Visual Analytics of Tuberculosis Detection Rat Performance

Joan Jonathan[1], Camilius Sanga[2], Magesa Mwita[1], Georgies Mgode[3, 4]

[1] Centre for Information and Communication Technology, Sokoine University of Agriculture, P.O Box 3218, Chuo Kikuu, Morogoro, Tanzania

[2]Sokoine National Agricultural Library (SNAL), Sokoine University of Agriculture, P.O.Box 3022, Morogoro, Tanzania

[3]Pest Management Centre, Sokoine University of Agriculture, P.O Box 3110, Chuo Kikuu, Morogoro, Tanzania

[4]APOPO TB Project, Sokoine University of Agriculture, Morogoro, Tanzania

Abstract

The diagnosis of tuberculosis (TB) disease remains a global challenge, and the need for innovative diagnostic approaches is inevitable. Trained African giant pouched rats are the scent TB detection technology for operational research. The adoption of this technology is beneficial to countries with a high TB burden due to its cost-effectiveness and speed than microscopy. However, rats with some factors perform better. Thus, more insights on factors that may affect performance is important to increase rats' TB detection performance. This paper intends to provide understanding on the factors that influence rats TB detection performance using visual analytics approach. Visual analytics provide insight of data through the combination of computational predictive models and interactive visualizations. Three algorithms such as Decision tree, Random Forest and Naive Bayes were used to predict the factors that influence rats TB detection performance. Hence, our study found that age is the most significant factor, and rats of ages between 3.1 to 6 years portrayed potentiality. The algorithms were validated using the same test data to check their prediction accuracy. The accuracy check showed that the random forest outperforms with an accuracy of 78.82% than the two. However, their accuracies difference is small. The study findings may help rats TB trainers, researchers in rats TB and Information systems, and decision makers to improve detection performance. This study recommends further research that incorporates gender factors and a large sample size.

Keywords: Data mining in healthcare, African giant pouched rats, Classification Technique in Tuberculosis diagnosis

*Correspondence: Joan Jonathan (joanjonathan@sua.ac.tz)

DOI: 10.5210/ojphi.v13i2.11465

1 Introduction

Tuberculosis (TB) is one of the life-threatening infectious diseases causing death worldwide [1]. The WHO report [2] shows that 10 million people are infected with TB each year. Microscopy is the widely used TB diagnostic tool in developing countries despite its lower sensitivity [3, 4]. Nucleic acid-based test such as GeneXpert MTB/RIF is now in use with higher sensitivity and specificity than microscopy. However, its full roll-out and utility is limited to some areas. There is a need for new cheap and rapid diagnostic approaches to enhance TB case detection in countries with a high TB burden. Since 2007, Anti- Personnel Landmines Detection Product Development (APOPO) and Sokoine University of Agriculture (SUA) have been exploring the potential application of the trained African giant pouched rats (HeroRATS) for detection of pulmonary TB in sputum samples [7]. Trained rats retest heat inactivated sputum samples after smear microscope and other hospital tests to detect missed TB cases. The study conducted by Poling et al. [6] evaluated sputum 21,600 from Tanzanians and 9,048 from Mozambicans which was already screened by the microscope. However, after the evaluation by rats it was revealed that there were 1,412 new patients with active TB in Tanzania and 645 new patients in Mozambique. The new detected cases increase the detection rate by 39% in Tanzania and 53% in Mozambique when compared to smear microscopy, the standard diagnostic for TB. Furthermore, trained rats increase pediatric TB detection by 68% as the additional of 23 children patients who tested TB positive from 982 children sputum samples [7]. The endorsed conventional tests such as concentrated smear microscopy offer higher sensitivity than the direct microscopy and thus are used to confirm detection rat results before patients start treatment [1]. These scent detection rats detect the specific volatile organic compounds produced by Mycobacterium tuberculosis bacterium that causes TB [5].

The usefulness of this scent detection technology is also due to the rats' rapid diagnostic speed in which rats can test up to 100 samples in 20 minutes that will take a laboratory technician about four days when using the microscopy to examine the recommended 30 samples per day [3]. The detection performance of trained rats may depend on rats' characteristics, which include age, sex, time of day, bacteria count, and weight [8]. The study conducted by Ellis et al. [3] identified that older rats do better than less old rats also time of day of training influences the detection performance. However, there were no significant differences between male and female detection performance. In another study of Mgode et al. [7] rats can detect TB in samples with a lower number of bacteria count likely to be missed by microscope. Based on the experience older rats and weighty rats have low detection performance. And as such, there is no empirical evidence on the main influencing factors, and the trend of their impact is not clear. Therefore, this study intends to use data mining techniques to predict factors that influence TB detection rats' performance. The scent detection technology at APOPO produces massive data that need an in-depth look to obtain insights on various valuable information using data mining techniques. Data mining is a useful field for discovering interesting patterns and information from multidimensional data. In healthcare, data mining techniques such as classification, clustering, and association are most used to solve health problems [9]. Most of the studies [10, 11] used classification technique in the diagnosis of tuberculosis to categorize and find the relationships among the manipulated variables. Furthermore, the study of Asha et al. [12] propose that the classification technique helps the health sectors to have better decision toward their operations.

1.1 Objective of the study

The objective of the study was to use Data Mining techniques to predict factors associated with TB detection performance of the rats. The goal is to provide a deeper understanding of the main factors influencing detection performance as well support decision making, improving human health, and scaling up of the detection technology. To further contribute to this body of knowledge the study focused on the following three different hypotheses:

Null hypothesis:

i) There is no measurable accuracy difference between the three algorithms of a classification technique in predicting the factors associated with TB detection performance in rats
ii) There is no measurable difference between different predicted factors of rats that affect TB detection performance
iii) There is no measurable difference between the ability of rats in TB detection performance

Alternate hypothesis:

i) There is a measurable accuracy difference between the three algorithms of a classification technique in predicting the factors associated with TB detection performance in rats
ii) There is a measurable difference between different predicted factors of rats that affect TB detection performance
iii) There is a measurable difference between the ability of rats in TB detection performance

2 Methods

This study used the Cross-Industry Standard Process for Data Mining (CRISP-DM) as an analytical framework for knowledge discovery. CRISP-DM involves a systematic and organized approach in the data mining process [9]. CRISP-DM consists of six phases, namely: (1) Business Understanding (2) Data Understanding (3) Data Preparation (4) Model Building (5) Testing and Evaluation (6) Deployment. These phases are explained in detail underneath.

