive therapy as programmatic management of TB preventive therapy is advanced. It is noteworthy that progress is being made in the provision of TB preventive therapy even though by 2019, the world was far off target. The small successes that are being made, however, need to be celebrated. For example, in 2019, TB preventive therapy was provided to 4.1 million people, which was nearly double the number that received this treatment in 2018. For HIV-infected individuals, 85% of those eligible to receive TB preventive therapy received this intervention, which is commendable. While provision of TB

preventive therapy to household contacts of people with active TB remained low, it seems TB programs across the world have begun to push forward with this intervention as evidenced by the increase in the number of household contacts who were provided with TB preventive therapy in 2019 (538, 396) compared to 2018 (423, 607). The development of new regimens that are shorter and equally effective such as a weekly dose of rifapentine and isoniazid for 3 months (3HP), a daily dose of rifampicin plus isoniazid for 3 months (3HR), a daily dose of rifapentine plus isoniazid for 1 month (1HP), a daily dose of rifampicin for 4 months (4R) (Sterling et al., 2011; Swindells et al., 2019; WHO, 2020) appeared to have helped to advance implementation of TB preventive therapy at the country level. We believe that combination therapies such as 3HP, 3HR and 1HP have been a truly major advance in that they may have allayed the fears held by some TB program managers and opinion leaders at the national level, that TB preventive therapy, using isoniazid preventive therapy could promote the development of resistance paving the way for acceleration of TB preventive therapy. Our view is that this moment should be seized to ensure TB preventive therapy services are rapidly scaled-up. By doing this TB preventive therapy, will contribute to the global desire to accelerate declines in TB incidence. Preventing future TB must not just be focused on finding and treating people with TB but should also include efforts to address social and other determinants of the disease. While efforts are being made to actively find people with TB and to provide TB preventive therapy, governments must ensure that the expansion of economies continues in this COVID-19 era (WHO, 2020) and percolate to all segments of the populations in every nation. The major drivers of TB – undernutrition, poverty, diabetes, tobacco smoking, and household air pollution (Dooley and Chaisson, 2009; Lee et al., 2020; Noubiap et al., 2019; Reid et al., 2019) must be addressed if the world is to expect to end TB as a public health threat by 2035. Moreover, since the year 2000, WHO has projected that 54 million people have survived tuberculosis (including a large proportion of children and adolescents) (Allwood et al., 2020).

However, these people are more likely to develop residual lung damage and recurrent tuberculosis (Allwood et al., 2019, 2020) and are at increased risk for all cause-mortality (standardized mortality ratios: 2.91 (95% CI 2.21-3.84) (Romanowski et al., 2019) compared with the general population or matched controls. We urge national TB programs to incorporate post-tuberculosis health and wellbeing interventions in the package of services provided to people with TB (Chakaya et al., 2016) and encourage the research community to undertake research intended to unravel the biomedical and social determinants impacting TB survivals' long-term prognosis (Romanowski et al., 2019).

There has been over two years since global leaders signed the UN general assembly high level meeting on TB declaration. It is disappointing and disheartening that we are not on track to reach the testing and treatment goals. While global political and public health systems have been severely shaken by the COVID-19 pandemic, which undoubtedly has dislodged TB from the number one slot in the year 2020 the expectation is that with the rollout of COVID-19 vaccines and public health measures, COVID-19 may be brought under control (Forni and Mantovani, 2021). The long-term socio-economic effects of the COVID-19 pandemic will further drive

poverty, malnutrition and poor living conditions, which are risk factors associated with TB prevalence. Thus, TB will likely quickly re-occupy the number one spot as the most common infectious cause of mortality worldwide. From first detection of COVID-19 as a new human pathogen, global coordination and political will with huge financial investments have led to the development of effective vaccines against SARS-CoV2 infection within 11 months. The world now needs to similarly focus on development of new vaccines for TB utilizing new technological methods. We urge the WHO, other UN agencies and partners to develop mechanisms that strongly push countries to ensure TB multi-sectoral accountability frameworks are not just developed but are pursued with vigour. Holistic programs for human development need to be developed and leaders made to account for their implementation. Global health inequities driving TB epidemiology, including the environment and climate control, gender, age, socio-economic status, and wealth as well as resource distribution, should also be addressed by multiple approaches and sectors. It is not yet too late to do this and on the World TB Day 2021, we expect that every leader and person of influence will get the message that it is time to reduce inequities as we work towards a TB free world. While there is a continued need to develop new prevention and J. Chakaya, M. Khan, F. Ntoumi et al. International Journal of Infectious Diseases G Model IJID-5190; No. of Pages 6 4treatment tools for TB, we strongly believe that the effective and efficient application of current tools can significantly dent the burden of TB and advance the push to end TB as a global public health threat. The time to do this is now. The world's TB control programs were already failing to meet the ambitious goals during the past 2-3 years, largely because of systemic weaknesses. The COVID-19 pandemic and its rapid global spread, highlights the intrinsic weaknesses of health care systems. The ultimate generic issue for all communicable diseases of public health importance is that of creating a systemically strong healthcare base upon which to build disease-specific programs such as for TB

