*IOSR Journal of Computer Engineering (IOSRJCE)*
*ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 5 (Sep-Oct. 2012), PP 05-13*
*www.iosrjournals.org*

# Estimating the Statistical Significance of Classifiers used in the Prediction of Tuberculosis

Asha.T[1], S. Natarajan[2,] K.N.B.Murthy[3]
[1] *(Department of Information Science & Engg., Bangalore Institute of Technology/VTU, INDIA)*
[2, 3]*(Department of Information Science & Engg., PES Institute of Technology/VTU, INDIA)*

**Abstract :** *Tuberculosis (TB) is a disease caused by bacteria called Mycobacterium Tuberculosis. It usually spreads through the air and attacks low immune bodies. Human Immuno deficiency Virus (HIV) patients are more likely to be attacked with TB. It is an important health problem in India as well. Diagnosis of pulmonary tuberculosis has always been a problem. Classification in medicine is an important task in the prediction of any disease. It even helps doctors in their diagnosis decisions. However the decision of best classification cannot just depend on accuracies or error rates. There is a need for critical statistical analysis of these classifiers based on some statistical tests. In this paper, a study on classification of Tuberculosis with statistical significance is realized at two stages. First stage is the comparison of accuracies by classifying TB data into two categories Pulmonary Tuberculosis(PTB) and retroviral PTB(RPTB) ie TB along with AIDS using basic learning classifiers such as C4.5 Decision Tree, Support Vector Machines (SVM), K-nearest neighbor, Bagging and Naïve Bayesian algorithms. Second stage is evaluating the performance of these classifiers using paired t-test to select the optimum model. Results for our datasets show that SVM and C4.5 Decision Tree are not statistically significant, whereas SVM with Naïve Bayes and K-nearest neighbor are statistically significant.*

**Keywords -** Classification, PTB, RPTB, Statistical significance,

## I. INTRODUCTION

Knowledge Discovery from Databases (KDD) has attracted increasing research interest in the past decade, with applications in medicine ranging from classification to rule generation [28] and to the discovery of logical connections between published results in the biomedical literature [27]. Modern hospitals nowadays are well equipped with monitoring as well as other data collection devices which provide relatively inexpensive means to collect and store the data in inter- and intra-hospital information systems. Extensive amount of data gathered in medical databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. In particular, the increase in data volume causes great difficulties in extracting useful information for decision support. Traditional manual data analysis has become inadequate, and methods for efficient computer-based analysis indispensable. To satisfy this need, medical informatics may use the technologies developed in the new interdisciplinary field of KDD [16], encompassing statistical, pattern recognition, machine learning, and visualization tools to support the analysis of data and the discovery of regularities that are encoded within the data. The KDD typically consists of the following steps [14,13]: understanding the domain, forming the dataset and cleaning the data, extracting regularities hidden in the data and formulating knowledge in the form of patterns or rules (this step in the overall KDD process is usually referred to as *data mining* (DM)), post-processing of discovered knowledge, and exploiting the results.

India has the world's highest burden of Tuberculosis (TB) with million estimated incident cases per year. It also ranks among the world's highest HIV burden with an estimated 2.3 million persons living with HIV/AIDS. Tuberculosis is much more likely to be a fatal disease among HIV-infected persons than persons without HIV infection [21]**.** It is a disease caused by mycobacterium which can affect virtually all organs, not sparing even the relatively inaccessible sites. The microorganisms usually enter the body by inhalation through the lungs. They spread from the initial location in the lungs to other parts of the body via the blood stream. They present a diagnostic dilemma even for physicians with a great deal of experience in this disease.

Data classification process using knowledge obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science and computer science. If used in medical data, they are able to help clinicians to process a huge amount of data available from solving previous cases and suggest the probable diagnosis based on the values of several important attributes [22].