2.1 Business understanding phase

This phase dealt with what APOPO TB center needs from a business perspective. APOPO is a Belgian Non-Government Organization (NGO) based in Morogoro, Tanzania, which aims at using rat odor detection technology to solve humanitarian problems. Extracting knowledge of the application domain was useful to create an understanding of the aim, requirements, and constraints of the center.

2.2 Data understanding phase

This phase focused on the access, description, and identification of the relevant data from the APOPO TB center. The given rats' detection performance data comprised of two datasets: Detection Rats Data and RAT_WEIGHT. Detection Rats Data dataset composed of 18 detection

performance variables (17 independent and 1 dependent) and 471,133 observations from 2011 to 2019 years. Meanwhile, the RAT_WEIGHT dataset contained four (4) independent detection performance variables and 1438 records from 2012 to 2019. This dataset also contained the five female rats. However, the fifth rat had no corresponding detection performance variables and thus disqualified. Hence, this study used the four (4) female rats' data from 2014 to 2018 years. Table 1 shows the two datasets with their respective variables.

Table 1: Rats Datasets Description

DetectionRatsDataDataset Description				
Number	Variable name	Data type	Description	Variable type
1	DOTS_NAME	String	Name of the DOTS center	Independent
2	DOTS_PATIENTS_NUMBER	Integer	Number of patients from DOTS center	Independent
3	ENTRY_YEAR	Integer	Year when patient attend DOTS center	Independent
4	ID_SAMPLE	Integer	Identification of the sample	Independent
5	ID_BL_DOTS	Integer	Identification of bacteria level from DOTS center	Independent
6	HIT	Boolean	TB detection rat performance (categorical variable)	Dependent
7	ID_BL_APOPO	Integer	Identification of bacteria level from APOPO center	Independent
8	ID_CONFIGURATION	Integer	Identification of the cage during training	Independent
9	ID_BL_FM	Integer	Identification of bacteria level by fluorescence microscope	Independent
10	ID_EVALUATION_SESSION	Integer	Identification of evaluation session	Independent

11	SESSION_DATE	Date	Date when a session performed	Independent
12	ID_RAT	Integer	Identification of the rat	Independent
13	RAT_NAME	String	Name of rat	Independent
14	GENDER	String	Sex of rat	Independent
15	Age	Integer	Age of rat	Independent
16	START_TIME	DateTime	Date and time when detection task started	Independent
17	END_TIME	DateTime	Date and time when detection task ended	Independent
18	DOB	Date	Date when rat was born	Independent
RAT_WEIGHT Dataset Description				
Number	Variable name	Data Type	Description	Variable type
1	ID_RAT	Integer	Identification of rat	Independent
2	RAT_NAME	String	Name of rat	Independent
3	WEIGHT_DATE	Date	Date when weight of the rat was measured	Independent
4	WEIGHT	Integer	Weight of the rat	Independent

2.3 Data preparation

Following the data understanding, this phase prepared the data into the well-formed data using the four main steps [9]. These steps were data consolidation, data cleaning, data transformation, and data reduction.

Data consolidation step, data were accessed from APOPO TB center based in Morogoro, Tanzania. The data were integrated into a single file to ease the data mining process. Data cleaning step, irrelevant variables, and empty rows were removed to prevent inconsistencies and outliers. As a result, the prepared data had 365,843 observations from 471,133 observations. Data transformation step, the new three (3) detection performance variables were created. These variables are Rat_Av_Weight_Per_Year, Session_Start_Time, and Session_Completion_Time

as shown in Table 2. It is important to note that all variables were converted to the required data type. Data reduction step, the prepared observations were reduced from 365,843 to 200,000 to ease the analysis by using random sampling method. The study conducted by Czarnowski et al. [13] shows that data reduction focused on reducing the volume of dataset while maintaining the integrity of data since the reduced dataset has the same acceptable amount of information as the original dataset. The main four steps were governed by R-language and RStudio. RStudio is a data mining tool and an integrated development environment for R, is a free programming language with extensive modeling and quality graphs resources [14]. Considering the given four (4) and three (3) new created detection performance variables, seven (7) prepared detection performance variables and 200,000 observations were used for analysis as shown in Table 2. Where six (6) are independent variables and one (1) is the dependent variable. Table 2 and Table 3 show description and descriptive statistical summary information of the variables used to build predictive models respectively.

Table 2: Description of the variables used to build predictive models

Variable	Description	Data type	Variable type	Values
DOTS_Name	Name of DOTS center	Factor	Independent	DOTS centre name
Rat_Name	Name of rat	Factor	Independent	Rat 1, Rat 2, Rat 3, Rat 4
Rat_Age	Age of rat in years	Numeric	Independent	0.79, 2.04, 3.22
Rat_Av_Weight_Per_Year	Average weight of rat per year	Numeric	Independent	846.35, 866.80
Session_Start_Time	Time of day when detection session started in 24 hours	Integer	Independent	12,13,14
Session_Completion_Time	Differences in minutes between session start time and session end time	Numeric	Independent	1,2,3
Performance	Performance of rat during the session	Factor	Dependent	TRUE, FALSE

Table 3: Descriptive statistical summary information of the independent variables used to build predictive models

	Age	Av_Weight_Per_Year (g)	Session_Start_Time (hours)	Session_Completion_Time (in min)
Min	0.79	843.7	8.00	1.0
Max	7.95	1054.8	18.00	129.0
Mean	3.83	899.4	12.16	10.5
Median	3.71	866.8	12.00	10.0
Range	7.16	211.16	10.00	128.0
SD	1.72	84.44	1.67	4.89
CI	0.0056	0.027	0.0041	0.016

