Journal of Intelligent Learning Systems and Applications, **2013, 5, 184-193**
http://dx.doi.org/10.4236/jilsa.2013.53020 Published Online August 2013 (http://www.scirp.org/journal/jilsa)

Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course

Sharareh R. Niakan Kalhori[1,2*], Xiao-Jun Zeng[3]

[1]Department of Public Health, School of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [2]Social Determinants of Health Research Center, School of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [3]Department of Machine Learning & Optimization, School of Computer Science, University of Manchester, Manchester, England.
Email: *niakan.sh@ajums.ac.ir

Received December 24th, 2012; revised January 24th, 2013; accepted February 4th, 2013

ABSTRACT

Tuberculosis treatment course completion is crucial to protect patients against prolonged infectiousness, relapse, lengthened and more expensive therapy due to multidrug resistance TB. Up to 50% of all patients do not complete treatment course. To solve this problem, TB treatment with patient supervision and support as an element of the "global plan to stop TB" was considered by the World Health Organization. The plan may require a model to predict the outcome of DOTS therapy; then, this tool may be used to determine how intensive the level of providing services and supports should be. This work applied and compared machine learning techniques initially to predict the outcome of TB therapy. After feature analysis, models by six algorithms including decision tree (DT), artificial neural network (ANN), logistic regression (LR), radial basis function (RBF), Bayesian networks (BN), and support vector machine (SVM) developed and validated. Data of training (N = 4515) and testing (N = 1935) sets were applied and models evaluated by prediction accuracy, F-measure and recall. Seventeen significantly correlated features were identified ($P <= 0.004$; 95% $CI = 0.001 - 0.007$); DT (C 4.5) was found to be the best algorithm with %74.21 prediction accuracy in comparing with ANN, BN, LR, RBF, and SVM with 62.06%, 57.88%, 57.31%, 53.74%, and 51.36% respectively. Data and distribution may create the opportunity for DT out performance. The predicted class for each TB case might be useful for improving the quality of care through making patients' supervision and support more case—sensitive in order to enhance the quality of DOTS therapy.

Keywords: Tuberculosis; Machine Learning; Prediction; Classification; DOTS

1. Introduction

Over recent years, Tuberculosis has been considered, particularly in low and middle-income countries as a global public health concern with the estimated two million deaths annually [1,2]. Up to 50% of all patients with TB do not complete treatment and fail to adhere to their therapy [3]. It has been estimated that in industrialized countries non-completion of treatment is around 20% [4] and according to the Centre of Disease Control and Prevention in the United States, 25% of patients fail to complete their chemotherapy [5]; in other words, the proportion of patients with active disease who completes therapy under standard conditions ranges from as little as 20% - 40% in developing countries to 70% - 75% in the USA [6]. Noncompliance is a significant factor leading to the persistence of tuberculosis in many countries and the consequences of this well recognized fact are prolonged infectiousness, relapse, prolonged and more expensive therapy due to multidrug resistance TB and death [7]. It has been revealed that noncompliance is associated with a 10-fold increase in the incidents of poor results from treatment and accounted for most treatment failures [8]. It is the most serious problem hampering tuberculosis treatment and control as patients with active disease who are non-compliant, sputum conversion to smear-negative will be delayed and relapse rates will be 5 - 6 times higher and drug resistance may develop [6]. That is, TB patients remained in the pool of active cases will in-

*Corresponding author.

crease the development of TB among latent cases who are prone to be infected or affected. This threats public health and makes huge costs that can be spent to improve public health improvement and promotion instead.

DOTS (directly observed treatment, short course) which is current international control strategy for TB control involves the case detection and completes the entire course of treatment successfully. In 2006, in order to improve DOTS quality, World Health Organization (WHO) has designed "Stop TB" Plan [2]. In this plan, it has been highlighted that health care services should identify and concentrate on interruption factors that halt TB treatment. Supervision which plays important role in patient treatment adherence and drug resistance prevention must be carried out in a context-specific and patient-sensitive manner.

Although WHO has highlighted the necessity of improving the quality of DOTS in terms of supervision and patient support in the "Stop TB" plan, there is no specific way to measure how intensive health providers' support and supervision should be and to whom of TB cases it should be provided. To make this clear, we may require a tool to predict the patient destination regarding TB treatment course completion. The tool may identify high risk TB patients for treatment course non-compliance as it is impossible to serve all TB patients with active supervision and support due to the cost consideration and limited resources. This may be used to define the level of supervision and support each patient needs based on the predicted outcome by an accurate predictive model. Currently, no system is available to estimate TB treatment course through using features of TB patients and designing a systematic method to predict the given outcome.

For the prediction of tuberculosis treatment course completion, the defined outcome related to each record of TB patient contains five potential classes: cure and completed treatment (desirable outcomes), failure, quit, and death (undesirable outcomes) [9]. Patients who are classified with undesired outcomes need more supervision and support. Public health interventions for TB control need to be set according to how many of TB patients would shift to unwelcome outcomes.

