A Classification Model for the Risk of Tuberculosis in Nigeria Using Data Mining Approach

1Egejuru, N.C., 2Fajimi, B.O., 3Omodunbi, T.O. & 4Idowu, P.A.
1,2Department of Computer Science, Hallmark University, Ijebu-Itele, Ogun State, Nigeria.
3,4Department of Computer Science and Engineering, Obafemi Awolowo University, Nigeria
Corresponding Author E-mail: paidowu1@yahoo.com

ABSTRACT

This study developed a model that can be used by medical non-experts for the early detection of the risk of tuberculosis based on the information about associated risk factors. The study identified the risk factors of tuberculosis, formulated and simulated the fuzzy logic model for predicting the risk of tuberculosis in Nigeria. This study performed structured interview performed with medical experts in order to identify the factors associated with the risk of tuberculosis following a review of relevant literature. The fuzzy logic model was formulated for the risk of tuberculosis using triangular membership functions to formulate the input risk factors and target risk. The inference rules of the fuzzy logic model was formulated using IF-THEN statements that associate the risk factors with the target variables. The fuzzy logic model was simulated using MATLAB® R2015a Fuzzy Logic Toolbox. The result showed that the eight risks factors which were Age of patient, HIV infection, History of TB infection, Exposure to relatives infected with TB, Body Mass Index (BMI), History of Renal Disease, Migration to TB Infected Area, Working in TB Treatment Center with a target class which was allowed the linguistic labels No risk, Low risk, Moderate risk and High risk were used to infer 768 rules in the fuzzy inference Engine. The inferred rules were simulated in MATLAB environment which was also used to confirm the validity of the inferred rules. The study concluded that it is better and cost effective to detect tuberculosis using the associated risk factors as a preventive approach of controlling the menace of tuberculosis rather than implementing a corrective approach for a patient that is already exhibiting symptoms of tuberculosis, which might require taking laboratory tests which might be expensive or unavailable.

Keywords: Tuberculosis, Predictive Model, Data Mining, Fuzzy Logic, Classification
1. INTRODUCTION

Tuberculosis (TB) is a life-threatening communicable disease that has serious health implications around the world (WHO, 2018). In 2013, approximately 90 million people contracted TB, about 1.5 million people died from TB, 360,000 of whom were HIV positive (WHO, 2014). TB was adjudged to be the largest single cause of adult illness from an infectious disease caused by the bacteria called *Mycobacterium tuberculosis* (Omisore *et al*., 2015). TB in most cases affects the lungs and this is referred to as Pulmonary Tuberculosis (PTB), this is called Extra-Pulmonary Tuberculosis when it affects other parts of the body besides the lungs (WHO, 2018). TB is contagious and airborne. In 2017, it was one of the top 10 major causes of mortality worldwide. It was also the leading killer of people with HIV and a major cause of deaths relating to antimicrobial resistance (WHO, 2017). A big percentage of people living with HIV are also suffering from Tuberculosis, According to records, in 2007, there were 465,000 reported cases of TB among people living with HIV, of whom 84% were on antiretroviral therapy (WHO, 2018). According to WHO, in 2017, 1.6 million people died from TB, including 0.3 million among people with HIV. There were an estimated 10 million new (incident) TB cases worldwide, of which 5.8 million were men, 3.2 million were women and 1 million were children. People living with HIV accounted for 9% of the total. Eight countries accounted for 66% of the new cases: India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and South Africa (WHO, 2018).

The U.S. Agency for International Development (USAID) in 2014 stated that Nigeria had a population of 177 million and it was among the highest tuberculosis (TB) burden countries in the world with 570,000 of new TB cases each year as well as one of the top 10 highest multidrug-resistant (MDRTB) countries globally. The United States Embassy in Nigeria (2012) declared that in 2010, Lagos, Kano, and Oyo states had the highest TB prevalence rate. Other states experienced a drop in cases notified, resulting in a 4% overall decline in 2010. Oyo increased by 46.5% from 2008 to 2010 while Benue has a high TB burden which is attributable to a high HIV prevalence. The Stop TB initiative in Nigeria has provided free TB diagnosis and treatment (Federal Ministry of Health, Nigeria., 2014). Despite this program, there was an incidence rate of 322 per 100,000 population in 2015 and Nigeria was among the 6 countries that accounted for 60% of the total TB burden in the world (WHO, 2017).