Table 3 depicts statistical summary information of the independent variables where the younger and older rats have ages of 0.79 and 7.95 years respectively with the mean, median, range, SD, and CI age of 3.83, 3.71, 7.16, 1.72, and 0.0056 years. Moreover, the rats' lowest and highest average weight per year are 843.7g and 1054.8g respectively, with the mean, median, range, SD, and CI age of 899.4, 866.8g, 211.16kg, 84.44 kg, and 0.027 kg. Besides, the table shows that their lowest and highest session start time are 8:00 and 18:00 hours, with the mean, median, range, SD, and CI age of 12:16, 12:00, 10:00,1.67, and 0.0041. Furthermore, the minimum and maximum session completion time is 1 and 129 minutes, with the mean, median, range, SD, and CI age of 10.5, 10.0, 128.0, 4.89, and 0.016 minutes. Since the mean and median are not equal, it manifests that the data used for this analysis lack normal distribution. Table 4 shows number of named rats, and the associated observations by factor of interest.

Table 4: Summary for number of rats, and the associated observations by factor of interest

Rat_Name	Gender	Age	Av_Weight_Per_Year (g)	Session_Start_Time (hours)	Session_Completion_Time (min)	No of observations
Sofia	F	0.7-6.3	846-877	11:00-14:00	4.0-14.0	50448
Catia	F	1-6	846-877	10:00-15:00	8.2-13.0	50271
Happy	F	1-8	844-1055	10:00-15:00	6.4-23.0	50035
Mkuta	F	1-6	844-1055	9:00-16:00	4.5-15.8	49246

The data from Table 4 depicts four female rats used in the analysis with their observations completed during the detection tasks. Sofia completed many numbers of observations compared to all since it started detection tasks early to the age of 0.7 years. Besides, Mkuta has few numbers of observations whereas its large average weight of 1055g may have caused this performance compared to Catia from which both started and end detection tasks at the age of 1-6 years, respectively. Moreover, Happy is older and has a few observations than Sofia and Catia. The large average weight of 1055g and the removal of irrelevant data from 2011 to 2013 and 2019 could have led this since there is a possibility that Happy had many observations in the irrelevant years. Furthermore, the table shows that there is sex inequality in the data given from the DOTS center, since all rats are female. Table 5 shows the dependent variable, and the associated observations used in this analysis.

Table 5: Summary for dependent variable (Performance) and the number of observations detected

Performance	No_Samples
FALSE	157686
TRUE	42314

From Table 5, there is performance inequality in the distribution of rats' detected observations. TRUE observations are far less by 21.2% than FALSE of about 78.8% for all observations. And as such, this analysis used more FALSE than TRUE.

It was also important to examine/measure the association of continuous independent variables with a dichotomous dependent variable (performance). Thus, the logistic regression analysis was used to describe data and explain the relationship between the dependent variable and independent variables as shown in Table 6.

Table 6: Association between Dependent and Independent variables

Variable	*Pr(>|z|)*
Age	< 2e-16
Session_Completion_Time	< 2e-16
Session_Start_Time	2.53e-09
Av_Weight_Per_Year	2.98e-07

With regard to the Table 6, the *Pr (>|z|)* column indicates the p-value corresponding to the z-statistics. The *p-values* for the independent variables are below 0.05, and implies that there is a relationship between independent variables and the dichotomous dependent variable, and variables are statistically significant. However, the data assumption of normality was not achieved since the corresponding values are less than 0.05. The normality was examined by prediction analysis (logistic regression) using the Kolmogorov Smirnov (KS) test under the *z test* statistic.

2.4 Model building

After obtaining the data with the required format, this phase used to select and apply the data mining technique and algorithms based on the nature of the data. This phase applied classification technique to build predictive models that assigned a class for each rat in the given data and predicted the factors that influence rats TB detection performance. Not only that but also the predictive models might be useful to place and predict the new instances (rats) with unknown labels into their respective classes. Classification is a supervised learning technique in data mining and machine learning that learn the relationship or patterns between independent variables (input) and the dependent variable (output) from the past data and classify each data item into a predefined class label.

Before presenting the prepared data to the algorithms, R was entirely used to partition the data (200,000 observations) using a simple split estimation method. The simple split estimation is the most popular method, which divided two-thirds of the data (134,000 observations) in the training data and one-third in the test data (64,000 observations) [9] as shown in Table 7. Therefore, the training data was used to build a predictive model while the test data used to assess the predictive model classification accuracy.

Table 7: Summary of a simple random data splitting

Type of data	Number of observations
Training data	134000
Testing data	66000

The data from Table 7 indicate that this analysis consisted of 67% training data and 33% test data. The training data were given many observations to build the predictive model while test data were used only to assess the performance of the model generated. Despite many classification algorithms used for prediction, this study used Decision Tree, Random Forest, and Naïve Bayes for prediction. A decision tree algorithm is a supervised classification algorithm which generates the decision tree automatically by examining the weight of each variable used to the extent that each leaf node has the same class [11]. Also, it generates rules that are easy to interpret and understand [16]. The decision tree is a tree-shaped diagram comprises many input variables that may have an impact on classifying different patterns. Additionally, it is known as a decision support algorithm which depends on the input to show the possible outcomes [12].

The decision tree was generated by recursively dividing the training data until each division consisted of the variables of the same class or values based on conditions. Following this, a split point used in each node to test the manipulated variables and decide the way to divide the data. The split decision focused on the amount of information a computed variable offered in the class (information gain) and its randomness (entropy). As a result, the variable with the highest information gain and the lowest entropy split and tested. The information gain and entropy determined the decision on the split of data and construction of the decision tree. However, the growth of the decision tree influenced deep learning. Control on the parameters used to overcome this problem through pruning [9]. Pruning is the process of reducing the size of decision trees by removing sections of the tree that provide little power to classify instances.