Statement of the Problem

The control of Drugs Resistant Tuberculosis (DR-TB) is becoming a threat worldwide. In most countries where mycobacterium is high, there is an increase likelihood for a Drugs Resistant Tuberculosis (DR-TB) epidemic to occur which can lead to loss of lives yearly at an increasing rate and many people tend to be affected with the TB through a contact with someone that is already affected with the virus especially through coughing when the droplets touches the body of another person, DR-TB spread mostly among family members or at office from a coworker or any other public places. COVID19 and HIV can increase the chances of having DR-TB which can quickly re-occupy the number one spot as the most common infectious cause of mortality worldwide because the pathogens of these people is already weak so any facial contact with a victim of TB can increase the chances of the spread of DR-TB, person can have latent TB or active TB and get recovered through medication. Due to having high number of people with COVID19 and HIV, the world now needs to similarly focus on development of new vaccines utilizing new technological methods for prediction of DR-TB. There must be mechanisms that strongly push countries to ensure TB multi-sectorial accountability frameworks are not just developed but are pursued with vigor. Holistic programs for human development need to be

developed and leaders made to account for their implementation. Global health inequities driving TB epidemiology, including the environment and climate control, gender, age, socio-economic status, and wealth as well as resource distribution, should also be addressed by multiple approaches and sectors, it is not yet too late to do this. The symptomatic feature presented by the patients is the clinical presentation of tuberculosis in a patient. This feature is an indication of disease course and therefore, has direct significant in directing the clinician on which decision to take. The disease sign and symptom has made tremendous achievement in predicting and diagnostic disease.

Aim and Objectives of the Study

Aim

The aim of this project is to design a model for prediction of Drug Resistant Tuberculosis in a patient using Data Mining Technique in order to reduce the time it takes to identify the state of a patient.

Objectives

The objectives are as follows:

1. To design a model for predicting drug resistant tuberculosis which utilizes large data obtained from hospital.
2. To model the drug resistant tuberculosis prediction using Naïve Bayes classifier in order to enhance the accuracy of the prediction of DR-TB
3. To find out the various stages of tuberculosis that causes the drug resistant tuberculosis in relation to Covid19 and HIV

This research I will be making use of tuberculosis data set obtained from FMC Yola

Significant of the Study

Tuberculosis remain one of today's global health challenges, ranking as the second leading infectious cause of death and of the most burden-inflicting disease in the world. The purpose of this study is to the following reasons:

1. To support countries to implement better tools that are needed to control spread of TB worldwide
2. The discovery of new markers for high and low risk individuals would allow evidence-based determination of who to treat, how to treat, and how long to treat both for prevention and cure.
3. Ability to predict the symptom of tuberculosis due to validated model to diagnose an individual

This model can assist specialist that have limited knowledge in discovering the symptoms of TB

Scope and Limitation of the Study

Despite notable progress made in the past decade, there is still need of an alternative way of diagnosis of resistant tuberculosis, within this context the overall goal is to develop a model that can be used for diagnosis of resistant tuberculosis symptom. The study highlighted the

significant of Naïve Bayes Classifier in supervised learning considering its independent assumption. Performance of the system was checked using matrix and ROC only.

LITERATURE REVIEW

Machine Learning

Machine learning, by its definition, is a field of computer science that evolved from studying pattern recognition and computational learning theory in artificial intelligence. It is the learning and building of algorithms that can learn from and make predictions on data sets. These procedures operate by construction of a model from example inputs in order to make data-driven predictions or choices rather than following firm static program instructions (Simon, et al., 2015). It is about learning how to do better in upcoming based on experience learned in the past (Cruz & Wisharts, 2006). For example, learns to act as an intelligent or predict disease accurately based on some number of observations (Raj & Prasanna, 2013). The machine learning can be of the following forms: supervised, unsupervised, semi supervised and reinforcement. In supervised Machine learning the program is "trained" on a pre-defined set of training examples, which then facilitate its ability to reach an accurate conclusion when given new data. In unsupervised machine learning, the program is given a bunch of data and must find patterns and relationships therein (Simon et al. 2015).

The semi-supervised machine learning is used for the same applications as supervised learning, but it uses both labeled and unlabeled data for training. While the Reinforcement machine learning is a computer program which interacts with a vibrant environment in which it must perform a certain goal. The program is provided feedback in terms of rewards and punishments as it navigates its problem space (Sharma & Kumar, 2017). Our objectives are to train a dataset which we can develop learning algorithms that can learn automatically without human assistance or intervention. Machine learning provides an alternative solution to a medical problem by using different techniques such as clustering and classification applied on previous real data to predict current disease. This approach was found stimulating by many researchers trying to use medical data to predict disease (Razak, 2015).

METHODOLOGY

Concept of Classification Technique

One of the data mining function that assigns items in a collection to target categories or classes is classification. It provides predictive data mining approach which makes a prediction about values of data using known results found from different data. The goal of classification is to accurately predict the target class for each case in the data (Bharathi & Deepankuma, 2014).

Several research studies have been conducted using classification algorithms. Recently, Badeji (2018). designed an automated model for tuberculosis identification, which predict tuberculosis based on the following parameters Coughing, sometimes with mucus or blood, chills, fatigue, fever, loss of weight, loss of appetite, night sweat, the prediction model used Naïve Bayes Classifier. This system considers only the parameters of tuberculosis disease, without considering checking accuracy as means of authenticating model.

In this study, I will develop a model for predicting Drugs resistant tuberculosis using Naïve Bayes classifier. This classifier Support independent assumption. The model can be train and test using large data set obtained from Hospital.