Making sure that TB patients finish the treatment course entirely is a main step to TB control and public health promotion. Actually, here the main question of concern is "can new TB cases at risk of failing in treatment course completion identified from their early registration"? For this purpose, machine learning methods have been already applied and they worked properly according to previous studies [10,11]. They infer a model through being trained by a set of historical examples in frame of data categories; thus, new examples with unknown classes could be assigned to one or more classes by pattern matching within the developed valid model [12].

This work used machine learning algorithms to achieve six predictive models of a crucial real-world problem which offer the medical research community a valid alternative to manual estimation of risky TB patients to quit or fail treatment course completion. This recognizes TB patients whom supervision in frame of DOTS therapy needs to be more active as they are risky for treatment course interruption.

2. Methods and Algorithms

2.1. Correlation Analysis

In the case of a large dataset, learning the dataset is not useful unless the unwanted features are removed since an irrelevant and redundant feature does not add anything positive and new to the target concept [13]. To identifying influential predictors, a bivariate correlation is used which is a correlation between two variables including independent and dependent parameters by calculating the correlation coefficient "*r*" [14]. As there is no specific prediction and hypothesis two-tailed test is carried out.

Suppose we have got two variables *X* (list of features) and *Y* (class), an optimal subset is always relative to a certain evaluation function. To discover the degree to which the variables are related, correlation criteria are applied [15]. It reflects the degree of linear relationship between two variables ranging from +1 to −1. A perfect positive linear relationship between variables is shown by +1; however, −1 implies an entire negative linear association. The following formula (1) is used to calculate the value of *r*:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left(n-1\right)S_x S_Y} \tag{1}$$

where there are two variables *X* and *Y* and their means *X*, *Y* and standard deviations including S_X and S_Y respectively; *n* is the number of TB instances. The correlation coefficient can be tested for statistical significance using special t-test through following formula.

$$t = r\sqrt{\left(n-2\right)\Big/\left(1-r^2\right)} \tag{2}$$

Degree of freedom for correlation coefficient calculation is equal to $n - 2$. From a t-table, we would find significant relationship between each of variables *X* and *Y* ($P < 0.05$).

2.2. Problem Modeling

To achieve the aim of this work, we evaluate a number of well-known classification algorithms on TB treatment course completion. In fact, prediction can be viewed as

the construction and use of a model to assess the class of an unlabeled sample, or to evaluate the value or value ranges of an attribute that a given sample is likely to have [16]. To train every of applied algorithms, we set a big matrices as follows:

$$X_i^j = \begin{bmatrix} x_1^1 & \cdots & x_1^{17} \\ : & : & : \\ x_j^1 & \cdots & x_j^i \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ : \\ y_j \end{bmatrix} \tag{3}$$

where X is a big matrice which is used to train the given algorithm by using train set in which i would be the *ith* number of samples and j is the *jth* number of predictor. x_j^i in our training set would be x_{4515}^{17} and in applied testing set has been x_{1935}^{17}. Y which is the dependent variable of TB treatment course in both training and testing sets addresses five classes where a label of "1" implies that TB case got cured and "2" means that s/he has completed the treatment course entirely; whilst a label of "3", "4", and "5" means that the patient belongs to the undesirable class such as quitting, failing the treatment course, or dead respectively.

Having compared the pros and cons of machine learning methods, to carry out the considered prediction task in this study six classifiers including decision tree (DT), Bayesian network (BN), logistic regression (LR), multilayer perceptron (MLP), radial basis function (RBF) and support vector machine (SVM) were selected. Next section is brief explanation about each of these selected classifiers. As shown in **Figure 1**, the dataset was split into training (two-third) and testing (the other one-third) datasets each containing seventeen significantly correlated attributes and the outcome variables for every record without any missing data. The six above named classifiers were applied to the training dataset to estimate the relationship among the attributes and to build predictive models. Afterwards, the testing dataset which was not used to model inference was utilized to calculate the predicted classes and compare the predicted values with the realones available in the testing dataset. The model which is the most fitted and accurate one will be selected to predict the outcome of Tb treatment course for new TB cases.

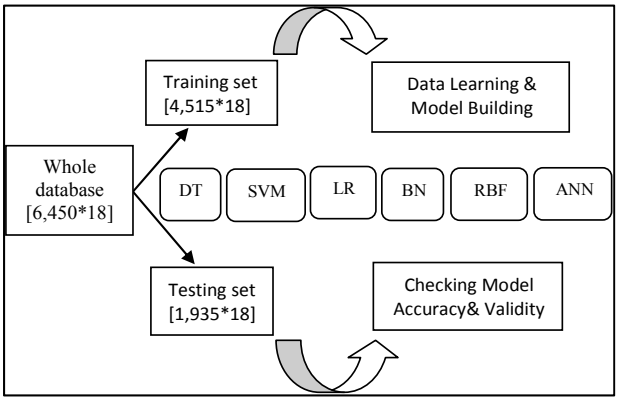

Figure 1. Schematic presentation of applied methodology of model development and validation process. In order to develop accurate predictive or classification models, firstly data of TB patients (in a huge matice with 6450 rows and 18 vectors) are divided as training and testing sets. Training set is used to learn data and build models through applying 6 algorithms including decision tree (DT), support vector machine (SVM), logistic regression (LR), Bayesian network (BN), radial basis function (RBF), and artificial neural network (ANN). To check the validity and accuracy of developed models, testing set is applied.