The generated decision tree consists of a root node, branches, and leaf nodes. The root node is the node at the top of the tree which implies the most important factor responsible for classifying the observations. The branches represent the pattern classification outcome of a test using one of the variables based on conditions. The leaf nodes placed either before or at the end of the decision tree imply the nodes without children. And as such, they identify the last class choice for a pattern. Moreover, the decision tree formed rules (IF-THEN statements) from the root node to the leaf nodes which are easy to interpret and understand. As a result, they enhanced the discovery of exploratory knowledge on the factors that influence rats TB detection performance. Furthermore, the random forest algorithm was also applied to predict the influencing factors and compare the prediction accuracy of the models.

A random forest algorithm is a supervised classification algorithm used to build multiple decision trees called forest in random during the training process. The choice of most of the trees determined the final decision of the algorithm based on the given manipulated variables. There is a relationship between the number of trees in the forest and the results. Thus, many trees, the more accurate the result. The motives behind this algorithm are that it can be used for both classification and regression problems and lowers the risk of overfitting [11]. Overfitting is a modeling error which occurs when the outcome of the analysis is limited only to specific data. As a result, instead of predicting the whole manipulated data, the model predicts only for that set of data [16]. In the random forest algorithm, the process of determining the root node and the splitting of the variable nodes were performed randomly from the training data. During the training, no control of parameter (pruning) involved preventing the decrease of the relationship between trees. However, pruning is of importance for the reduction of complexity in variable computation during the training. As a result, the algorithm handled about 500 trees in the ensemble and identified the error rate based on the training data. Following this, the random forest algorithm predicted the factors for detection by pinpointing the mean decrease in Gini values for each variable. Furthermore, the Naive Bayes algorithm was applied to compare their predictive accuracy by using the same test data, and finding the best algorithm with high classification accuracy rates for the given data.

Naive Bayes is the supervised classification algorithm that uses a probability theory (Bayesian Theorem) to generate the classification model. Moreover, to place an instance in the desired class. This theory supported to calculate a set of probabilities by counting the frequency and values of the manipulated variables from the given data [11]. The Naive Bayes algorithm is a well-performed algorithm owing to its simplicity in execution time. And as such, it can build a final model that can learn rapidly different classification problems [17]. However, this algorithm assumed that all variables were independent of the given data while few real-world applications may agree with this [12]. The main advantages of the Naive Bayes algorithm compared to the other two algorithms are the run-time speed on large and complex datasets. Hence, most healthcare field researchers across the world use this algorithm due to its better speed and accuracy. This algorithm identified a priori probability for the dependent variable and conditional probabilities for every independent variable based on the manipulated data. The Naive Bayes algorithm does not show the weights of each variable included in the classification, but it has been used purposely to compare its prediction performance with the results generated from the decision tree and random forest algorithms [17].

2.5 Testing and Evaluation

This phase was used to test and assess the classification performance of the three generated predictive models. The assessment was based on the accuracy metric to show the predictive accuracy of the model from the confusion matrix. However, the confusion matrix has several assessment measures such as sensitivity and specificity. The confusion matrix is a table used to describe a classification model performance based on the test data. The accuracy measure was used to assess the ability of the models to accurately predict the class label of the test data. The accuracy entailed the matching between actual class labels of the test data and the class labels of the predicted models. The accuracy measurement focused on the accuracy rate, the percentage of test instances that were accurately classified by the predictive model as shown in Table 8. The accuracy (acc) and error rate (err) values of the classification matrix rated the predictive model performance. The error rate (err) implies the fraction of the sum of FALSE positives and FALSE negatives and the sum of the total number of all the predictions made. Table 8 presents the comparison of predictive models classification accuracy rate and error rate between training data and test data for all three algorithms. Since the predictive models learned to classify the rats TB detection performance into TRUE or FALSE, the positive class is a FALSE value since it has many observations of about 157,686 samples as reported in Table 5. Therefore, the following formulas measured the percentage accuracy rate and error rate respectively for the positive class.

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

$$\text{Error rate (err)} = \frac{FP + FN}{TP + FP + TN + FN}$$

TP, TN, FP, FN mean True Positive, True Negative, False Positive and False Negative respectively.

Table 8: Comparison of predictive models' classification accuracy rate between training data and test data for all algorithms

Training data			
Evaluation criteria	Predictive model		
	Decision tree	Random forest	Naive Bayes
Accuracy (%)	78.94%	79.00%	78.86%
Error rate (%)	21.06%	21.00%	21.14%
Correctly classified observations (TP)	105573	105878	105674
Incorrectly classified observations (FN)	28227	28142	28326
Mcnemar's Test *p-value*	<2e-16	<2e-16	<2e-16
Test data			
Evaluation criteria	Predictive model		
	Decision tree	Random forest	Naive Bayes
Accuracy (%)	78.78%	78.82%	78.71%
Error rate (%)	21.22%	21.18%	21.29%
Correctly classified observations (TP)	51997	52019	51946
Incorrectly classified observations (FN)	14003	13981	14054
Mcnemar's Test *p-value*	<2e-16	<2e-16	<2e-16

From Table 8, during the building of the predictive model, the decision tree algorithm correctly classified 105573 observations equal to the accuracy rate of 78.74% and incorrectly classified 28227 observations equal to the error rate of 21.06%. In other hands, random forest correctly classified 105878 observations equal to the accuracy rate of 79.00% and incorrectly classified 28142 observations equal to the error rate of 21.00%. Furthermore, the naïve Bayes correctly classified 105674 observations equal to the accuracy rate of 78.86% and incorrectly classified 28326 observations equal to the error rate of 21.14%. With regards to the test data, the decision tree algorithm correctly classified 51997 observations equal to the accuracy rate of 78.78% and incorrectly classified 14003 observations equal to the error rate of 21.22%. Additionally, random forest correctly classified 52019 observations equal to the accuracy rate of 78.82% and incorrectly classified 13981 observations equal to the error rate of 21.18%. Again, the naïve Bayes correctly classified 51946 observations equal to the accuracy rate of 78.71% and incorrectly classified 14054 observations equal to the error rate of 21.29%. Thus, the random forest algorithm outperforms both during the building of the predictive model and assessing the classification performance. However, the ability to overcome overtraining problem might have led to this. Additionally, data overlapping, and the random nature of the modeling algorithms presumed to affect the overall performance of the three predictive models.