Naïve Bayes Classification

Naïve Bayes classifiers can be train very efficiently in a supervised learning setting. It can also work as statistical based technique for classification (Bharathi & Deepankuma, 2014). It uses Bayes theorem to find the probability of another event that has already occurred. It is a popular classifier among other classifiers such as decision tree and support Vector Machine. Despite its simplicity, it competes favorably with other classifiers (Georgina, 2018). Naïve Bayes classification is simple and particularly suited when the dimensionality of the input is high. Despite its simplicity, it can outperform more sophisticated classification method. It provides perspective for understanding many learner algorithms and works on the assumptions that: is easy to construct, classifying categorical data, occurrences of an event (attributes) are independent and can be trained in a supervised manner (Path et al., 2016). The major advantage of Naïve Bayes in classification is its simplicity and its ability to approximate probabilities for a class on any given instance (Kononenko, 1991).

Software Design Phase

The proposed model is implemented using Python programming environment version 3.2. Python is a general-purpose, high-level programming language which is widely used in the recent times; its constructs enable the user to write clear programs on both a small and large scale. Python supports a dynamic type system and automatic memory management and has a large and comprehensive standard library. Python interpreters are available for many operating systems. Python program has a good number of programming packages for Naïve Bayes classification such pandas, nympy, matplotlib, GaussianNB, Metrics, and many more. It also known to have an abundance of libraries that assists with data analysis and scientific computing. In python programming you will also find package for determining the accuracy of your model such as confusion matrix and ROC (Srinath, 2017).

Software Requirement

The software is designed using Python programming language version 3.2 since it supports all the needed library for the packages that Naïve Bayes Classifier required. It's also has packages that support the accuracy of a model such as Confusion Matrix and ROC. This language was chosen over other in order to have maximum control on interface design, better presentation of statistical result and flexibility.

Hardware Requirement

The hardware requirements are;

1. All windows operating system above windows 7, 64bits for PC and iOS 8, 10 for Macintosh Operating system.
2. All CPUs
3. Preferably 4GB RAM and 40GB hard disk drive free space.

Naïve Bayes Algorithm

The Naïve Bayes Classifier works as follows:

1. Let T be a training set of samples, each with their class labels. There are k classes, Ci, C2,…, Ck. Each sample is represented by an n-dimensional vector, X= {xi, x2,…., xn}, depicting n measured values of the n attributes, Ai, A2, . . ., An, respectively.

2. Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class Ci if and only if:

P (C |Xi) > P (CXj) for ij Sm, j # i.

Thus we fund the class that maximizes P (C |Xi). The class C for which P (C |Xi) is maximized is called the maximum posteriori hypothesis.

By Bayes' theorem P(CIXi) = P(XiIC)P(C) / P(X).

3. As P(X) is the same for all classes, only P (XiC)P(C) need be maximized. If the class a priori probabilities, P(C), are not known, then it is commonly assumed that the classes are equally likely, that is, P (Cl) =P(C2) =. . .=P (Ck), and we would therefore maximize P(X |Ci). Otherwise we maximize P (X |Ci) P (Ci).

4. Given data sets with many attributes, it would be computation-ally expensive to compute P(XCi). in order to reduce computational in evaluating P(X |Ci)P(Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one an-other, given the class label of the sample. Mathematically this means that:

$$P(\frac{X}{Ci}) = \sum_{k=0}^{n} P(Xk|Ci) \quad K=1\ldots\ldots n$$

The probabilities P(x1ICi). P (x2K'i P(xn |Ci)…., P(xn |Ci) can eais1y be estimated from the training set. Recall that here XK refers to the value of attribute AK for sample X. If Ak is categorical, then P (xk|Ci) is the number of sample of class Ci in T having the value xk for attribute AK, divided by freq(Ci,T), the number of class Ci in T.

5. In order to predict the class label of X, P(X |Ci) P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of X is Ci if and only if it is the class that maximizes P (X |Ci) P(Ci).

The Proposed Model

The proposed model (Fig. 3.1) is based on the trends of predicting drug resistant tuberculosis in a suspected patient. It involves several steps, which include data collection, data cleaning and transformation, classification and prediction, and finally, interpretation, evaluation and knowledge discovery.