2.3. Modeling Algorithms

2.3.1. Decision Tree

The decision tree is a nonparametric estimation algorithm that input space is divided into local regions defined by a distance measure like the Euclidean norm; it is a flowchart-like tree structure where the internal node, branches, and leaf node means concepts associated with our training tuples. In this hierarchical data structure, the local region is identified in a sequence of recursive splits that in a smaller number of steps by implementing divide-and-conquer strategy. This powerful classifier is famous for its intuitive explainability [17]. In this research, C4.5 classification task-oriented algorithm has been applied through calculating:

$$I\left(s_1, s_2, \cdots, s_m\right) = -\sum_{i=1}^{m} p_i \log_2\left(p_i\right) \tag{4}$$

$$p_i = s_i / s \tag{5}$$

where p_i is the probability that an instance belongs to class C_i. Having calculated the entropy $E\left(A_i\right)$, which addresses the expected information according to partitioning by attribute A_i we have:

$$E\left(A_i\right) = \sum_{j=1}^{q} \frac{s_{ij} + s_{ij} + \cdots + s_{mj}}{s} I\left(s_{ij}, \cdots, s_{mj}\right) \tag{6}$$

$$I\left(s_{1j}, s_{2j}, \cdots, s_{mj}\right) = -\sum_{i=1}^{m} p_{ij} \log_2\left(p_{ij}\right) \tag{7}$$

where $p_{ij} = \frac{s_{ij}}{\left|s_j\right|}$ and $\left|s_j\right|$ is the number of samples in subset s_j. In the next step, the encoding information that would be gained by branching on A_i is:

$$\text{Gain}\left(A_i\right) = I\left(s_1, s_2, \cdots, s_m\right) - E\left(A_i\right) \tag{8}$$

The attribute A_i with the highest information gain is chosen as the root node, the branches of the root node is formed according to various distinctive values of a_{ij}. The tree grows until if all the samples are all of the same class, and then the node becomes a leaf and is labeled

with that class. From the tree, understandable rules can be extracted in a quick processing [17]. Here, the task of decision tree development is conducted using C4.5 classification algorithm where each tree leaf is allocated to a class of chemotherapy outcome along with number of misclassified cases.

2.3.2. Bayesian Networks

Generally, Bayesian classifiers are statistical approaches capable of predicting class membership likelihoods like the probability of the training set belonging to a specific class. It is based on Bayes theorem and well known for its high accuracy and speed when applied to a large data collection. Here, simple estimator Bayesian network has been used.

Let X be a sample whose class is to be defined and let H be a hypothesis that X belongs to class C. In this classification task, $P(H|X)$ needs to be defined which is the probability that the hypothesis H holds based on the data sample X. It can be calculated through:

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)} \tag{9}$$

where $P(H)$ and $P(X)$ are the prior probability of H and X respectively. $P(X|H)$ is the posterior probability of X given on the H. In model training these three values are calculated and the probability for sample X to be in hypothesis H can be determined. There are two essential features that define a Bayesian network: a directed cyclic graph in which each node denotes a random variable, either discrete or continuous values, as well as a set of conditional probability tables. Let $x=(x_1,\cdots,x_n)$ be a data tuple related to the correspondent attributes $y_1,y_2,\cdots,y_n$. Given that every attribute is conditionally independent on its non-descendent and parents in the network, the following equation is a representation of joint probability distribution.

$$P(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}P\left(x_i\left|\text{parents}(y_i)\right.\right) \tag{10}$$

where the values for $P\left(x_i\left|\text{parents}(y_i)\right.\right)$ are related to the entries for y; $P(x_1,x_2,\cdots,x_n)$ is the probability of a specific combination of values of X. Probability distribution, with the probability of each class may be the output of this classification process [18].

2.3.3. Logistic Regression

Logistic regression is used primarily for predicting binary or multi-class dependent variables. This algorithm's response variable is discrete and it builds the model to predict the odds of its occurrence. This method's restrictive assumptions on normality and independence induced an increased application and popularity of machine learning techniques for real-world prediction problems. In this study, multinomial logistic regression is applied which is an algorithm that constructs a separating hyperplane between two sets of data by using the logistic function to express distance from the hyperplane as a probability of dichotomous class membership:

$$P(Y)=\frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_nx_n+\varepsilon)}} \tag{11}$$

In this equation, X_i symbolizes discrete or continuous predictor variables with numeric values; in the case of depending variable (Y) being dichotomous, we use this algorithm. The constants $\beta_0,\beta_1,\beta_2,\cdots,\beta_n$ are the regression coefficients estimated from training data which are typically computed by using an iterative maximum likelihood technique. Normally, this formula's justification is that the log of the odds, a number that goes from $-\infty$ to $+\infty$, is a linear function. Particularly by using this model, stepwise selection of the variables can be made and the related coefficients calculated. In producing the LR equation, the statistical significance of the variables used to be determined by the maximum-likelihood ratio [19].