2.6 Deployment

This is the last phase that was used to organize and present the knowledge gained to the end-user for real application using visualization techniques, such as a plot. The knowledge obtained is explicitly aimed at helping users to predict rats' factors that influence TB detection performance and the classes of new data instances (where the class label is unknown). Table 6 pinpoints the association between a dichotomous variable and independent variables. Moreover, Figure 2 shows a variable importance plot used for proper interpretation and ease understanding of the knowledge gained.

3 Results and Analysis

The data mining process aimed to elicit knowledge from the given structured data and present it to the end-user for the real application. And as such, this process was managed by the classification technique and algorithms that helped to learn the relationship between the patterns. However, the classification technique used three algorithms which are decision tree, random forest, and naïve Bayes to build the predictive models. Thus, this section presents results and analysis based on the formulated different three hypotheses as follows:

• *There is no measurable accuracy difference between data mining algorithms in predicting the factors that associated with TB detection performance in rats*

With regards to the first hypothesis, our findings show that there is in fact a measurable difference in between the three data mining algorithms. According to McNemar's Test, the test checked if there was significant difference between the counts in two cells made in both predictive models. by capturing the errors made by both models. Hence, Table 8 shows that the errors made by both models in the test data are not the same, and thus the result of the test is significant and the null hypothesis is rejected. Additionally, the Mcnemar's Test *p-values* for both models is <2e-16 which are below the 0.05 leading to the rejection of the null hypothesis that the data mining algorthms are statistically significant.

• *There is no measurable difference between the predicted factors of rats that affect TB detection performance*

The classification technique was used to build predictive models that predicted the class for each rat and the factors that influence TB detection performance. However, this technique applied three algorithms to learn the relationship between variables. The independent variables (input) include Age, Av_Weight_Per_Year, Session_Start_Time, and Session_Completion_Time while the dependent variable (output) is Performance. Therefore, all three algorithms applied these variables separately to build predictive models on factors that influence rats' TB detection performance. Starting with the decision tree algorithm, it generated a decision tree where the top node (root node) shows the most significant factor that influences TB detection performance. The ability of the algorithm to seek optimal splits in variable values has led to this. Moreover, the leaf nodes indicate the class of every instance from the observations.

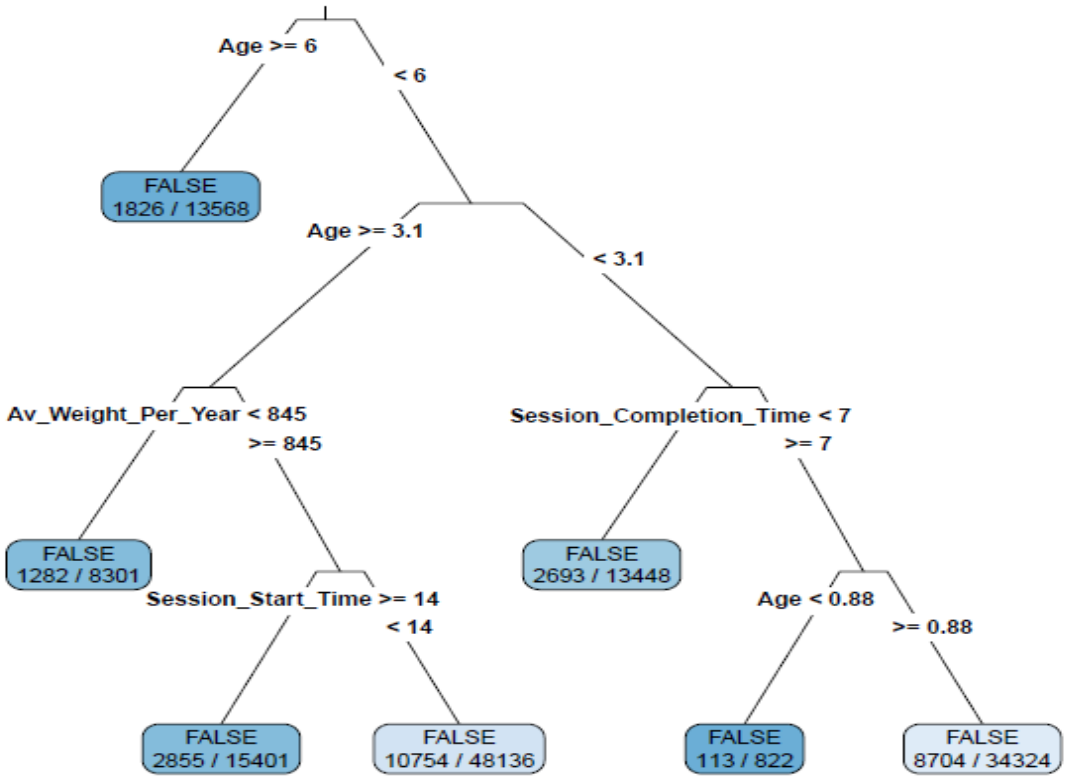

Figure 1: Decision tree with rats factors that influence TB detection performance

Figure 1 depicts the hierarchy of variables where the variable with a high correlation (Age) with the prediction, split on first. Thus, the age of the rat is the most significant factor. However, the other predicted factors are shown on the leaf nodes which indicate the class of every instance. Moreover, the decision tree algorithm-generated rules which are easy to interpret and understand. These rules are the result of the IF-THEN statements from the root node to the leaf nodes as reported in Table 9.