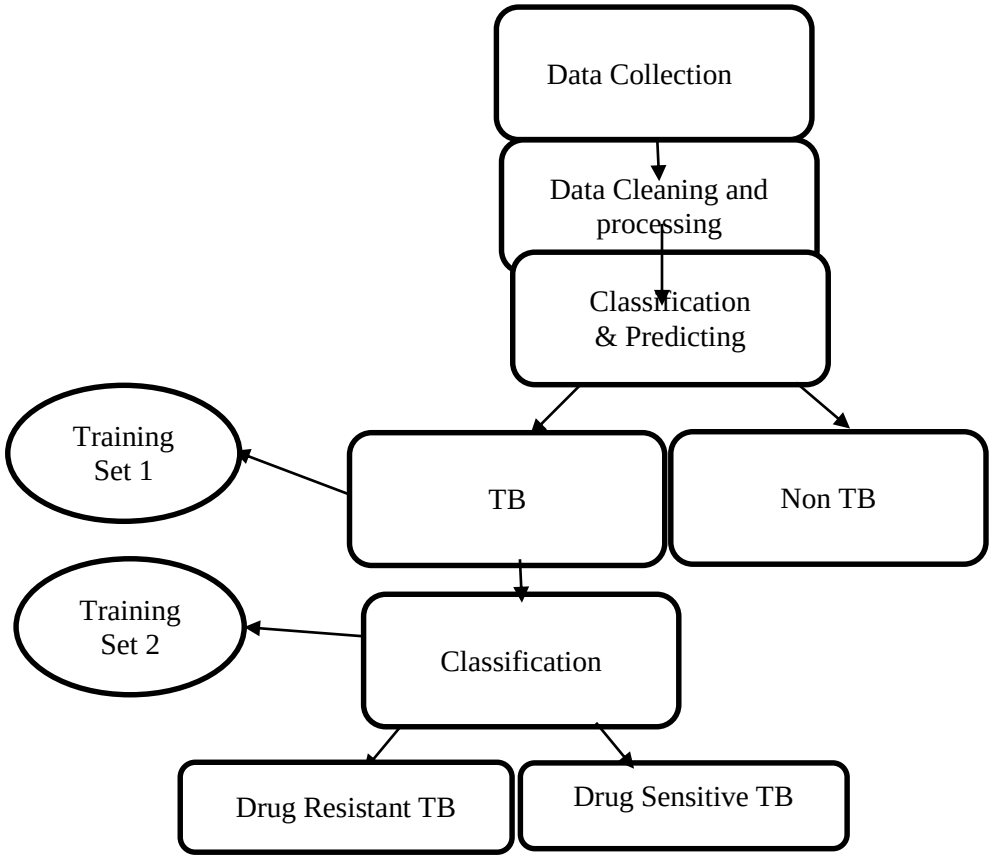


Figure 1: Model Architecture of Drug Resistant TB

Prediction Procedure

Data Collection: The data collected would be divided into two portions; one portion of the data is extracted as a training set, while the other portion would be used for testing. The training portion is taken from a table stored in a C:\dr_dataset.xlsx (say Table 4.la) is called as datal which is *training set1*, while the training portion taking from another table stored in a C:\dr_datasetl.xlsx (called Table 4. 1b) is called as data 2 which is *training set2.*

Data Preprocessing: Data preprocessing was done to remove noise, and outlier.

Transformation: The processed data is transform from hard copy to soft copy.

Classification and prediction: Based on the nature of variable in our dataset, we will use Naïve Bayes classification techniques twice; *C1assfication phase 1 and C1assfIcation phase 2.* Working of the framework is illustrated as follows:

- Data collection and preprocessing are done.
- Preprocessed data is stored in a training set I and training set 2. These datasets are used during classification.
- Test data set is stored in database test data set.
- Part of test data set is compared for classification using classifier 1 and rest part is classified using classifier 2 as follows:

Classifier phase I: classify into positive or negative class label. If the patient is having tuberculosis, then the patient is classified as positive (P), while a patient is classified as negative (N) if the patient does not have tuberculosis

Classifier phase *2:* classify only that data set that has been classified as positive by classifier I and then further classify them into drug resistant tuberculosis and drug sensitive tuberculosis class label

The system would be designed in such a way that the core parameters as a determining factor should be supplied their value

Data Collection

The data set used in this research was collected from the record office, guided by the ethics of the specialist Hospital Yola, Adamawa state, and then transformed from the stored C:\data\data called dataset. Each patient file will be extracted and reviewed for signs and symptoms of malaria then check for laboratory confirmation result from diagnosis. The data is divided into two tables: the first is called dataset which contain data used in phase I of the classification, while the second table called dataset] which contain data used in phase 2 of the classification.

Experimental Setup

The data was divided into two part; one part is used for the training the model. Our model consists of two training set. First training set was used for classifying patient to either have tuberculosis (positive) or not (negative). The second training set would further classify those that were positive tuberculosis in the first classification as either having drug resistant tuberculosis (DR) or drug sensitive tuberculosis (DS). Using naïve Bayes classification technique, we have predicted a new patient having a set of parameter. Naïve Bayes Classifier works as follows:

Using:

Naive Bayes; $P(c/x) = \frac{p(x/c)P(c)}{P(x)}$

Where c: is the class

P (clx): is a posterior prob. of class given predictor.

P (c): is the past (prior) prob. of class.

P (xlc): the prob. of predictor given class.

P (x): past prob. of the predictor."

Positive class label (P): Patient may have tuberculosis (positive) if the probability of selected features points out that the probability of positive class is greater than negative class.

$$P(positive/patient) = P(c/x) = \frac{P(patient/P)P(P)}{P\left(\frac{patient}{P}\right)+\left(\frac{Patient}{N}\right)*P(N)}$$

Negative class label (N): patient may not have tuberculosis (negative) if the probability of selected features points out the probability of negative class is greater than positive class

$$P(Negative/patient) = \frac{P(patient/P)P(P)}{P\left(\frac{patient}{P}\right)*P(P)+\left(\frac{Patient}{N}\right)*P(N)}$$

Drug Resistant Tuberculosis (BR): Patient may have a drug resistant case of tuberculosis if the probability of selected features points out that the probability of drug resistant class is greater than drug sensitive class.

$$P(DR/patient) = \frac{P(patient/DR)P(DR)}{P\left(\frac{patient}{DR}\right)*p(DR)+p\left(\frac{Patient}{DR}\right)*P(DS)}$$

Drug Sensitive (PS): patients may have Drug sensitive case of tuberculosis if the probability of selected features points out that the probability of Drug sensitive class is greater than Drug resistant class.