2.3.4. Artificial Neural Networks

Artificial Neural Network (ANN) is biologically inspired analytical method which is able of modeling extremely complex non-linear functions. Here, a common architecture named multi-layer perceptron (MLP) with learning by back-propagation algorithm is built. A neural network is a compound of linked input/output units in which every link has an associated weight. Adjusting the weights is the core phase for predicting the correct class label of input through iterative learning. This method is popularly used in classification and prediction tasks with high tolerance to noise and the ability to classify unseen patterns [19]. Here, algorithm has conducted the learning process 500 times and the network structure is shown in a simple way when only two attributes have been applied as inputs. The structure of a two-input, one hidden layer neural network for our task with five outputs has been presented in **Figure 2**.

2.3.5. Radial Basis Function

A radial basis function network (RBFN) is an artificial intelligence network in which its activation function is simply radial basis in a linear combination. This type of network was designed to view a problem in curve-fitting (approximation) and high dimensional space. The real inspiration behind the RBF technique is finding amultidimensional function that offers the best fit to training tuple and then applies this multidimensional surface to interpolate the test data through regularization [20]. Gaus-

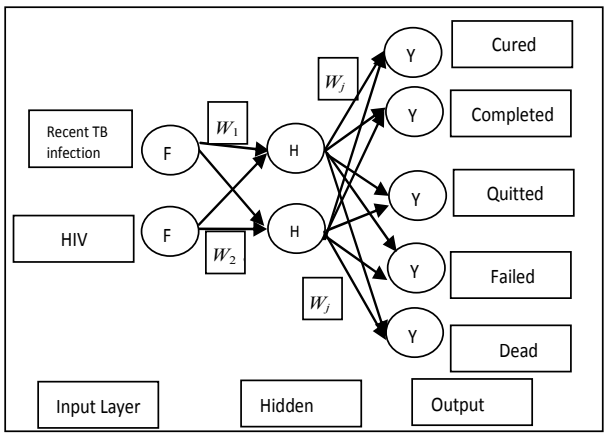

Figure 2. The simplified structure of applied neural network only with two of our attributes "HIV" and "recent TB infection", one hidden layer and five classes in output layer. The structure is much simplified neural networks are popularly applied in classification and prediction, as they have advantages such as high tolerance to noise, and the ability to classify unseen patterns.

sian radial basis function is applied algorithm in this study.

2.3.6. Support Vector Machine

Support vector machine (SVM) is a new classification method for both linear and non-linear data. SVM applies nonlinear mapping to transform the original training tuple into a higher dimension. It seeks the optimal linear separation hyperplane which is the decision boundaries separating the tuple based on their class labels.Polykernel support vector machine has been applied in this investigation.Let the dataset *A*be considered as (x_1, y_1), $\left((x_2 x_2, y_2), \cdots (x|A|), y|D|\right)$ where (x_i) is the set of learning data with correspondent class label y_i. For a two-class related training dataset, for instance, every y_i can take either +1 or −1. This could also be generalized to *n* dimensions and the SVM duty is to find the best dividing lines that can be drawn and divide all of the tuples of every class from the others. For multidimensional classes the hyperplanes should be found asdecision boundaries. This can be arranged by defining the maximum marginal hyperplane (MMH) since models with larger hyperplane are more accurate at classification [18].

2.4. Accuracy Measurements

In order to evaluate the prediction rate, the following related parameters need to be measured. Prediction accuracy percentage (model Accuracy), model fitness, recall, Precision, F-measure, and ROC area are considered criteria used to assess the models' validity. In confusion matrix (**Figure 3**) in which the columns denote the actual cases and the rows denote the predicted cases, the accuracy of model obtained from training set (model fitness) and testing set (prediction accuracy) are calculated. The prediction accuracy is the percentage of correct prediction (true positive + true negative) divided by the total number of predictions. In machine learning, sensitivity is simply termed recall (*r*) and precision is the positive predicted value. F-measure is a harmonic means of precision and recall and the higher value of it merely reveals the better performance of a prediction task.

Another comparative criterion is Roc curve which is a two-dimensional graph where the true positive (*TP*) rate on the *Y* axis is plotted on the false positive (*FP*) rate on the *X* axis [21].

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{13}$$

$$\text{Prediction Accuracy}\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$\text{Precision}(\%) = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall}\ (\%) = \frac{TP}{TP + FN} \tag{16}$$

2.5. The Database

The available dataset has been driven from gathered records by health practitioners, nurses, and physicians at local TB control centres throughout Iran in 2005. In tuberculosis control centres, health deputies of each province in a network system collected data in every appointment. By using "Stop TB" software, more than 35 parameters were collected and through applying bivariate

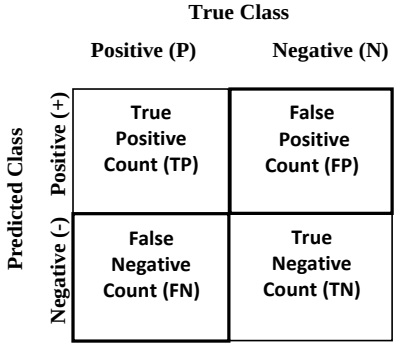

Figure 3. A simple confusion matrix, a two-by-two table, values in the main diagonal (TP and TN) represent instances correctly classified, while the other cells represent instances incorrectly classified (FP and FN errors).

correlation method; those independent variables which are significantly correlated with target outcome are selected as predictors ($P <= 0.05$). They are presented and defined in **Table 1**. The distribution of corresponding outcomes of TB treatment course based on their frequency in training and testing sets are shown in **Table 2**. Due to normal and non-normal distribution variables which were tested by both Kolmogorov-Smirnov test and visual shape of attribute's distribution, Pearson and Spearman *rho* methods were applied to find highly correlated factors.