Table 9: Classification rules generated from decision tree algorithm

Rule number	Rule	Performance Decision		
		TRUE	FALSE	Number of Observations in %
1	IF Age >= 6 ⇒	0.13	0.87	10%
2	IF Age < 0.88 & Session_Completion_Time>= 7 ⇒	0.14	0.86	1%
3	IF Age is 3.1 to 6	0.15	0.85	6%

	& Av_Weight_Per_Year< 845 ⇒			
4	IF Age is 3.1 to 6 & Av_Weight_Per_Year >= 845 & Session_Start_Time >= 14 ⇒	0.19	0.81	11%
5	IF Age < 3.1 & Session_Completion_Time< 7 ⇒	0.20	0.80	10%
6	IF Age is 3.1 to 6 & Av_Weight_Per_Year >= 845 & Session_Start_Time < 14 ⇒	0.22	0.78	36%
7	IF Age is 0.88 to 3.1 & Session_Completion_Time >= 7 ⇒	0.25	0.75	26%

From Table 9, the first rule implies that older rats (with ages greater or equal to 6 years) had a performance chance of 0.13, TRUE and 0.87, FALSE, and detected fewer observations (10%). Considering the second rule, rats with the age of fewer than 0.88 years and at least 7 minutes (session completion time) had a performance chance of 0.14 TRUE, 0.86 FALSE, and detected 1% of the observations. Hence, older and less young rats portrayed low performance. The sixth rule has 36% of the detected observations. Rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, and the session start time before 14:00 hours had a detection performance chance of 0.22 TRUE, and 0.78, FALSE. This rule is consistent with the fourth one except for the session start time. Since the sixth rule had many observations than the fourth, the session starts time before 14:00 hours are the most performed one. Furthermore, the fifth rule has 10% detected observations, which imply rats with ages of 3.1 years and session completion time of fewer than 7 minutes had a performance chance of 0.20 TRUE and 0.80 FALSE. When comparing this rule with the second one, rats with a session completion time of fewer than 7 minutes depicted potentiality in detection since this rule had many observations compared to the second one. Therefore, the results pinpointed in Table 11 manifest that rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, the session start time before 14:00 hours, and fewer than 7 minutes as the session completion time performed well.

However, it is of importance to understand the extent to which each factor contributed to the prediction. The random forest algorithm pinpointed the predictor variables that are important in predicting the outcome based on the mean decrease in Gini (impurity), as shown in Figure 2. Mean Decrease in Gini is the average (mean) of a variable total decrease in the likelihood of incorrect classification of a new instance of a random variable from the data set.

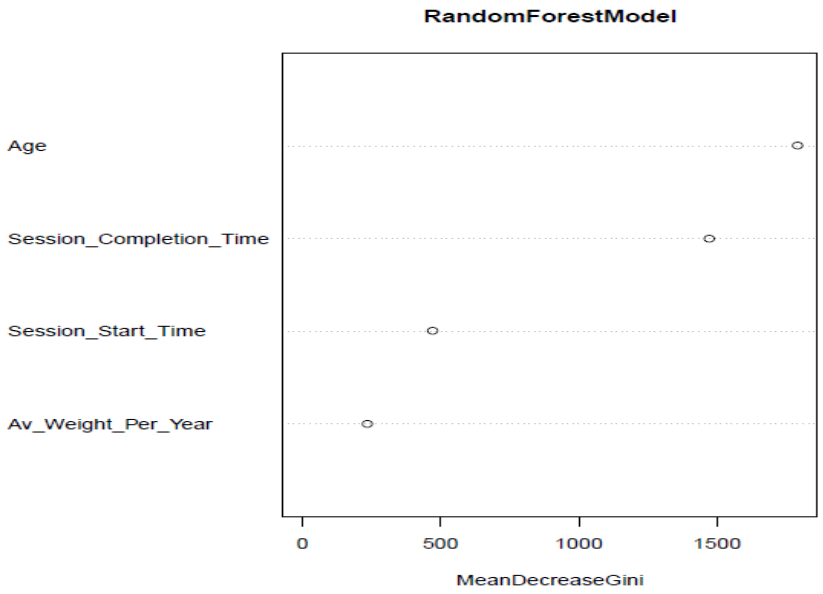

Figure 2: Variable importance generated by Random Forest algorithm

From Figure 2, a higher (1791.9167) and lower (233.5753) mean Decrease in Gini portrays greater and less variable importance, respectively. Hence, Age and Av_Weight_Per_Year are the most and least significant factors. Both decision tree and random forest have indeed shown Age as the most significant factor. Thus, the random forest algorithm and decision tree algorithm have predicted the factors that influence rats' TB detection performance by using the classification technique. However, the naive Bayes algorithm was used to create the model and compare their classification accuracy since it measures the probabilities of the variables and not their weights. Therefore, regarding the second hypothesis, as the *p-value*s shown in Table 10 are less than 0.05, we reject the null hypothesis and conclude that there is a measurable difference between the predicted factors of rats that affect TB detection performance. Hence, the predicted factors are statistically significant.

Table 10: Predicted factors with their corresponding p_values

Factor	*p_value*
Age	< 2e-16
Session_Completion_Time	< 2e-16
Session_Start_Time	2.53e-09
Av_Weight_Per_Year	2.98e-07

• *There is no measurable difference between the ability of rats in TB detection performance*

From the given data and the aim of the study, the rats' performance implies their ability to detect a sample with either TB, TRUE (Sensitivity) or without TB, FALSE (Specificity). Table 3 manifests that the youngest and oldest rats had the ages of 0.79 and 7.95 years respectively with the median age of 3.71 years. Meanwhile, the less weighty and weighty rats had the weights of 843.7g and 1054.8g respectively with the median of 866.8g. And as such, rats with ages and

weights below and above the median refer to younger, older, less weighty, and weighty rats respectively. Additionally, early detection conducted at 8:00 hours while the late detection was done at 18:00 hours with the minimum completion time of 1 minute and maximum completion time of 129 minutes respectively with the median completion time and start time of 10 minutes and 12:00 hours. Since the given data had many numbers of observations with FALSE values than TRUE values as shown in Table 5, the rats' high performance in these data had a FALSE value. In this conception, rats' performance depended on the number of observations accomplished. Therefore, the results pinpointed in Table 3 manifest that rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, the session start time before 14:00 hours, and fewer than 7 minutes as the session completion time performed well. With regards to these results, it is obvious that there is difference in ability of rats to detect TB samples. This is also evidenced by Table 10 which indicated the *p_values* of each predicted factor which are less than 0.05, and thus makes the results significant and reject the null hypothesis.