The predictive ability of the classification algorithm is typically measured by its predictive accuracy (or error rate, which is 1 minus the accuracy) on the testing examples. In many data mining applications, however, accuracy is not enough. A desired property of a good classifier is the power of generalization to new, unknown instances. The detection and characterization of statistically significant predictive patterns is crucial for obtaining a good classification accuracy that generalizes beyond the training data. Unfortunately, it is very often the case that the number of available data points with labels is not sufficient. Data from medical or biological applications, for example, are characterized by high dimensionality thousands of features) and small number of data points (tens of rows). An important question is whether we should believe in the classification accuracy obtained by such classifiers. The most traditional approach to this problem is to estimate the error of the classifier by means of cross-validation or leave-one-out cross-validation, among others. This estimate, together with a variance-based bound, provides an interval for the expected error of the classifier. The error estimate itself is the best statistics when different classifiers are compared against each other (Hsing et al., 2003). However, it has been argued that evaluating a single classifier with an error measurement is ineffective for small amount of data samples (Braga-Neto and Dougherty, 2004; Golland et al., 2005; Isaksson et al., 2008). Also, the classical generalization bounds are not directly appropriate when the dimensionality of the data is too high. Hence there is a need for critical statistical analysis of classifiers in order to evaluate them.

The article discusses the statistical significance of classifiers which are used to classify Tuberculosis data [37,38]. The paper is organized as follows: section 2 presents related work, section 3 describes in brief about the data source, algorithms used, different evaluation techniques such as cross validation and significance tests. Section 4 explains the results obtained followed by conclusions.

## II. Related Work

The well developed information infrastructure of modern hospitals provides relatively inexpensive means to store the data, which can become widely available via internet. Medical information, knowledge and data keep growing on a daily basis. The ability to use these data to extract useful and new information for quality healthcare is crucial. Medical informatics plays a very important role in the use of clinical data. The development of machine learning tools for medical diagnosis and prediction was frequently motivated by the requirements for dealing with these medical datasets.

Tuberculosis (TB) is a bacteria that usually causes disease in the lung. Many people become symptom-free carriers of the TB bacteria. Although common and deadly in the third world, tuberculosis was almost non-existent in the developed world, but has been making a recent resurgence. Certain drug-resistant strains are emerging and people with immune suppression such as AIDS or poor health are becoming carriers.

Orhan Er. and Temuritus[1] present a study on tuberculosis diagnosis, carried out with the help of MultiLayer Neural Networks (MLNNs). For this purpose, an MLNN with two hidden layers and a genetic algorithm for training algorithm has been used. Data mining approach was adopted to classify genotype of mycobacterium tuberculosis using c4.5 algorithm [2]. Our proposed work is on categorical and numerical attributes of TB data with data mining technologies.

Jin Huang and Charles X. Ling [3] establish formal criteria for comparing two different measures AUC (Area Under Curve) and accuracy for learning algorithms and show theoretically and empirically that AUC is a better measure (defined precisely) than accuracy. Tuan D. Pham [4] applied the computational theories of statistical and geostatistical linear prediction models to extract effective features of the mass spectra and simple decision logic to classify disease and control samples for the purpose of early detection of cardiovascular events. A.-L. Boulesteix et.al.,[5] carefully review various statistical aspects of classifier evaluation and validation from a practical point of view. The main topics addressed are accuracy measures, error rate estimation procedures, variable selection, choice of classifiers and validation strategy. Demsar[6] reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed Ranks Test for comparison of two classifiers and the Friedman Test for comparison of more than two classifiers over multiple data sets. Polina Golland and Bruce Fischl [7] demonstrate a non-parametric technique for estimation of statistical significance in the context of discriminative analysis (i.e., training a classifier function to label new examples into one of two groups).