According to Goldman (2018), risk factors for TB includes poverty, HIV infection, homelessness, being in jail or prison (where close contact can spread infection), substance abuse, taking medication that weakens the immune system, kidney disease and diabetes, organ transplants, working in healthcare, air pollution, cancer, smoking tobacco, pregnancy, age (specifically babies, young children, and elderly people). Data mining is the field of computational sciences that deals with extracting trends and potentially useful information from large volumes of data by using techniques such as clustering, classification, association, regression, generalization, characterization, evolution, pattern matching, data visualization and meta-rule guided mining (Liao *et al*., 2012; Gera *et al*., 2015; Idowu *et al*., 2015). Data mining enables firm and organization decisions by assembling, accumulating, analyzing and accessing corporate data.

It uses variety of tools like query and reporting tools, analytical processing tools and decision support system (DSS) tools (Rehman, 2017). Data mining is one of a number of analytical tools which allows which allow users to search and analyze data from many different sources and transform it into useful information that can aid decision-making (Khan *et al*., 2014). Data mining uses the trends and patterns extracted from the analysis of large data sets to predict or forecast the likelihood of future events (Crocket *et al*., 2017). Data mining has been applied in a number of fields such as sales retails, bioinformatics, counter-terrorism, education and medicine (Baker and de Carvalho, 2008).
Generally, data mining can be classified into tasks of description and prediction, description aims at finding patterns that can be easily interpreted by humans and associations, after considering the data as a whole and constructing a model. According to Idowu et al. (2015) fuzzy logic systems have the ability to decide and control a system using the knowledge of an expert. There is a need to develop a model to forecast likelihood of tuberculosis in Nigeria which will assist in knowing the possibility of having tuberculosis.

2. RELATED WORKS

Soundararajan et al. (2006) used fuzzy logic algorithms to develop a decision support system to find the probable class of TB a patient may have. A Rule-based fuzzy diagnostic decision support system was used to assign class labels for tuberculosis and fuzzy logic was used for the class assignment process. TB symptoms and class details are updated in a rule based system and learning and testing operations were performed by this process. The designed system was limited to pulmonary physicians that were focusing on tuberculosis and patients that have already been diagnosed with TB. The proposed system suits the need of pulmonary physicians and reduce the time consumed in making diagnosis, however, it cannot predict TB from a range of other possible diseases and in this case a corrective approach can only be possible because the patient already has TB.

Walia et al. (2015) used data mining techniques to develop a fuzzy expert system for the diagnosis of TB. The research focus on the development of system architecture and algorithm used to find the stage of tuberculosis a patient may have. Mamdani’s MAX-MIN fuzzy inference engine was used to infer from the rules developed. This resulted in the establishment of degree influencing input variables on the output. The technique allows for less, mild, moderate, high, yes and no symptoms to be applied in order to get the estimation result. The nine input variables considered were Coughing, Chest pain, Haemoptysis, Fever, loss of Appetite, Smoking addiction, BCG Vaccine, Weight loss, Malaise. This work presented diagnosability for TB and formalized reasoning in a rule-based system. The data set of 35 patients were collected from government health clinic in order to obtain the best prediction model for TB and the fuzzy inference system was constructed with 9 inputs. This system will aid medical experts in the diagnosis of TB, however, it cannot be used to implement a preventive approach to tackling tuberculosis.

Gumpy et al (2018) used data mining techniques to develop a Neuro-fuzzy for diagnosing TB. The system was designed to accept symptoms as inputs and are used to automatically generate rules that are fed into the knowledge base where the system will use it to make decisions and draw a conclusion. MATLAB 7.0 was used to implement this experiment using fuzzy logic and neural network toolbox. In the experiment linguistic variables were evaluated using Gaussian membership function. The developed system will offer potential assistance to medical practitioners in making prompt and accurate decisions during the diagnosis of tuberculosis. An emblematic approach using Neuro-fuzzy methodology was presented that describes to forecast the presence of Mycobacterium and provides support for researches in related fields.