3. Experiments and Results

3.1. Feature Selection

Significant Kolmogorov-Smirnov test confirmed with the shape of their distribution showed the distribution of our variables; Spearman *rho* was applied for those with non-normal distribution.

Having considered the results, it is revealed that males are more likely to not get cure and complete the course of TB treatment ($r = -0.082$, $P < 0.0001$).

As patients get older there is more probability to get undesirable result from treatment course ($r = 0.158$, $P < 0.0001$).

The more under-weight the patients are, they are more likely at the risk of unwelcome outcome ($r = -0.056$, $P < 0.0001$).

TB cases with extra-pulmonary TB are more likely at the risk of outcomes like quitting, failing or death ($r = -0.066$, $P < 0.0001$).

There is 0.127 more probability that immigrants from Iraq, Pakistan, and Afghanistan to quit, get failed in treatment course ($r = 0.127$, $P < 0.0001$). TB patients who are living in abroad as well as mobile cases are −0.027% more possible to get undesirable outcome compared with

Table 1. Patients' attributions applied for experiments, their definitions and their range of values.

Variable	Variable Definition	Categories of Values
	Demographic Characteristics	
sex	Gender of TB case	Male (1)/Female (2)
Age	Age of TB patient	(Continuous var.) 0.05 - 99
Weight	Weight of TB patient	(Continuous var.) 4 - 110
Nationality	Nationality of TB patient	Iranian (1), Central Asians (2), Iraqi (3), Pakistani (4), Afghani (5)
Area	Area of residence	Abroad (1), Mobile (2), Rural (3), Urban (4)
Prison	Current stay in prison	No (1) /Yes (2)
	Clinical Features	
Case type	Type of TB that patient is belonged to	New (1), imported (2), cure after absence (3), returned (4)
Treat cat	Category of treatment that is conducting	A (1)/B (2)
TB type	The part of body that has been affected	Pulmonary (1)/extra-pulmonary (2)
RTBinf	Whether patient has recently TB affected	No (1)/yes (2)
Diabetes	Whether TB patient isaffected by diabete	No (1)/yes (2)
HIV	Whether TB patient has been known as HIV+	No (1)/suspected (2)/yes (3)
Length	Length (month) of being affected by TB	(Continuous var.) 0.03 - 90.77
LBW	Low Body Weight	No (1)/yes (2)
	Social Risk Factors	
Imprisonment	The history of living in prison	No (1)/suspected (2)/yes (3)
IV drug using	Whether patient is using the (IV) drugs	No (1)/suspected (2)/yes (3)
Risky sex	Whether patient has history of risky sex	No (1)/suspected (2)/yes (3)

Table 2. The real outcomes happened for Tuberculosis patients for initial, training and testing sets with their corresponding outcomes.

Data set	Cured	Completed	Quit	Failed	Dead	Total
Training set	1510	1274	790	462	479	4515
Testing set	572	841	179	207	136	1935
Initial dataset	2082	2115	969	669	615	6450

those who are living in urban and rural areas ($r = -0.027$, $P < 0.0001$).

Prison residency and treatment outcome completion are significantly correlated ($r = -0.026$, $P < 0.0001$).

The more time TB patients spend affected with this infectious disease, there is 0.073 times more chance of non-desirable outcomes of tuberculosis treatment course.

Diabetic or HIV^+ TB cases are positively prone to have worse result of treatment course (*rho diabetes* = 0.029, $P < 0.001$ & *rho HIV* = 0.045, $P < 0.05$).

Also, the probabilities of having imprisonment history in his/her life, consuming drugs through intravenous, or having unprotected sex increase the likelihood of unwanted outcome with $r = 0.157$, 0.0172, and 0.16 respectively ($P < 0.000$).

3.2. Classification Algorithms Analysis

In this work, six classifiers including DT, BN, LR, MLP, RBF and SVM were applied to the patient dataset. Model development is conducted in two main steps including model fitness and model accuracy. To calculate the model fitness criteria we used the data of training set; however, to compute the model accuracy measurements, data of testing set is applied which is merely much more valuable to judge about our models accuracy. Related results of these experiments are demonstrated in **Table 3**.

Model fitness assessment by evaluating training accuracy is 84.45% for C4.5 decision trees; it is considerably less for Bayesian net where the value is 58.56%. The values of model fitness for logistic regression (56.5%), MLP (64.93%), RBF (50.65%), and SVM (53.04%) have been close. Having compared the Roc curves reveal that the area under curve for C4.5 has the most value for model fitness and accuracy with 0.96 and 0.97 respectively. This measurement is less for Bayesian net (0.85), logistic regression (0.82), MLP (0.81), RBF (0.79), and SVM (0.76) in terms of model accuracy. C4.5 decision tree has been able to build a model with greatest accuracy since the model fitness and prediction accuracy are 84.45% and 74.21% respectively.