$$P(DS/patient) = \frac{P(patient/DS)P(DS)}{P\left(\frac{patient}{DR}\right)*p(DR)+p\left(\frac{Patient}{DS}\right)*P(DS)}$$

Table 1: Predicting patient to be either have drug resistant tuberculosis (DR) or drug sensitive tuberculosis (DS)

	Gender	Contact DR	Smoking	Alcohol	Cavitary pulmonary	Diabetes	TBoutside	Class	Age_ < 45years	Age_ ≥ 45years	Nutritional - Normal	Nutritional_underweight
1	0	1	1	0	0	0	0	0	0	1	1	0
2	0	0	1	0	0	0	0	0	1	0	1	0
3	0	1	0	0	0	1	0	1	0	1	0	1
4	1	1	1	0	0	0	0	0	0	1	1	0
5	0	1	1	0	0	1	0	0	1	0	1	0

Where DS (1) = Drugs Sensitive, and DR (0) = Drugs Resistant

Performance Measure of Classifier

Accuracy check

"The accuracy check of Naïve Bayes algorithm can be check using confusion Matrix (Goyal and Mahta, 2012) and Receiver Operating Characteristic (ROC) analysis.

Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance is such systems is commonly evaluated using the data in the matrix. The following tables shows the confusion matrix for a two class classifier (Goyal and Mahta 2012). It classifies each instance in to one of two classes. The classes are true and false; this gives rise to four possible classifications for each instance as listed below.

1. True-Positive (TP) means positive pattern seen as positive
2. False-Positive (FP) means negative pattern seen as positive
3. False- negative (TP) means positive pattern seen as negative
4. True- negative (TN) means negative pattern seen as negative

Table 2: A 2 x 2 Confusion Matrix for two Class Classifier

Confusion Matrix		Actual class	
		Positive(1)	Negative(0)
Predicted class	Positive(1)	True Positive (TP)	False Negative(FN)
	Negative(0)	True Negative(TN)	True Negative(TN)

From the table above the classification that lies along the major diagonal that is TP and TN are the correct classifications. While the remaining fields that is FN and FP signify model error. From Confusion Matrix many model performance metrics can be derived, popular among the metric is accuracy, which is defined as:

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

Here accuracy rate is all the correctly classified patterns divided by total number of patterns. Other performance matrices include precision and recall defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall measures the negative pattern correctly identified as:

$$\text{Recall} = \frac{TN}{TN+FN}$$

Positive Predictive Value- values of positive

$$\text{PPV} = \frac{TN}{TP+FN}$$

Negative Predictive Value – where Negative occurs

$$\text{NPV} = \frac{TN}{TN+FP}$$

ROC and Area under Curve

ROC is a graph representing the performance of a classification model at all classification thresholds (Vihinem, 2012). This curve plot two parameters (TPR) on the vertical axis and false positive rate (FPR) on the horizontal axis, this signifies the relationship between true positive and true negative rates of a classifier through minimizing sensitivity (false positive) and specify (false negative) (Raj and Prasanna, 2013). ROC curve are accessed by either smearing the classification rule on test dataset with known classes or by using a sample of reused method e.g. cross-validation. It presents better performance in a number of ways through decreasing standard error as both the number of test sample and area under curve (AUC) increase and increases sensitivity when performing analysis of variance test (Hand and Till, 2001). A good classification rule

design by a classifier is reflecting on thr ROC curve by lying at the upper left triangle of the square area area (Vihinen, 2012).

RESULTS AND DISCUSSION

Models Assessment of Drug Resistant Tuberculosis

The dataset was divided into 70%, 300% for training, validation and testing respectively. Naïve bayes classifier were used for the study. Each of the models where trained using the training dataset and validated using the validation dataset. To test the resistant ability of Tuberculosis, we tested it using data it has not previously seen I.e. the test dataset. The results from each of the models were organized and assessed in terms of the magnitude of the statistical error between the measured result and the real data.

The results from the models were organized and assessed in terms of the magnitude of the statistical error between the measured result and the real data. This was achieved by measuring the average of the Mean Square Errors (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for the models used for Drugs Resistant of Tuberculosis.

]:

	Gender	Age	Contact DR	Smoking	Alcohol	Cavitary pulmonary	Diabetes	Nutritional	TBoutside	Class
0	Male	>= 45 years	Yes	Yes	No	Yes	No	Underweight	No	DR
1	Female	>= 45 years	Yes	Yes	No	Yes	No	Underweight	No	DR
2	Female	>= 45 years	No	No	No	Yes	No	Underweight	No	DR
3	Male	>= 45 years	Yes	Yes	No	Yes	No	Normal	No	DR
4	Female	< 45 years	No	No	No	Yes	No	Underweight	No	DR
...	...	...	...	...	...	...	...	...	...	...
95	Female	< 45 years	No	No	No	No	No	Normal	No	DS
96	Male	< 45 years	No	No	No	No	No	Normal	No	DS
97	Female	< 45 years	No	No	No	No	No	Normal	No	DS
98	Female	< 45 years	Yes	No	No	No	No	Normal	No	DS
99	Female	< 45 years	No	No	No	No	Yes	Normal	No	DS