Idowu et al. (2013) used data mining techniques to develop a mathematical model to predict the trends in six (6) immunize-able diseases that affects children under the ages 0-5. In this experiment three (3) data mining techniques were put together in order to investigate the occurrence of viral disease such as Tuberculosis, Measles, Poliomyelitis and Hepatitis B and contagious bacteria such as Tuberculosis and Whooping Cough. Three data mining techniques that were used to compose the mathematical model were Artificial Neural Networks (ANN), Decision Tree Algorithm and Naïve Bayes Classifier using MATLAB’s Toolbox.
Hidden trends and knowledge within database were discovered by means of these data mining techniques which can facilitate the prediction of future disease occurrence in tested locations. Results obtained showed that diseases have peak periods depending on their epidemicity, therefore it is necessary to administer immunization to the right places at the right time. This created model will enhance the effectiveness of routine immunization in Nigeria. The gap in this work is that it does not cater for other age groups from age 5 and above.

Goni et al. (2018) using data mining approach developed an intelligent system for diagnosing TB using adaptive neuro-fuzzy method. Eleven symptoms of TB which are persistent cough for more than two weeks, cough with blood, weight loss, tiredness, chest pain, fever, difficulty in breathing, loss of appetite, lymph node enlargement, history of TB contact and night sweat are assigned with weights which were categorized based on the level of severity as mild, moderate, severe and very severe, yes and no which serve as inputs to the adaptive neuro-fuzzy inference system (ANFIS). MATLAB 7.0 is used to implement this experiment, Trapezoidal Membership function was used, back propagation algorithm was used for training and testing, the error obtain is 0.41777 at epoch 2 which shows that the training performance is exactly 99.58223 and testing performance of the system are 99.58197 at epoch 2. The experiment was tried on different epoch and it was found that at epoch 2 yielded better result than others and minimized error to a minimum level. It was observed that in using ANFIS the learning duration is shorter and excellent results were obtained. However, it cannot be used to implement a preventive approach to tackling tuberculosis.

3. METHODS

3.1 Identification of Associated Risk factors
Following the process of the review of related works, a number of associated risk factors were identified which were validated by four medical experts with more than 10 years’ experience. The risk factors identified by the doctors interviewed were poverty, HIV infection, homelessness, being in jail or prison (where close contact can spread infection), substance abuse, taking medication that weakens the immune system, history of renal disease, body mass index, diabetes, organ transplants, working in TB treatment center, air pollution, cancer, smoking tobacco, pregnancy, age (specifically babies, young children, and elderly people) alcohol, malnutrition. This work was limited to the eight risk factors which includes Age (adult, teenager, aged, child), HIV infection (no, yes), history of TB infection (no, yes), exposure to relatives infected with TB (no, yes), body mass index (normal, underweight, overweight), history of renal disease (no, yes), migration to TB infected area (no, yes), working in TB treatment center (no, yes).

Table 3.1 describes a summary of the crisp values, linguistic variables of each identified variables alongside the corresponding center crisp value of the triangular membership functions. The age of the patient was labelled using the linguistic values: adult, teenager, aged and child with central crisp values of 0, 1, 2 and 3 respectively in the order of increasing risk of TB. Therefore, adults are less likely to have TB compared to teenagers whom are less likely to have TB compared to the aged who are also less likely to have TB compared to children. HIV infection was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients without HIV infection are less likely to have TB compared to patients with HIV infection.
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Linguistic Variable</th>
<th>Crisp Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of Patient</td>
<td>Adult</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Teenager</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Aged</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>3</td>
</tr>
<tr>
<td>HIV Infection</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>History of TB Infection</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Exposure to Relatives Infected with TB</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Body Mass Index (BMI)</td>
<td>Normal</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Underweight</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Overweight</td>
<td>2</td>
</tr>
<tr>
<td>History of Renal Disease</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Migration to TB Infected Area</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Working in TB Treatment Center</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Risk of TB</td>
<td>No Risk</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Moderate Risk</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>High Risk</td>
<td>3</td>
</tr>
</tbody>
</table>