Prediction accuracy for Bayesian networks has been calculated by 62.06%. Model accuracies obtained from other classifiers are different as this value for LR, MLP, RBF, and SVM have been 57.88%, 57.31%, 53.74%, and 51.36% respectively.

Figure 4 is the comparative Roc curves based on the given outcome of tuberculosis treatment including cure, complete, quit, failed or dead. This figure shows six Roc curves for six developed models based on the given outcome. For the outcome "cure", C4.5 has outperformed than other classifiers with area under curve 0.958. Similarly, the most accurate result for outcomes 'completed' and "quit" obtained for C4.5 with 0.966 and 0.956 respectively. C4.5 has also performed the best for the outcome "failed" and "dead" by classifying 0.986 and 0.963 of cases correctly. Overall, these results of area under curve reveals better performance of C4.5 decision tree classification algorithm. **Table 3** and **Figure 4** present obtained results including model fitness and accuracy as well as produced area under ROC.

4. Discussion

Of the six investigated methods, decision tree has achieved the best performance while other classifiers have given relatively close results in lower ranks. According to previous studies [22,23], the technique with the best classification performance might behave differently from another one and there is no single best method for every circumstance. Decision trees that classify instances by sorting them based on feature values has outperformed other methods. C4.5 performs variously in different conditions. It has been reported that there is an association between the performance of applied tools and following issues including the type of problem we are analyzing, the type of input data (discrete or continuous), and finally emerging overlapping in outcome classes.

Due to the fact that our available dataset is a large volume of data with dimensional structure, normally it is expected that SVMs, neural networks, and decision tree outperform others. However, the dataset is mainly com-

Table 3. Comparison on model fitness and model accuracy of six various applied machine learning algorithms.

	Model Fitness (through using training set)					Model Accuracy (through using testing set)				
Classifiers	PA*	Recall	Precision	F-measure	ROC Area	PA**	Recall	Precision	F-measure	ROC Area
C4.5	84.4	0.845	0.845	0.843	0.96	74.21	0.742	0.753	0.746	0.97
BN	58.5	0.586	0.591	0.579	0.83	61.70	0.621	0.659	0.621	0.85
LR	56.5	0.566	0.574	0.553	0.81	57.82	0.579	0.628	0.578	0.82
MLP	64.9	0.649	0.68	0.644	0.86	57.82	0.573	0.677	0.57	0.81
RBF	50.6	0.507	0.503	0.491	0.77	53.74	0.537	0.554	0.536	0.79
SVM	53.0	0.53	0.555	0.503	0.76	57.47	0.514	0.621	0.50	0.76

*training accuracy (%); **prediction accuracy percentage (%).