4 Discussion

4.1 Characteristics of Data

Considering data understanding phase, the given data consisted of many variables and observations, but the sample size for characterizing a TB rat is therefore only four female rats. The given number of rats was the ones found with the requested data and was expected to address the aim of the study of finding the influencing factors based on the number of observations as shown in Table 4 and not comparing the performance of every rat which would require large sample sizes. Moreover, Table 4 reported that, there was no gender equality in the given data since all rats were female. However, for the future it is advantageous to analyze data with large sample size and both male and female rats to understand which gender influences detection performance.

Based on dependent variable performance, Table 5 demonstrated that data consisted of many FALSE values than TRUE values. Since it was the target class for classification, it is presumed to have an impact on the results. Thus, when one value has many samples than the other, its performance is also higher. It was valuable if the data would have an estimation of about an equal number of values of the observations in the detection performance class. And as such, it would reduce the suspicion that the results might rely on one group of the data and limits generalization. Furthermore, Table 6 shows the logistic regression analysis which examined the association of independent variables with a dichotomous dependent variable (performance). The *p-values* for the independent variables are below 0.05, and implies that there is a relationship between independent variables and the dichotomous dependent variable. Hence, the variables are statistically significant.

4.2 Factors Influencing Rats TB Detection Performance

The results depict the strength of the Age factor in the detection performance. Figure 1 shows that Age split first due to the highest information gain ratio. As a result, it has appeared in all generated rules in Table 9. Contrary to the other variables that are shown only once in the generated rules. Moreover, in the variable importance of random forest depicted in Figure 2, the decrease mean Gini of Age was higher than the other variables. The results manifested that rats

between the ages of 3.1 to 6 years positively affected the performance. However, it may limit the generalization of the results since it referred to female rats. The study of Brushfield et al. [8] proposes that detection performance may depend on rats' characteristics such as age. Nevertheless, successful training and growth progress might have led to good detection performance. Furthermore, the results show that older rats portrayed a low detection performance. And as such, the olfactory deficit might have caused this since detection performance depends on the rats' olfactory sensitivity [13].Moreover, the results provide new insight into the relationship between time differences when the rat starts and ends detection tasks (session completion time). And as such, good performers were the rats that completed the number of observations with less than 7 minutes. Since these rats have a high-speed of detecting 100 samples in 20 minutes, good performers were the rats that completed the number of observations with less than 7 minutes. However, the given data might have influenced the session's completion time since the samples contained many values of FALSE (samples without TB bacteria). The study conducted by Mgode et al. [7] pinpoints that during the training, rats learn to pause for a long time of about 3 seconds to the sample hole with TB bacteria and take a short time of about 1 second to the sample hole without TB bacteria.

Not only that but also, the results contribute a clearer understanding of the influence of average weight per year (Av_Weight_Per_Year) on detection performance. Rats with an average weight per year of greater or equal to 845g performed better. According to the study conducted by Beyene et al. [5], the weight range of adult rats' females ranges from 1 to 1.5kg. Therefore, one can argue that young rats with at most 1.05 kg were the most performers. However, presumed the reliability of the results could increase with the corresponding weight rather than the applied average. However, these results may limit generalization since they refer to female rats. Therefore, rats TB trainers and decision-makers must consider these results to utilize the usefulness of this technology and should maintain it for sustainability. On the other hand, the results reveal that for the three different algorithms used, the classification accuracy was much more in the random forest (78.82%) than decision tree (78.78%) and naive Bayes (78.71%). Conversely, the predictive models' accuracies differences are small. The nature of data and algorithms used might have caused this in the sense that random forest and decision tree algorithms fit in skewed data different from naive Bayes which, do better in normally distributed data [9]. Moreover, in the random forest, the ability to assembly several trees and make the final decision from several trees might influence this highest classification accuracy [12].

Despite the found results based on the dependent and independent variables given from the data, other factors presumed the influence on these results. These factors may include training procedures, trainers or recorders (data recording), experimental setup, and laboratory technicians (quality control) [7, 9, 15]. The study conducted by Reither et al. [15] argue that since rats are trained based on the conditioning techniques which support to change their behavior such as learning to recognize sound during the training, it is useful to have the justifiable rules to avoid incorrect results. Likewise, Mgode et al. [7] demonstrate that rats' successful and consistent training procedures are most important in TB healthcare centers that apply rat as odor-detection technology. With this regards, it is presumed that rats from the given data succeeded in the training procedures and thus manifested better performance. Moreover, observing precision in data recording during detection tasks is highly emphasized to avoid false results. Since rats' trainers and recorders are the ones performing data recording and training, they should have skills in getting consistent records. Hence, a well-organized experiment setup may facilitate rats

to portray better performance [1]. Additionally, before presenting the sample in a cage for detection, a standard heat is applied into it to kill infectious microorganisms and enhance quality control. Hence, quality control may determine the effectiveness of rats in detection performance. Consequently, despite the outcome of the dependent and independent variables, the mentioned confounding variables might influence the results. Therefore, rats' detection performance depends on the main and confounding factors.

5 Conclusion and Recommendations

This study has focused on the prediction of factors influencing rats' TB detection performance using data mining techniques. Also, building predictive models for predicting the class for every new instance (rat). While this study also concentrated on understanding the relationship of the manipulated variables, the results indicate that Age, Session_Completion_Time, Session_Start_Time, and Av_Weight_Per_Year are the factors influencing rats TB detection performance. However, the results show that the age of the rat was the most influencing factor. The results also pinpoint that rats with the age of 3.1 to 6 years, at least 845g of the Av_Weight_Per_Year, before 14:00 hours as the session start time, and less than 7 minutes as the session completion were the best performers. These results are useful to rats' trainers and decision-makers in understanding the potential factors that may affect the detection performance and hence increase TB detection performance. Ultimately to support decision making, scaling up of the detection technology and improve human health.