History of TB infection was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients without a history of TB infection are less likely to have TB compared to patients with a history of HIV infection. Exposure to relatives infected with TB was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients who are not exposed to relatives with TB are less likely to have TB compared to patients who are exposed to relatives with TB infection. The Body Mass Index (BMI) of a patient was labelled using the linguistic values: Normal, Underweight and Overweight with central crisp values of 0, 1 and 2 respectively in the order of increasing risk of TB.
Therefore, patients with normal BMI are less likely to have TB compared to patients with overweight BMI. History of renal disease was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients who have no history of renal disease are less likely to have TB compared to patients with a history of renal disease. Migration to TB infected area was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients who have never migrated to TB infected area are less likely to have TB compared to patients who have migrated to TB infected areas. Working in TB treatment centers was labelled using the linguistic values: no and yes with central crisp values of 0 and 1 respectively in the order of increasing risk of TB. Therefore, patients who do not work in TB treatment centers are less likely to have TB compared to patients who do not work in TB treatment centers.

The risk of tuberculosis was determined as a cumulative of the crisp values assigned to the label of the risk factors used to assess the risk of tuberculosis. Thus, the value of the sum of the crisp values of all the labels identifying response to a risk factor was required for assessing the interval of risk to which tuberculosis is classified. The risk of tuberculosis was classified into 4 linguistic variables, namely: No Risk, Low Risk, Moderate Risk and High Risk with crisp values of 0, 1, 2, 3 and 4 respectively. Following the identification of the variables that were proposed for this study, the membership function that was used to formulate the fuzzy logic model was determined.

### 3.2 Method of Fuzzy Logic Model Formulation for Risk of Tuberculosis (TB)

For the purpose of developing a classification model for the risk of tuberculosis using fuzzy logic theory, each variable identified was fuzzified using a triangular membership function. The triangular membership function required the provision of 3 parameters which consisted of the left-hand base of triangle \((a)\), the central apex of the triangle \((b)\) and the right-hand base of the triangle \((c)\). The values \((a, b, c)\) of the triangular membership function corresponded to an interval of \(a \leq b \leq c\) such that the parameters are numeric valued.

The interval of this parameter was used to define the crisp interval within which each crisp value required for calling the linguistic variable was assigned. As a result of this, since there were 2, 3 or 4 linguistic variables defined for each risk factor identified then there were 2, 3 or 4 triangular membership functions such that one was assigned to each linguistic variable identified for each risk factor as appropriate. Therefore, 2, 3 or 4 triangular membership functions were formulated for each risk factor that was identified in this study based on the mathematical expression in equation (3.1). The expression shows how the triangular membership function was used to formulate the label of a variable called \(variable\_label\) by fitting a numerical value \(x\) into a crisp interval of \((a, b, c)\).

\[
Variable\_label(x; a, b, c) = \begin{cases} 
0; x \leq a \\
\frac{x - a}{b - a}; a < x \leq b \\
\frac{c - x}{c - b}; b < x \leq c \\
0; x > c
\end{cases}
\] (3.1)

Using 2, 3 or 4 triangular membership functions, the labels of the identified risk factors were formulated using the crisp intervals of \([-0.5, 0.5]\) with crisp center value of 0, \([0.5, 1.5]\) with crisp center value of 1, \([1.5, 2.5]\) with crisp center of 2 and \([2.5, 3.5]\) with crisp center value of 3 to model the linguistic variables such that the values 0, 1, 2 and 3 are the centers \(b\) of each interval \([a, b]\) as shown in Table 3.1.
3.2.1 Fuzzification of the risk of TB

Following the identification and the fuzzification of the risk factors of tuberculosis, there was a need to formulate the target variable that was used to define the risk of tuberculosis. The triangular membership function was used to formulate the fuzzy logic model for the target variable by assigning crisp values of 0, 1, 2 and 3 to the target class labels, namely: No risk, low risk, Moderate risk and High risk using the intervals (-0.5, 0.5), (0.5, 1.5), (1.5, 2.5) and (2.5, 3.5) respectively.