Radiological feature extraction for breast tumour classification was carried out utilizing supervised and unsupervised machine learning methods. The features are extracted from Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) time-course data. They employed k-Nearest Neighbor classifiers (k-NN), Support Vector Machines (SVM) and Decision Trees (DT) to classify features using a Computer Aided Design (CAD) approach.[8]. T.S. Subashini et. al[9] compares the use of polynomial kernel of SVM and Radial Basis Function Neural Network (RBFNN) in ascertaining the diagnostic accuracy of cytological data obtained from the Wisconsin breast cancer database. This research has demonstrated that RBFNN outperformed the polynomial kernel of SVM for correctly classifying the tumours [9]. Evaluation of the performance of two decision tree procedures and four Bayesian network classifiers as potential Decision Support Systems in the cytodiagnosis of breast cancer was carried out[10]. Tzung-I Tang et al., [11] discuss medical data classification methods, comparing Decision Tree and system reconstruction analysis as applied to heart disease medical data mining. Under most circumstances, single classifiers, such as Neural Networks, SVM and DT, exhibit worst performance. In order to further enhance performance combination of these methods in a multi-level combination scheme was proposed that improves efficiency [12]. R.E. Abdel-Aal [15] demonstrates the use of adductive network classifier committees, trained on different features for improving classification accuracy in medical diagnosis.

The paper entitled “Mining Several Data Bases with an Ensemble of Classifiers” [17] analyze the two types of conflicts, one created by data inconsistency within the area of the intersection of the data bases and the second created when the meta method selects different data mining methods with inconsistent competence maps for the objects of the intersected part and their combinations . The authors suggest ways to handle them. Seppo Puuronen et al.,[20] propose a similarity evaluation technique that uses a training set consisting predicates that define relationships within the three sets: the set of instances, the set of classes, and the set of classifiers. Lei Chen and Mohamed S. Kamel[18] propose the scheme of Multiple Input Representation-Adaptive Ensemble Generation and Aggregation(MIR-AEGA) for the classification of time series data. Kai Jiang et.al.[19] propose a Neural Network ensemble model for classification of incomplete data. In the method, the incomplete dataset is divided into a group of complete sub datasets, which is then used as the training sets for the neural networks.

## III. Materials And Methods

3.1 Data Source

The medical dataset used for classification includess 700 real records of patients suffering from TB obtained from a state hospital. The entire records have been digitized. Each record corresponds to most relevant information of one patient. Initial queries by doctor as symptoms and some required test details of patients have been considered as main attributes. Totally there are 11 attributes (symptoms) and one class attribute. The symptoms of each patient such as age, chroniccough(weeks), loss of weight, intermittent fever(days), night sweats, Sputum, Bloodcough, chestpain, HIV, radiographic findings, wheezing and class are considered as attributes.

Table 1 shows names of 12 attributes considered along with their Data Types (DT). Type N-indicates numerical and C is categorical.

**Table 1 : List of Attributes and their Data Types**

| No | Name | DT |
|---|---|---|
| 1 | Age | N |
| 2 | Chroniccough(weeks) | N |
| 3 | WeightLoss | C |
| 4 | Intermittentfever | N |
| 5 | Nightsweats | C |
| 6 | Bloodcough | C |
| 7 | Chestpain | C |
| 8 | HIV | C |
| 9 | Radiographicfindings | C |
| 10 | Sputum | C |
| 11 | Wheezing | C |
| 12 | Class | C |

3.2 Classification Algorithms

Let $X$ *be a data set with* $n\times m$ data matrix. We denote the $i$-th row vector of $X$ by $X_i$ and the $j$-th column vector of $X$ by $X_j$. Rows are called observations or data points, while columns are also called attributes or features. Observe that we do not restrict the data domain of $X$ and therefore the scale of its attributes can be categorical or numerical. Associated to the data points $X_i$ we have a class label $\mathrm{Y}_i$. Let $D$ be the set of labeled data $D = \{( X_i ,\mathrm{Y}_i )\}n_i = 1$.In a traditional classification task the aim is to predict the label of new data points by training a classifier from $D$. The function learned by the classification algorithm is denoted by $f : X$ maps $Y$. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. A test statistic is typically computed to evaluate the classifier performance: this can be either the training error, cross-validation error or jackknife estimate, among others.