Considering predictive models, the random forest predictive model has the highest classification performance accuracy of 78.82%. Followed by the decision tree with 78.78% and naive Bayes is the last model with 78.71% and thus makes the random forest predictive model the best model for the study. Since this study implemented data mining techniques in a social setting by predicting factors that influence rats in detecting TB disease, it is also helpful to the academic society of Information systems. However, confounding factors such as training procedures, trainers or recorders (data recording), experimental setup, and laboratory technicians (quality control) might have an impact on the results. Therefore, to maximize the effectiveness and efficiency of these results, several criteria for future research will have to be optimized. First, a dataset with large sample size and many desirable variables for rats TB detection performance is valuable to increase the number of known factors. Moreover, to predict significant sex differences, the dataset should balance gender distribution.

Acknowledgments

Data for this study was supported by APOPO TB Training and Research center in Morogoro, Tanzania. Many staff from the APOPO TB center have provided advice and appreciated suggestions. Colleague's critiques and comments have consistently improved the paper.

Conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

6 References

1. Poling A, Weetjens B, Cox C, Beyene N, Durgin A, et al. 2011. Tuberculosis Detection by Giant African Pouched Rats. *Behav Anal*. 34(1), 47-54. PubMed https://doi.org/10.1007/BF03392234

2. World Health Organization. Global tuberculosis report 2018. New York, United States of America: WHO; 2018.

3. Ellis H, Mulder C, Valverde E, Poling A, Edward T. 2017. Reproducibility of African giant pouched rats detecting Mycobacterium tuberculosis. *BMC Infect Dis*. 17, 298. doi:. PubMed https://doi.org/10.1186/s12879-017-2347-3

4. Weetjens BJ, Mgode GF, Machang'u RS, Kazwala R, Mfinanga G, et al. 2009. African pouched rats for the detection of pulmonary tuberculosis in sputum samples. *Int J Tuberc Lung Dis*. 13, 737-43. PubMed

5. Beyene, N., Mahoney, A., Coxi, C., Weetjens, B., Makingi, G., Mgode, G, et al. (2012). APOPO's tuberculosis research agenda: achievements, challenges and prospects. *Tanzania Journal of Health Research.* doi: 10.4314/thrb.v14i2.5

6. Poling A, Valverde E, Beyene N, Mulder C, Cox C, et al. 2017. Active Tuberculosis detection by pouched rats in 2014: More than 2,000 new patients found in two countries. *J Appl Behav Anal*. PubMed https://doi.org/10.1002/jaba.356

7. Mgode GF, Cox CL, Mwimanzi S, Mulder C. 2018. Pediatric tuberculosis detection using trained African giant pouched rats. *Pediatr Res*. 84(1). doi:. PubMed https://doi.org/10.1038/pr.2018.40

8. Brushfield A, Luu T, Callahan B, Gilbert P. 2008. A comparison of discrimination and reversal learning for olfactory and visual stimuli in aged rats. *Behav Neurosci*. 122(1), 54-62. PubMed https://doi.org/10.1037/0735-7044.122.1.54

9. Sharda, Delen & Turban (2014). Business Intelligence and Analytics (Tenth edition).

10. Nagabhushanam D, Naresh N, Raghunath A, Praveen Kumar K. 2013. Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient's Data. Int J Cloth Sci T*echnol*. 4, 262-65.

11. Suresh, N. & Arulanandam, D. (2018). A Mining Approach for Detection and Classification Techniques of Tuberculosis Diseases.

12. Asha, T., Natarajan,S., Murthy, K.N.B., (2011). Effective Classification Algorithms to Predict the Accuracy of Tuberculosis-A Machine Learning Approach.

13. Czarnowski I., Jędrzejowicz P. 2018. An Approach to Data Reduction for Learning from Big Datasets: Integrating Stacking, Rotation, and Agent Population Learning Techniques. *Complexity*. doi:. https://doi.org/10.1155/2018/7404627

14.Hussain, S. (2015). Educational Data Mining using R Programming and R Studio. Journal of applied and fundamental sciences

15.Reither, K., Jugheli, L., Glass, T.R., Sasamalo, M., Mhimbira, F.A., Weetjens, B.J, et al. (2015). Evaluation of Giant African Pouched Rats for Detection of Pulmonary Tuberculosis in Patients from a High-Endemic Setting.

16.Chaurasia V, Pal S. (2013). Data Mining Approach to Detect Heart Disease. International Journal of Advanced Computer Science and Information Technology Volume 2, Issue 4, ISSN: 2296-1739.

17.Ameri H, Alizadeh S, Hadizadeh M. 2014. Assessing the Effects of Infertility Treatment Drugs Using Clustering Algorithms and Data Mining Techniques [Persian]. *J Mazandaran Univ Med Sci*. 24, 26-35.

18.Ayas, S. &Ekinci, M. (2014). Random forest-based tuberculosis bacteria classifications in images of ZN-stained sputum smear samples. doi: .

19.Kraemer S, Apfelbach R. 2014. Olfactory sensitivity, learning and cognition in young adult and aged male Wistar rats. *Physiol Behav*. PubMed

20.Mahoney A, Edwards TL, Weetjens BJ, Cox C, Beyene N, et al. 2013. Giant African pouched rats (CricetomysGambianus) as detectors of Tuberculosis in human sputum: Two operational improvements. *Psychol Rec*. 63, 583-94. https://doi.org/10.11133/j.tpr.2013.63.3.012

21.Maniya H, Hasan MI, Patel PK. 2011. Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis [IJCA]. *Int J Comput Appl*.

22.Mulder C, Mgode GF, Reid SE. 2017. Tuberculosis diagnostic technology: an African solution … think rats. *Afr J Lab Med*. 6(2), https://doi.org/10.4102/ajlm.v6i2.420

23.PrasannaDesikan. Kuo-Wei Hsu, Srivastava,J. (2011). Data Mining for Healthcare Management. SIAM International Conference on Data Mining.

24.World Health Organization. Make every mother and child count. Geneva, Switzerland: WHO; 2005.