Therefore, four (4) triangular membership functions were used to formulate the fuzzy logic model required to describe the 4 labels of the target class that was used to describe the risk of tuberculosis using the identified crisp as shown in table 3.2. Using the description provided in Table 3.3, the relationship between the risk factors and the risk of tuberculosis was proposed using the fuzzy inference system.

Table 3.2: Description of Crisp Intervals used during Fuzzy Model Formulation

<table>
<thead>
<tr>
<th>Crisp Value</th>
<th>Interval</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>[-0.5, 0.5]</td>
<td>-0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>[0.5, 1.5]</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>[1.5, 2.5]</td>
<td>1.5</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>[2.5, 3.5]</td>
<td>2.5</td>
<td>3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 3.3: Formulation of the Risk of Tuberculosis

<table>
<thead>
<tr>
<th>Target Class</th>
<th>Interval</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Risk</td>
<td>(-0.5, 0.5)</td>
<td>-0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Low Risk</td>
<td>(0.5, 1.5)</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Moderate Risk</td>
<td>(1.5, 2.5)</td>
<td>1.5</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>High Risk</td>
<td>(2.5, 3.5)</td>
<td>2.5</td>
<td>3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

3.2.2 Fuzzy inference system design

Following the formulation of the fuzzy logic model using triangular membership functions to model the risk factors and the risk of TB, the fuzzy inference engine was implemented. For the purpose of establishing a relationship between the identified non-invasive parameters, rules were inferred from the experts in order to determine the relationship between the parameters identified and the risk of tuberculosis. In order to construct the knowledge base of the classification model using fuzzy logic, a number of IF-THEN rules were used by combining the risk factors as the precedence while the risk of tuberculosis was used as the consequent variable. Using the risk factors that were identified for assessing the risk of tuberculosis, the process of inference rule generation usually follows the fuzzification process.
A typical rule that can be inferred is as follows:

\[
\text{IF (Age = "Adult") AND (HIV Infection = "No") AND (History of TB Infection = "No") AND (Exposure to Relatives with TB = "No") AND (Body Mass Index (BMI) = "Normal") AND (History of Renal Disease = "No") AND (Migration to TB Infected Area = "No") AND (Working in TB treatment Center = "No") THEN (Risk of Tuberculosis = "No Risk")}
\]

The number of rules that were required to be formulated for the fuzzy model were estimated from the product of the number of linguistic variable for each variable. Therefore, since age had 4 linguistic variables, HIV infection had 2 linguistic variables, history of tuberculosis had 2 linguistic variables, exposure to relatives with TB had 2 linguistic variables, BMI had 3 linguistic variables, history of renal disease had 2 linguistic variables, migration to TB infected area had 2 linguistic variables, and working in TB treatment center had 2 linguistic variables. Therefore, the total number of rules were 768 rules.

3.3 Simulation Environment Used

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Toolboxes are comprehensive collections of MATLAB functions, commands and solvers that expand the MATLAB environment to solve particular classes of problems. Fuzzy Logic Toolbox™ provides MATLAB functions, graphical tools, and a Simulink® block for analyzing, designing, and simulating systems based on fuzzy logic.