3.2.1 C4.5 Decision Tree

Decision Tree used in data mining and machine learning, is a predictive model which maps observations about an item to conclusions about the item's target value. It is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. C4.5 Decision Tree algorithm used to generate the tree is developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. The Decision Trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. It has many advantages such as avoiding overfitting the data , reducing error pruning, handling continuous attributes and missing values[25].

1. Check for base cases
2. For each attribute *a*

- Find the normalized information gain from splitting on *a*

3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*

J48 is an open source java implementation of the C4.5 algorithm in the Weka data mining tool developed by Prof. Witten of Waikato University in New Zealand.

### 3.2.2 k-Nearest Neighbor (k-NN)

The *k*-Nearest Neighbor algorithm (*k*-NN) is a method for classifying objects based on closest training examples in the feature space. *k*-NN is a type of instance-based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. Here an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive, typically small)[23].

<u>Algorithm</u>

Input: D, the set of K training objects and test object z = x′, y′
Process:
Compute d(x′,x ) the distance between z and every object, $x,y\in(\text{belongs})D$
Select $D_z$ , a subset of D, the set of K closest training objects to z
Output: Majority Voting:
y′=argmax $\sum_v^{x,y\in(\text{belongs})D} I(v = y_i)$ where *v* is a class label, $y_i$ is the class label for the *i*th nearest neighbors, and *I (·)* is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

### 3.2.3 Naive Bayesian Classifier

It is Bayes classifier which is a simple probabilistic classifier based on applying Baye's theorem(from Bayesian statistics) with strong (naive) independence assumptions[24]. In probability theory Bayes theorem shows how one conditional probability (such as the probability of a hypothesis given observed evidence) depends on its inverse (in this case, the probability of that evidence given the hypothesis). In more technical terms, the theorem expresses the posterior probability (i.e. after evidence E is observed) of a hypothesis H in terms of the prior probabilities of H and E, and the probability of E given H. It implies that evidence has a stronger confirming effect if it was more unlikely before being observed.

Naïve assumption is mainly based on attribute independence
$P(x_1,\ldots,x_k|C) = P(x_1|C)\cdot\ldots\cdot P(x_k|C)$
If i-th attribute is categorical:
$P(x_i|C)$ is estimated as the relative freq of samples having value $x_i$ as i-th attribute in class C
If i-th attribute is continuous:
$P(x_i|C)$ is estimated thru a Gaussian density function computationally easy in both cases.

### 3.2.4 Bagging

Bagging (Bootstrap aggregating) was proposed by Leo Breiman in 1994 to improve the classification by combining classifications of randomly generated training sets.The concept of bagging (voting for classification, averaging for regression-type problems with continuous dependent variables of interest) applies to the area of predictive data mining to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data.

```
1. for $m$ = 1 to $M$     // $M$ ... number of iterations
a) draw (with replacement) a bootstrap
sample $S_m$ of the data
b) learn a classifier $C_m$ from $S_m$
2. for each test example
a) try all classifiers $C_m$
b) predict the class that receives the highest number of votes.
```

### 3.2.5 SVM

The original SVM algorithm was invented by Vladimir Vapnik. The standard SVM takes a set of input data, and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier.

A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Let data D be ($X_1$, $Y_1$), …, ($\mathbf{X}_{|D|}$, $Y_{|D|}$), where $\mathbf{X}_i$ is the set of training tuples associated with the class labels $Y_i$ .There are infinite lines (hyperplanes) separating the two classes but we want to find the best one (the one that minimizes classification error on unseen data).SVM searches for the hyperplane with the largest margin, i.e., Maximum Marginal Hyperplane (MMH).A separating hyperplane can be written as $\mathbf{W} \bullet \mathbf{X} + b = 0$ where $\mathbf{W}=\{w_1, w_2, \ldots, w_n\}$ is a weight vector and b a scalar (bias).For 2-D it can be written as $w_0 + w_1 x_1 + w_2 x_2 = 0$.The hyperplane defining the sides of the margin:

$H_1$: $w_0 + w_1 x_1 + w_2 x_2 \geq 1$ for $y_i = +1$, and

$H_2$: $w_0 + w_1 x_1 + w_2 x_2 \leq -1$ for $y_i = -1$

Any training tuples that fall on hyperplanes $H_1$ or $H_2$ (i.e., the sides defining the margin) are support vectors. This becomes a constrained (convex) quadratic optimization problem. The complexity of trained classifier is characterized by the # of support vectors rather than the dimensionality of the data. The support vectors are the essential or critical training examples —they lie closest to the decision boundary (MMH).If all other training examples are removed and the training is repeated, the same separating hyperplane would be found. The number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality. Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high

Instead of computing the dot product on the transformed data tuples, it is mathematically equivalent to instead applying a kernel function $K(\mathbf{X_i}, \mathbf{X_j})$ to the original data, i.e., $K(\mathbf{X_i}, \mathbf{X_j}) = \Phi(\mathbf{X_i})\ \Phi(\mathbf{X_j})$.Typical Kernel Functions are :

$$\text{Polynomial kernel of degree } h: \quad K(X_i, X_j) = (X_i \cdot X_j + 1)^h$$

$$\text{Gaussian radial basis function kernel}: \quad K(X_i, X_j) = e^{-\|X_i - X_j\|^2/2\sigma^2}$$

$$\text{Sigmoid kernel}: \quad K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta)$$

SVM can also be used for classifying multiple (> 2) classes and for regression analysis (with additional user parameters).

### 3.3 Cross-Validation

Cross-Validation (CV) is the standard Data Mining method for evaluating performance of classification algorithms, mainly to evaluate the error rate of a learning technique. In CV a dataset is partitioned in n folds, where each is used for testing and the remainder used for training. The procedure of testing and training is repeated n times so that each partition or fold is used once for testing. The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use a stratified 10-fold cross-validation. Stratification implies making sure that when sampling is done each class is properly represented in both training and test datasets. This is achieved by randomly sampling the dataset when doing the n fold partitions.

In a stratified 10-fold Cross-Validation the data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. The learning procedure is executed a total of 10 times on different training sets, and finally the 10 error rates are averaged to yield an overall error estimate. When seeking an accurate error estimate, it is standard procedure to repeat the CV process 10 times. This means invoking the learning algorithm 100 times. Given two models M1 and M2 with different accuracies tested on different instances of a data set, to say which model is best, we need to measure the confidence level of each and perform significance tests.

### 3.4 Significance Tests / Hypothesis Testing

Estimating statistical significance of classifiers used in medical data is a challenging problem due to the high dimensionality of the data and the relatively small number of training examples. Our approach adopts significance tests/hypothesis tests, first developed in classical statistics for hypothesis testing, to estimate how likely we are to obtain the observed classification performance.

Suppose someone suggests a *hypothesis* that a certain population of data samples is 0. Recalling the convoluted way in which statistics works, one way to do this would be to

- construct a confidence interval(CI) for the population mean and
- *reject the hypothesis* if the interval failed to include 0.
- We would *fail to reject the hypothesis* if the interval contained 0.

We fail to reject the hypothesis if

$$\bar{x} - 1.96\ SEM \leq 0 \leq \bar{x} + 1.96\ SEM$$

which can be rewritten as

$$-1.96 \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le +1.96$$

On the other hand, we reject the hypothesis if

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \le -1.96 \quad \text{or} \quad \frac{\bar{x}-\mu}{s/\sqrt{n}} \ge 1.96$$

The statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is denoted by the symbol *t*. The test can be summarized as: Reject the hypothesis that the mean of population of data samples is 0 if and only if the absolute value of *t* is greater than 1.96.

There is a 5% chance of obtaining a 95% CI that excludes 0 when it is in fact the population mean. For this reason, we say that this test has been performed at the 0.05 level of significance. Had a 99% CI been used, we would say that the test had been performed at the 0.01 level of significance, that is, the *significance level* (or simply the *level*) of the test is the probability of rejecting a hypothesis when it is true.