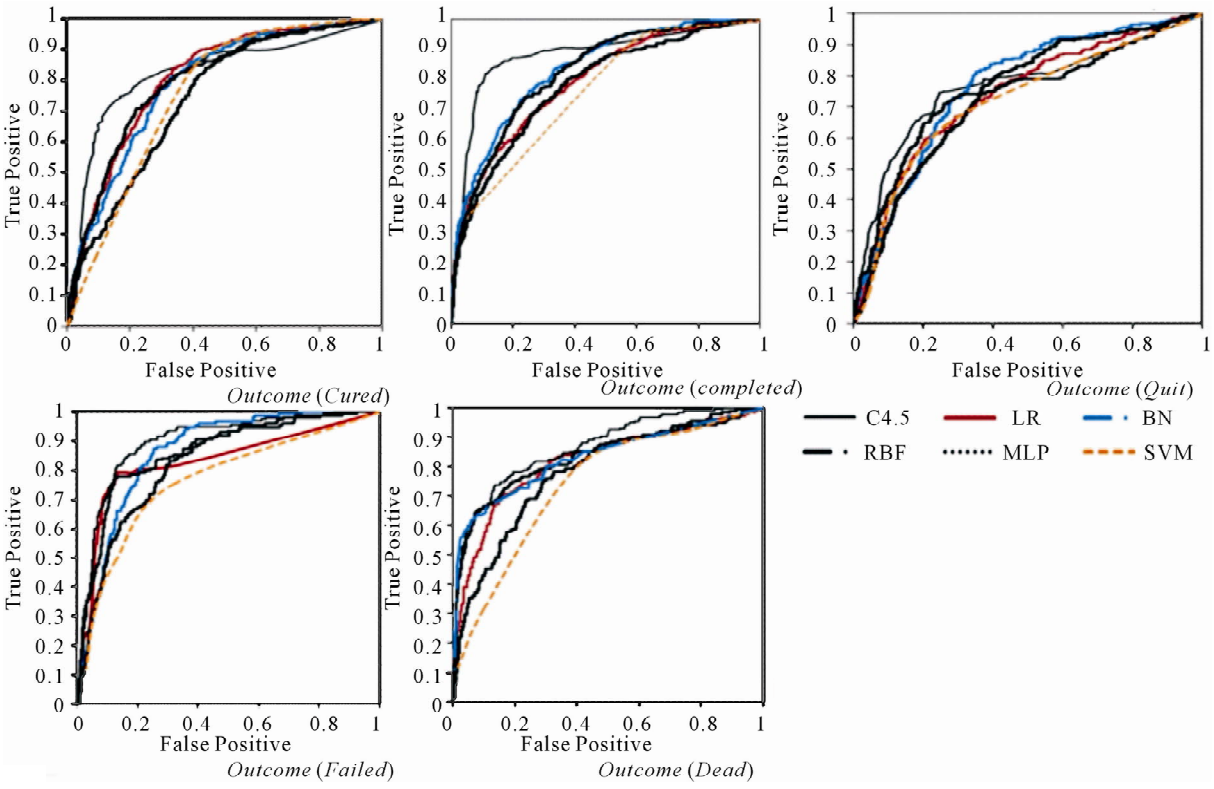

Figure 4. ROC curve for C4.5 decision tree, logistic regression (LR), Bayesian networks (BN), radial basis function (RBF), multilayer perceptron neural networks(MLP), and support vector machine on the task of classifying outcome of TB treatment course based on the target outcome including cured, completed, quit, failed, and dead. In all cases decision tree has outperformed other methods with area under curve 0.92 for cured, 0.94 for completed, 0.88 for quit, 0.96 for failed, and 0.85 for dead.

posed of fourteen discrete variables and three continuous attributions (age, weight, and length of disease); in this case, decision tree has produced the most promising results due to its dual ability to tackle both continuous and discrete/categorical predictors which is superior to other aforementioned techniques that are good at handling only continuous variables.

In the case of emerging relationship among attributions, BNs don't perform well to manage learning properly. There have been significant relationships between seventeen predictors and outcome class which may cause the weaker performance of Bayesian networks rather than C4.5; furthermore, discriminant algorithms like logistic regression also fail on this type of data with high correlation between the attributes. In this study, there are many correlations among variables, like weight and nationality ($r = -0.052$, $P < 0.001$), LBW and sex ($r = -0.047$, $P < 0.001$), imprisonment and sex ($r = -0.156$, $P < 0.001$), prison and weight ($r = 0.065$, $P < 0.001$), length and nationality ($r = 0.099$, $P < 0.001$). Those correlations in addition to applying fourteen discrete inputs might cause weaker results from BN and LR rather than C4.5. is good at coping with irrelevant data. This might be the case in this particular study since there are some variables with very low correlation coefficient that decision tree has not taken them very much to built the model and not at all in the main root nodes. For example, area (−0.027), prison (−0.026), diabetes (0.029) have low correlation coefficient, where this value for recent TB infection, imprisonment, IV drug using, sex are 0.250, 0.151, 0.172, 0.160 respectively. The variables with high correlation coefficient like recent TB infection, length, imprisonment, and treatment category have played a major role as root nodes whereas the variables with small correlation coefficient have been recognized as less important factors placed in very sub-nodes close to leaves which can be even pruned. Decision tree's ability of utilizing significant input factors in basis of their degree of contribution to estimate outcome of tuberculosis treatment course creates greater predictive model than those classifiers, such as MLP, RBF, and SVM which count every of input variable uniformly by weightening affecting the results' transparency.

The higher values for Kurtosis (>7) and skew (>1) denote that our variables are far from normality and decision tree and other symbolic methods (nonparametric

schemes) tackle robustly with distributions with large kurtosis and skew [22]. In this research's dataset, the average values of skew and kurtosis are 2.169 and 7.469 respectively ($P < 0.05$) confirming the non-normal distribution. Hence, the only available nonparametric symbolic learning algorithm in the current study is decision tree which performed well to partition the input space. In actuality, high skew (>1) or kurtosis (>7) along with the presence of binary/categorical variables, using relevant and correlated predictors without any missed instances or noised data have prepared the best opportunity for decision tree to predict more accurately than other applied algorithms. In view of the rank of other employed classifiers, BN outperforms others and four remaining classifiers work relatively similar with prediction accuracy percentage ranging from 53.74% to 57.82%. In a study [24], Tu has reviewed a number of researches concluding that LR & MLP perform closely; it is the case here where they performed with identical prediction accuracy (57.82). RBF is actually a type of neural network and it might be a postulation that based on their algorithm similarities and data type entity results are comparable.

Decision tree with flowchart-type structure is more likely method to be understandable for general users with low level of specialized knowledge about TB. Produced results of decision tree can be simply interpretable and applicable; their rules can be understood either by doctors or health practitioners who implement DOTS in rural area and make decision alone.

5. Conclusion

To sum up, available big body of real data related to TB patients has created an opportunity to generate accurate models which can predict outcome of DOTS therapy. This provides us information about outcome of treatment course for each patient and defines who needs high level of supervision and support; this is valuable as it is not possible to give every single of patients a full supervision and support distinctly. The decision tree model can be used to screen risky patients for fail in treatment course completion in population using general data gathering in routine general practice. This will help healthcare practitioners especially in rural regions to evaluate the risks of MDR-TB among their patients quickly, inexpensively, and noninvasively. TB control through totally implemented DOTS therapy is such a crucial stage in public health improvement and promotion.