100 rows × 10 columns

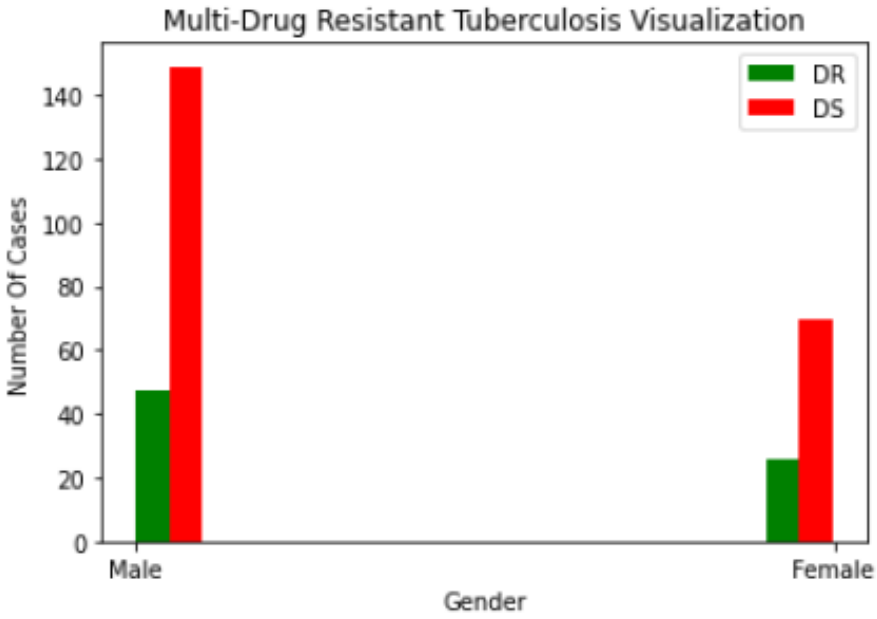
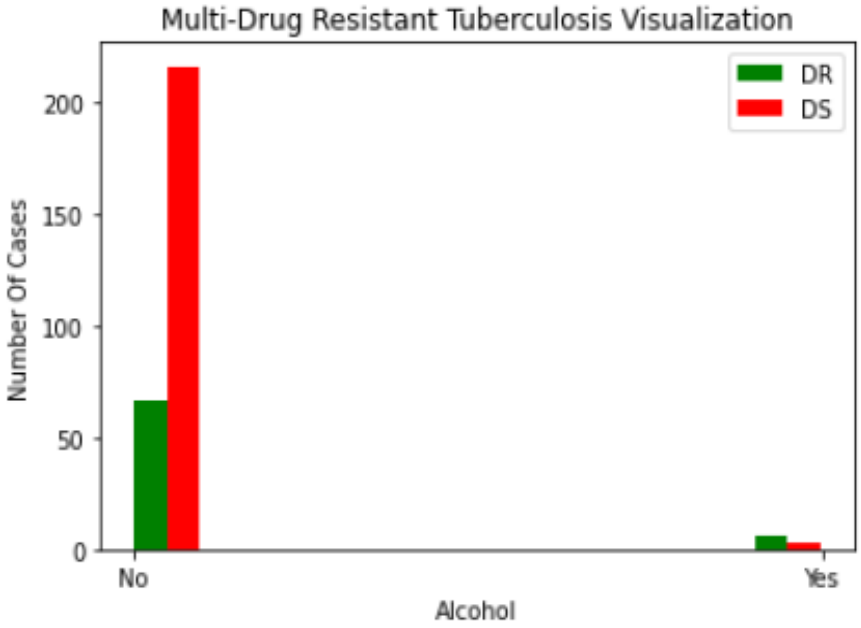


Figure 2: 4.0 Graphical representative of Visual Gender



```
def print_unique_col_values(df):
       for column in df:
            if df[column].dtypes=='object':
                print(f'{column}: {df[column].unique()}')
```

```
print_unique_col_values(df1)
```

```
Gender: ['Male' 'Female']
Age: ['>= 45 years' '< 45 years']
Contact DR: ['Yes' 'No']
Smoking: ['Yes' 'No']
Alcohol: ['No' 'Yes']
Cavitary pulmonary: ['Yes' 'No']
Diabetes: ['No' 'Yes']
Nutritional: ['Underweight' 'Normal']
TBoutside: ['No' 'Yes']
Class: ['DR' 'DS']
```

```
from sklearn.naive_bayes import GaussianNB
model = GaussianNB()
```

```
model.fit(X_train, y_train)
```

```
GaussianNB()
```

```
model.score(X_test,y_test)
```

```
0.7159090909090909
```

```
X_test[:10]
```