Statistical theory says that in many situations where a population value is estimated by drawing random samples, the sample and population values will be within two standard errors of each other 95% of the time. That is, 95% of the time,

*-1.96 SE* ≤*population value - sample value* ≤*1.96 SE* [*]

This is the case for means, differences between means, proportions, differences between proportions, and regression coefficients. After an appropriate transformation, this is the case for odds ratios and even correlation coefficients.

We have used this fact to construct 95% confidence intervals by restating the result as

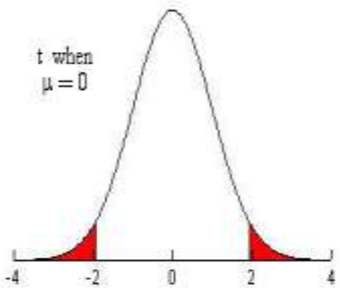
*sample value - 1.96 SE* ≤*population quantity* ≤*sample value + 1.96 SE*

For example, a 95% CI for the difference between two population means, $\mu_x$-$\mu_y$, is given by

$$(\bar{x}-\bar{y}) - 1.96\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}} \le \mu_x - \mu_y \le (\bar{x}-\bar{y}) + 1.96\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}.$$

When we perform significance tests, we reexpress [*] by noting that 95% of the time

$$-1.96 \le \frac{\text{sample value} - \text{population quantity}}{SE} \le 1.96$$



Suppose you wanted to test whether a population quantity were equal to 0. You could calculate the value of

$$t = \frac{(\text{sample value}) - 0}{SE}$$

which we get by inserting the hypothesized value of the population mean difference (0) for the population_quantity. If *t<-1.96* or *t>1.96* (that is, |*t*|>*1.96*), we say the data are not consistent with a population mean difference of 0 (because *t* does not have the sort of value we expect to see when the population value is 0) or "we reject the hypothesis that the population mean difference is 0". If t were -3.7 or 2.6, we would reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unusual if the hypothesis were true.

If -*1.96*≤*t*≤*1.96* (that is, |*t*|≤*1.96*), we say the data are consistent with a population mean difference of 0 (because *t* has the sort of value we expect to see when the population value is 0) or "we fail to reject the hypothesis that the population mean difference is 0". For example, if t were 0.76, we would fail reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unremarkable if the hypothesis were true.

## IV. Results

The open source tool Weka was used in different phases of the knowledge discovery process. Weka is a collection of state-of-the-art data mining algorithms and data preprocessing methods for a wide range of tasks such as data preprocessing, attribute selection, clustering, and classification. Weka has been used in prior research both in the field of clinical data mining and in bioinformatics.

Table 2 summarizes the mean accuracies of all five classifiers. Though researchers theoretically and practically say that SVM is the best classifier, for our data set there is no much significant difference in the accuracies of SVM and Decision trees. SVM proves to be with 99.07 accuracy against decision tree with 99%.

Bagging proves to be better than K-nearest neighbor and lastly Naïve Bayes. Table 3 shows the result of all five classifiers over each set which is required for computing statistical tests.

**Table 2: Comparison of Mean Accuracy of All Classifiers**

| Classifier | Accuracy |
|---|---|
| SVM | 99.07% |
| C4.5 decision Trees | 99.00% |
| Bagging | 98.96% |
| Naïve Bayes | 96.63% |
| K-NN | 97.89% |