With respect to this study, there five primary GUI tools (Elements of the MATLAB Fuzzy Logic System) for building, editing, and observing fuzzy inference systems in the toolbox.

i. **Fuzzy Inference System (FIS) Editor** handles the high-level issues for the system: it was used to define the names and number of input and output variables for the proposed model. For this study, 6 input variables and 1 output variable were defined.

ii. **Membership Function Editor** was used to define the shapes of all the membership functions associated with the linguistic variables of each variable. In this study, the triangular membership function was used to formulate all linguistic variables defined for the inputs and output variable. For this study, 2, 3 or 4 membership functions were used to formulate each input variable with 2, 3 or 4 linguistic variables respectively while 4 membership functions were used to formulate the linguistic variables namely: no, low, moderate and high risk of the output variable.

iii. **Rule Editor** was used for editing the different rules that defined the behaviour of the system using a set of IF-THEN statements which combined the identified risk factors with the risk of tuberculosis labels. In this study, 768 rules were constructed using the IF-THEN based on the number of possible combination of the linguistic variables of input variables needed for determining their respective risk of tuberculosis.

iv. **Rule Viewer** is a MATLAB technical computing environment based display of the fuzzy inference diagram which was used as a diagnostic. It shows which rules are active, or how individual membership function shapes are influencing the results. This interface was need for testing the validity of the consistency of the fuzzy model based on the inferred rules constructed for detecting the risk of tuberculosis.
4. RESULTS AND DISCUSSION

4.1 Results of the Formulation of the Fuzzy Model for Risk of Tuberculosis

This section presents the results of the formulation of the fuzzy logic model using triangular membership functions based on the crisp intervals defined for each linguistic variable identified in this study. 2, 3 or 4 triangular membership functions were formulated for the labels of each risk factor identified as appropriate while 4 triangular membership functions were formulated for the labels of the target class which defined the risk of tuberculosis. Therefore, since the same crisp interval was used to define the labels of each risk factor using 2 with centers 0 and 1; 3 with centers 0, 1 and 2 or 4 with centers 0, 1, 2 and 3 using triangular membership function with respective centers. The mathematical representation of the fuzzy logic model used to formulate the predictive model for the risk of TB is presented in the following paragraphs.

As stated earlier, various numbers of triangular membership functions were used to formulate the fuzzy logic model for the linguistic values (or labels) of each risk factors with centers of 0, 1, 2 and 3. Also, the allocation of the values was done based on the increasing effect of the labels of the identified risk factors used in this study. Therefore, the results of the mathematical representation of the fuzzy logic model formulation using the triangular membership function for each of the labels is presented in equation (4.1).

\[ \text{linguisticLabel}_0(x; -0.5, 0, 0.5) = \begin{cases} 0; x \leq -0.5 & \frac{x + 0.5}{0.5} < 0.5 < x \leq 0 & \frac{0.5 - x}{0.5}; 0 < x \leq 0.5 & x > 0.5 \end{cases} \]  
\[ \text{linguisticLabel}_1(x; 0.5, 1, 1.5) = \begin{cases} 0; x \leq 0.5 & \frac{x - 0.5}{0.5} < 1.5 < x \leq 1 & \frac{1.5 - x}{0.5}; 1 < x \leq 1.5 & x > 1.5 \end{cases} \]
\[ \text{linguisticLabel}_2(x; 1.5, 2, 2.5) = \begin{cases} 0; x \leq 1.5 & \frac{x - 1.5}{0.5} < 2.5 < x \leq 2 & \frac{2.5 - x}{0.5}; 2 < x \leq 2.5 & x > 2.5 \end{cases} \]
\[ \text{linguisticLabel}_3(x; 2.5, 3, 3.5) = \begin{cases} 0; x \leq 2.5 & \frac{x - 2.5}{0.5} < 3.5 < x \leq 3 & \frac{3.5 - x}{0.5}; 3 < x \leq 3.5 & x > 3.5 \end{cases} \]

Also, the classification of the risk of tuberculosis was classified into 4 linguistic variables, namely: No risk, Low risk, Moderate risk and High risk using crisp values with centers of 0, 1, 2 and 3 respectively. Using the 4 triangular membership functions stated in equations (4.2a) to (4.2d), the linguistic variables of the risk of tuberculosis was formulated.