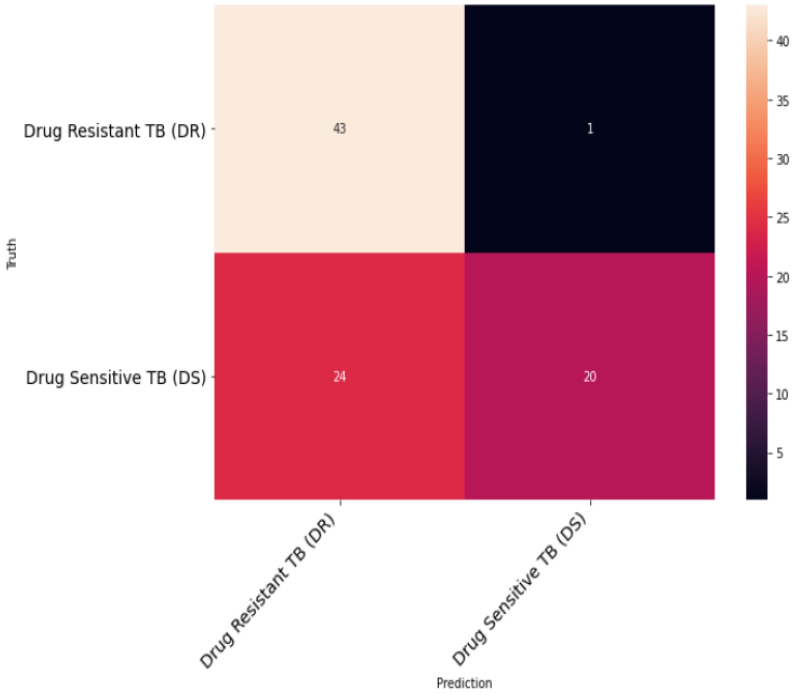


Figure 3: Confusion Matrix for Drugs resistant test data using Naïve Bayes.

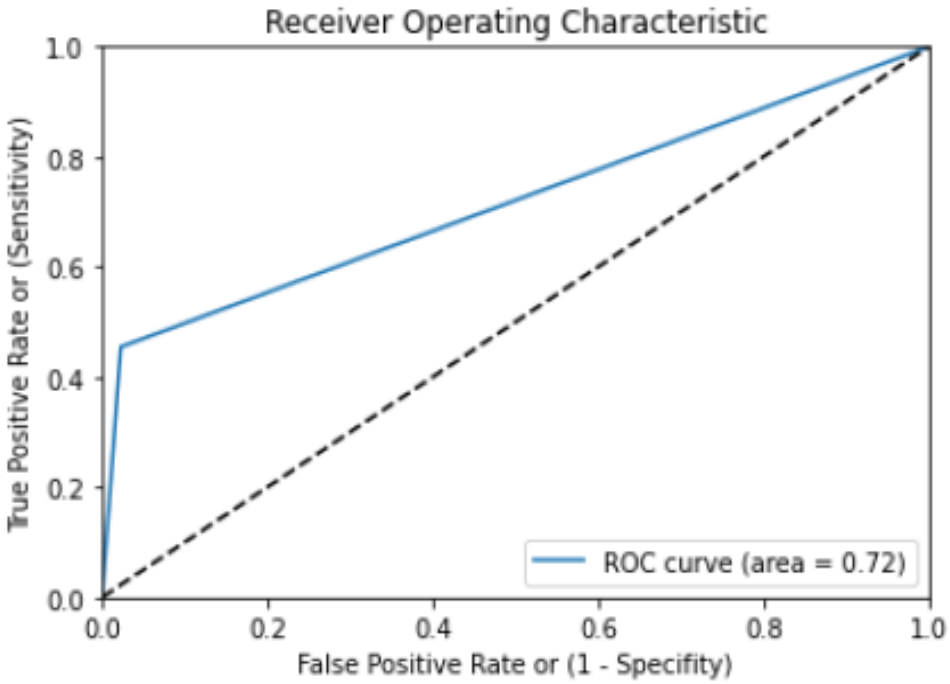


Figure: 4: Drugs Resistant Prediction

References

Alipanah, N., Jarlsberg, L., Miller, C., Linh, N.N., Falzon, D., & Jaramillo, E. (2018). Adherence interventions and outcomes of tuberculosis treatment: A systematic review and meta-ana!ysis of trials and observational studies. *PLoS Medicine, 15*(7), e1002595, http://dx.doi.org/10.l371/iounial.pmed.1002595

Allwood, B., van der Zalm, M., Makanda, G., Mortimer, K., Andre, F.S.A., & Uzochukwu, E. (2019). The long shadow post-tuberculosis. *The Lancet Infectious Diseases, 9*(11), 1170-1. http://dx.doi.org/I0.1016/S1473-3099(19)30564-X.

Allwood, B.W., Van Der Zalm, M.M., Amaral, A.F.S., Byrne, A., Datta, S., & Egere, U. (2020). Post tuberculosis king health: Perspectives from the First International Symposium. *International Journal of Tuberculosis and Lung Disease, 24*(8), 820-8, http://dx.doi.org/l0.5588/ijtld.20.0067

Badeji, B. (2018). Bayesian Classification Model in Predicting Tuberculosis Infection. *Journal of Computer Engineering, 20*(4), 06-16. http://dx.doi.org/10.9790/0661-2004010616

Barathi, A., & Deephankuma, E. (2014). Survey on Classification Techniques in Data Mining. *International Journal on Recent and Innovation Trends in Computing and Communication, 2*(7)

Bisimwa, B.C., Nachega, J.B., Warren, R.M., Theron, G., Metcalfe, J.Z., & Shah, M. (2020). Xpert Mycobacterium tuberculosis/Rifampic in-Detected Rifampicin Resistance is a Suboptimal Surrogate for Multi drug- resistant Tuberculosis in Eastern Democratic Republic of the Congo: Diagnostic and Clinical Implications. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 26.* http://dx.doi.org/10.1093/cid/ciaa873

Bulabula, A.N.H., Nelson, J.A., Musafiri, E.M., Machekano, R., Sam-Agudu, N.A., Diacon, A.H. (2019). Prevalence, Predictors, and Successful Treatment Outcomes of Xpert MTB/ RIF-identified Rifampicin-resistant Tuberculosis in Post-conflict Eastern Democratic Republic of the Congo, 2012-2017: A Retrospective Province-Wide Cohort Study. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 69*(8), 1278-87. htttp://dx,doi.org.1093cid/ciy1105

Chakaya, J., Kirenga, B., Getahun, H. (2016). Long term complications after completion of pulmonary tuberculosis treatment: A quest for a public health approach. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, 3*, 10-2. http://dx.doi.org/10.1016/j.jctube.2016.03.001

Cherian, V., & Bindu. M. S. (2017). Heart Disease Prediction using Naive Bayes algorithm and laplace smoothing technique. *International Journal of Computer Science Trends and Technology (IJCST), 5*(2), 68-73.