6. Acknowledgements

We would like to express our gratitude to Iranian Ministry of Health and Medical Education for funding and data access. Furthermore, the authors declare that they have no conflict of interest.

REFERENCES

[1] A. D. Harries and C. Dye, "Tuberculosis," *Annals of Tropical Medicine and Parasitology*, Vol. 100, No. 5, 2006, pp. 415-443. doi:10.1179/136485906X91477

[2] World Health Organization, "The Stop TB Strategy, Building on and Enhancing DOTS to Meet the TB-Related Millennium Development Goals," 2006.

[3] W. D. Cuneo and D. J. Snider, "Enhancing Patient Compliance with Tuberculosis Therapy," *Clinics in Chest Medicine*, Vol. 10, No. 3, 1989, pp. 375-380.

[4] H. G. Tangüis, J. A. Caylà, P. García, J. M. Jansà and M. T. Brugal, "Factors Predicting Non-Completion of Tuberculosis Treatment among HIV-Infected Patients in Barcelona (1987-1996)," *The International Journal of Tuberculosis and Lung Disease*, Vol. 4, No. 1, 2000, pp. 55-60.

[5] W. W. Yew, "Directly Observed Therapy Short-Course: The Best Way to Prevent Multidrug-Resistant Tuberculosis," *Chemotherapy*, Vol. 45, No. 2, 1999, pp. 26-33. doi:10.1159/000048479

[6] J. Legrand, A. Sanchez, F. Le Pont, L. Camacho and B. Larouze, "Modeling the Impact of Tuberculosis Control Strategies in Highly Endemic Overcrowded Prisons," *Plos One*, Vol. 3, No. 5, 2008, Article ID: e2100. doi:10.1371/journal.pone.0002100

[7] S. Thiam, A. M. Le Fevre and F. Hane, "Effectiveness of a Strategy to Improve Adherence to tuberculosis Treatment in a Resource-Poor Setting: A Cluster Randomized Controlled Trial," *Journal of the American Medical Association*, Vol. 297, No. 4, 2007, pp. 380-386. doi:10.1001/jama.297.4.380

[8] W. J. Burman, D. L. Cohn, C. A. Rietmeijer, F. N. Judson, J. A. Sbarbaro and R. R. Reves, "Noncompliance with Directly Observed Therapy for Tuberculosis. Epidemiology and Effect on the Outcome of Treatment" *Chest*, Vol. 111, No. 5, 1997, pp. 1168-1173. doi:10.1378/chest.111.5.1168

[9] P. D. O. Davies, "The Role of DOTS in Tuberculosis Treatment and Control," *American Journal of Respiratory Medicine*, Vol. 2, No. 3, 2003, pp. 203-209. doi:10.1007/BF03256649

[10] A. V. Sitar-Taut, D. Zdrenghea, D. Pop and D. A. Sitar-Taut, "Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation," *Journal of Applied Computer Science & Mathematics*, Vol. 5, No. 3, 2009, pp. 29-32.

[11] M. Lazarescu, A. Turpin and S. Venkatesh, "An Application of Machine Learning Techniques for the Classification of Glaucomatous Progression," Vol. 2396, Springer-Verlag, Berlin, 2006.

[12] J. I. Serrano, M. Tomécková and J. Zvárová, "Machine Learning Methods for Knowledge Discovery in Medical Data on Atherosclerosis," *European Journal of Biomedical Informatics*, 2006.

[13] I. Guyon and A. Elissee, "An Introduction to Variable and

Feature Selection," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1157-1182.

[14] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, Vol. 1, No. 3, 1997, pp. 131-156. doi:10.1016/S1088-467X(97)00008-5

[15] A. Field, "Discovering Statistics Using SPSS," 2nd Edition, SAGE Publication LTD, London, 2005.

[16] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2nd Edition, Morgan Kaufmann Publishers, Burlington, 2006.

[17] E. Alpaydin, "Introduction to Machine Learning," 1th Edition, The MIT Press, Cambridge, 2004.

[18] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, Vol. 31, No. 3, 2007, pp. 249-268.

[19] E. Vittinghoff, S. C. Shiboski, D. V. Glidden and C. E. McCulloch, "Regression Methods in Biostatistics, Linear, Logistic, Survival, and Repeated Measures Models," Springer, Berlin, 2005.

[20] S. Marsland, "Machine Learning: An Algorithmic Perspective," 1st Edition, Chapman and Hall, London, 2009.

[21] L. Olson and D. Delen, "Advanced Data Mining Techniques," Springer, Berlin, 2008.

[22] R. D. King, C. Feng and A. Sutherland, "Statlog: Comparison of Classification Algorithms on Large Real-World Problems," *Applied Artificial Intelligence*, Vol. 9, No. 3, 1995, pp. 289-333.

[23] I. Kurt, M. True and A. T. Kurum, "Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease," *Expert System Application*, Vol. 34, 2008, pp. 366-374. doi:10.1016/j.eswa.2006.09.004

[24] J. V. Tu, "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes," *Journal of Clinical Epidemiology*, Vol. 49, No. 11, 1996, pp. 1225-1231. doi:10.1016/S0895-4356(96)00002-9