\[ \text{Crisp} - \text{no\_risk}(x; -0.5, 0, 0.5) = \begin{cases} 0; x \leq -0.5 & \frac{x + 0.5}{0.5} < 0.5 < x \leq 0 & \frac{0.5 - x}{0.5}; 0 < x \leq 0.5 & x > 0.5 \end{cases} \]
\[ \text{Crisp} - \text{low\_risk}(x; 0.5, 1, 1.5) = \begin{cases} 0; x \leq 0.5 & \frac{x - 0.5}{0.5} < 1.5 < x \leq 1 & \frac{1.5 - x}{0.5}; 1 < x \leq 1.5 & x > 1.5 \end{cases} \]
4.2 Results of the Simulation of the Fuzzy Model for Risk of Tuberculosis

Using the triangular membership functions stated in equations (4.1a) to (4.1d), the linguistic values (or the labels) of the identified risk factors were simulated. Also, the triangular membership functions stated in equations (4.2a) and (4.2d) were used to simulate the risk of tuberculosis using the MATLAB software. The results of the simulation of the membership functions and of the inference rules used to generate the final .fis file of the Fuzzy Logic Model for the Risk of Tuberculosis is presented in the following sections.

4.2.1 Results of simulation of the fuzzification of input and output variables using the membership function editor

Therefore, the results of the simulation of the model for the age of patient is shown in Figure 4.1 such that the interval [-0.5, 0.5] was used to model adult, interval [0.5, 1.5] was used to model teenager, interval [1.5, 2.5] was used to model child. The results of the simulation of the model for HIV infection is shown in Figure 4.2 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes. The results of the simulation of the model for the history of TB infection is shown in Figure 4.3 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes. The results of the simulation of the model for the exposure to relatives infected with TB is shown in Figure 4.4 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes.

\[
\text{Crisp} - \text{moderate}_\text{risk}(x; 1.5, 2, 2.5) = \begin{cases} 0; & x \leq 1.5 \\ \frac{x - 1.5}{0.5}; & 1.5 < x \leq 2 \\ 2.5 - x \frac{0.5}{0.5}; & 2 < x \leq 2.5 \\ 0; & x > 2.5 \end{cases} \quad (4.3c)
\]

\[
\text{Crisp} - \text{high}_\text{risk}(x; 2.5, 3, 3.5) = \begin{cases} 0; & x \leq 2.5 \\ \frac{x - 2.5}{0.5}; & 2.5 < x \leq 3 \\ 3.5 - x \frac{0.5}{0.5}; & 3 < x \leq 3.5 \\ 0; & x > 3.5 \end{cases} \quad (4.2d)
\]
Figure 4.2: Fuzzification of HIV Infection

Figure 4.3: Fuzzification of History of Tuberculosis Infection
Figure 4.4: Fuzzification of Exposure to Relatives with TB

The results of the simulation of the model for the body mass index (BMI) of patients is shown in Figure 4.5 such that the interval [-0.5, 0.5] was used to model normal, the interval [0.5, 1.5] was used to model underweight while the interval [1.5 2.5] was used to model overweight. The results of the simulation of the model for the history of renal disease is shown in Figure 4.6 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes. The results of the simulation of the model for migrations to areas affected by TB is shown in Figure 4.7 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes. The results of the simulation of the model for working in TB treatment centers is shown in Figure 4.8 such that the interval [-0.5, 0.5] was used to model no while [0.5, 1.5] was used to model yes. The results of the simulation of the model for the risk of tuberculosis is shown in Figure 4.9 such that the interval [-0.5, 0.5] was used to model none, [0.5, 1.5] was used to model low risk, [1.5 2.5] was used to model moderate risk while [2.5 3.5] was high risk.
Figure 4.5: Fuzzification of Body Mass Index

Figure 4.6: Fuzzification of Renal Disease History
Figure 4.7: Fuzzification of Migration to TB Infected Areas

Figure 4.8: Fuzzification of Working in TB treatment Centers
Figure 4.9: Fuzzification of Risk of Tuberculosis (TB)

4.2.2 Results of simulation of Fuzzy Logic Model for Risk of Tuberculosis

Figure 4.10 shows the process of importing the completed source file of the fuzzy model for the risk of tuberculosis using the MATLAB Fuzzy Logic Toolbox. The 768 rules that were inferred for determining the risk of tuberculosis were defined using the rule editor interface. Figure 4.11 shows the complete illustration of the 768 rules that were inferred for determining the risk of tuberculosis. It is clear that each rule inferred is unique and does not contain linguistic variables occurring in the same pattern in any of the rules defined. Therefore, for any given set of rules $r$ and $s$ within the 768 rules there is no rule $r$ that has the set of linguistic variables as another rule $s$. This is also shown by the rule viewer in Figure 4.12. Figure 4.12 displays the graphical region of each variable selected by each rule with respect to the linguistic variables of the risk of tuberculosis.