Cruz, J. A., & Wishart, D. S. (2006). Application of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2*, 59

Diacon, A.H., Pym, A., Grobusch, M.P., de Los Rios, J.M., Gotuzzo, E., & Vasilyeva, I. (2014). Multidrug-Resistant Tuberculosis and Culture Conversion witk Bedaquiline. *New England Journal of Medicine,371*, 723- 32, http://dx.doi.org/1I.i 056/ NEJMoal313865

Garcia-Basteiro AL, Brew J, Williams B, Borgdorff M, Cobelens F. What is the true tuberculosis mortality burden? Differences in estimates by the World Health Organization and the Global Burden of Disease study. *International Journal of Epidemiology, 47*(5), 1549-60, http://dx.doi.org/10.1093/ije/dyyI44

Gler, M.T., Skripconoka, V., Sanchez-Garavito, E., Xiao, H., Cabrera-Rivero, J.L., VargasVasquez, D.E. (2012). Delamanid for multidrug-resistant pulmonary tuberculosis. *New England Journal of Medicine, 366,* 2151-60. http://dx.doi.org/10.1056/NEJMoall 12433/

Grace AG, Mittal A, Jain S, Tripathy JP,Satyanarayana S, Tharyan P, et al. Shortened treatment regimens versus the standard regimen for drug-sensitive pulmonary tuberculosis.

Parwati, C.G., Farid, M.N., Nasution, H.S., Sulisty, Basri, C., Lolong, D. (2020). Estimation of subnational tuberculosis burden: Generation and application of a new tool in Indonesia. *International Journal of Tuberculosis and Lung Disease, 24*(2), 250-7. http://dx.doi.org/10.5588/ijtld.l9.0139/

Raj. T. F. M., & Prasanna. S. (2013). Implementation of ML using Naive Bayes algorithm for identifying disease-treatment relation in bio-science text. *Research Journal of Applied Sciences. Engineering and Technology, 5*(2), 421-426.

Razzak, M. I. (2015). Automatic detection and Classification of malaria parasite. *International Journal of Biometrics and Bioinformatics (IJBB), 9*(1), 1-12.

Romanowski, K., Baumann, B., Basham, C.A., Ahmad Khan, F., Fox, G.J., & Johnston, J.C. (2019). Longterm all-cause mortality in people treated for tuberculosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases, 19(*10), 1129-37. http://dx.doi. org/10.1016/81473-3099(19)30309-3/

Sharma, D., & Kumar, N. (2017). A Review on Machine Learning Algorithms Tasks and Applications. *International Journal of Advance Research in Computer Engineering and Technology*, *6*(10), 1548-1552.

Simon, A., Deo, M., Venkatesan, and S.. Babu. R. (2015). An overview of Machine Learning and its Applications. *International of Electrical Sciences and Engineering (IJESE), 1*(1). 2224.

Srinath, K. R. (2018). Python - The Fastest Growing Programming Language. *International Journal of Engineering and Technology, 4*(12), 354-357

Sterling, T.R., Villarino, M.E., Borisov,. A.S., Shang, N., Gordin, F., Bliven-Stzemore, E. (2011). Three Months of Rifapentlne and Isoniazid for Latent Tuberculosis Infection. *New England Journal of Medicine, 365,* 2155-66, http://dx.doi.org/10.1056/nejmoal 104875.

Subbaraman, R., Nathavitharana, R.R., Mayer, K.H., Satyanarayana, S., Chadha, V.K., & Arinaminpathy, N. (n.d.). Constructing care cascades for active tuberculosis: A strategy for program.

Trebucq, A., Schwoebel, V. (2016). Numbers of tuberculosis cases: Dreams and reality. *International Journal of Tuberculosis and Lung Disease, 20*(10), 1288-92. http://dx.doi.org/10.5588/ijtld. 35.0873

Volmink, J., Garner, P. (2007). Directly observed therapy for treating tuberculosis. *Cochrane Database of Systematic Reviews*. http://dx.doi.org/10.1002/l4651858.CD00334 3.pub.3

WHO. Global Task Force on TB Impact Measurement, World Heal Organ. 2019. https//www.who.int/tb/areas-of-work/monitoring-evaliiatioiVinipact_measurement_taskforce/ed

Zakhmi, R., & Arora, J. (2016). Review on Tuberculosis Detection Using Data Mining Techniques. *International Journal of Engineering and Technology, 3*(4), 840-843.

Zumla, A., Marais, B.J., McHugh, T.D., Maeurer, M., Zumla, Adam., & Kapata, N. (2020). COVID-19 and tuberculosis-threats and opportunities. *International Journal of Tuberculosis and Lung Disease, 24*(8), 757-60, http://dx.doi.org/10.5588/ijtld.20.0387