As shown in the bottom left part of the figure, the crisp values entered were 1, 0, 1, 0, 2, 0, 0, 0 which were consistent with the linguistic values namely: teenager for age, no for HIV infection, yes for history of TB infection, no for exposure to relatives with TB, overweight for BMI, no for history of renal disease, no for migration to TB infected area and no for working in TB treatment center.
According to rule # 250, the combination of these linguistic variables should yield low risk of tuberculosis which amounted to a crisp value of 2 which is the interval of low risk of tuberculosis defined as [1.5, 2.5].
4.3 Discussion

The results of the formulation and simulation of the fuzzy logic model for determining the risk of tuberculosis has been presented. The discussion of the results are presented in terms of the centers of the crisp intervals defined with respect to the linguistic values using triangular membership functions of the risk factors of tuberculosis that were associated with the target class which is the risk of tuberculosis. The vagueness and uncertainties associated with the prediction of the likelihood of tuberculosis are accommodated by the implementation of the Fuzzy Logic model for this task.

In this research 2, 3 or 4 triangular membership function were formulated for the labels of each risk factor identified as appropriate, while 4 triangular membership functions were formulated for the labels of the target class which defined the risk of tuberculosis. The crisp interval used to define the labels of each risk factor were 2 with centers 0 and 1; 3 with centers 0, 1 and 2 or 4 with centers 0, 1, 2 and these values were allocated based on the increasing effect of the labels of the identified risk factors used in this study. Also, the classification of the risk of tuberculosis was classified into 4 linguistic variables, namely: No risk, Low risk, Moderate risk and High risk using crisp values with centers of 0, 1, 2 and 3 respectively.
Using the triangular membership functions stated in equations (4.1a) to (4.1d), the linguistic values (or the labels) of the identified risk factors were simulated. Also, the triangular membership functions stated in equations (4.2a) and (4.2d) were used to simulate the risk of tuberculosis using the MATLAB software.

The MATLAB model used was the Mamdani model version 2.0 which is suitable for discrete variables, the number of input variables were 8 and with one output variable and they were used to generate 768 rules in the Fuzzy Inference Engine. The defuzzification method used is the centroid method. The 768 rules were inferred such that were unique and does not contain linguistic variables occurring in the same pattern in any of the rules defined. Therefore, for any given set of rules \( r \) and \( s \) within the 768 rules there is no rule \( r \) that has the set of linguistic variables as another rule \( s \).

5. CONCLUSION

This study focused on the development a predictive model that can be used by medical non-experts for the early detection of the risk of tuberculosis in patients based on the information about associated risk factors. The study identified the risk factors that were associated with the risk of tuberculosis following which the fuzzy membership function were used to formulate linguistic labels and crisp intervals for eight input risk factors and target risk using triangular membership functions. The study formulated 768 inference rules of the fuzzy logic model using IF-THEN statements that associated the risk factors with the target variables which were allowed to take the labels “no risk”, “low risk”, “moderate risk” and “high risk”. The study concluded that a number of risk factors were associated with forecasting the risk of tuberculosis based on the on the inferred rules in the Fuzzy Inference System. The study concluded that it is far better and cost effective to make use of the identified variables associated with the risk of tuberculosis, to formulate and simulate a fuzzy logic model for forecasting the likelihood of tuberculosis disease for an individual, who is not necessarily infected with tuberculosis or exhibiting any of the symptoms of TB. The model formulated can serve as a timely intervention and a preventive approach for dealing with menace of tuberculosis in South-west Nigeria.
REFERENCES