**Table 3: Comparison of Classification Accuracy Over Ten Sets**

| Dataset | SVM | C4.5 DT | Naive Bayes | Bagging | K-NN |
|---|---|---|---|---|---|
| 1 | 99.14 | 99.00 | 96.43 | 98.86 | 98.43 |
| 2 | 99.14 | 99.00 | 96.57 | 99.00 | 97.86 |
| 3 | 99.14 | 99.00 | 96.71 | 99.00 | 98.00 |
| 4 | 99.14 | 99.00 | 97.00 | 98.86 | 98.00 |
| 5 | 98.86 | 99.00 | 97.14 | 99.00 | 97.71 |
| 6 | 99.00 | 99.00 | 96.57 | 99.00 | 97.57 |
| 7 | 99.14 | 99.00 | 96.43 | 99.00 | 98.00 |
| 8 | 99.14 | 99.00 | 96.57 | 98.86 | 98.00 |
| 9 | 99.14 | 99.00 | 96.14 | 99.00 | 97.86 |
| 10 | 98.86 | 99.00 | 96.71 | 99.00 | 97.43 |

Proving the Statistical significance based on paired t-statistic

t-statistic is given by $d_{avg}$/sqrt(var/k) where $d_{avg}$ = errorrate$_i$(mean-avg)- errorrate$_j$(mean-avg) [errorrate=1-accuracy], var is given by $\sum(d- d_{avg})^2$ /k where d = errorrate$_i$ – errorrate$_j$ for each set of k=10 fold cross validation.

Table 4 shows the calculated errorate comparison of SVM and Naïve Bayes and Table 5 is comparison between SVM and Decisiontree.

**Table 4: Comparing SVM and Naive Bayes**

| **errorrate$_i$** | **errorrate$_j$** | **D** | **$d_{avg}$** | **d- $d_{avg}$** | **(d- $d_{avg)}$$^2$** |
|---|---|---|---|---|---|
| 0.0086 | 0.0357 | 0.0271 | | 0.0027 | 0.00000729 |
| 0.0086 | 0.0343 | 0.0257 | | 0.0013 | 0.00000169 |
| 0.0086 | 0.0329 | 0.0243 | | 0.0001 | 0.00000001 |
| 0.0086 | 0.03 | 0.0214 | | 0.003 | 0.000009 |
| 0.0114 | 0.0286 | 0.0172 | 0.0244 | 0.0072 | 0.00005184 |
| 0.01 | 0.0343 | 0.0243 | | 0.0001 | 0.00000001 |
| 0.0086 | 0.0357 | 0.0271 | | 0.0027 | 0.00000729 |
| 0.0086 | 0.0343 | 0.0257 | | 0.0013 | 0.00000169 |
| 0.0086 | 0.0386 | 0.03 | | 0.0056 | 0.00003136 |
| 0.0114 | 0.0329 | 0.0215 | | 0.0029 | 0.00000841 |

variance of two models SVM and Naïve Bayes is 0.000011859.

Hence t-statistic value is 22.19.

Looking into t-distribution table for k-1 intervals at 0.005 significance level, the confidence limit z=3.2. since t (t-statistic value) > z, we reject the null hypothesis that two classifiers are same and conclude that there is a significant difference between them.

**Table 5: Comparing SVM and C4.5 Decision Tree**

| | **$\text{errorrate}_j$** | **D** | **$d_{avg}$** |
|---|---|---|---|
| 0.0086 | 0.01 | 0.0014 | |
| 0.0086 | 0.01 | 0.0014 | |
| 0.0086 | 0.01 | 0.0014 | |
| 0.0086 | 0.01 | 0.0014 | |
| 0.0114 | 0.01 | 0.0014 | 0.0007 |
| 0.01 | 0.01 | 0.0000 | |
| 0.0086 | 0.01 | 0.0014 | |
| 0.0086 | 0.01 | 0.0014 | |
| 0.0114 | 0.01 | 0.0014 | |

$\text{errorrate}_i$(mean_avg) = 0.0093 and $\text{errorrate}_j$(mean_avg) = 0.01,hence $d_{avg}$ = 0.0007, which is almost zero and follows null hypothesis that there are no significant differences between them.

## V. Conclusions

There is a great concern regarding prediction of Tuberculosis along with HIV infected patients. Machine learning techniques such as classification can be applied with a great deal to this data set. Performance of a classifier in the classification of tuberculosis is dealt with statistical significance tests. It was found that SVM and Decision trees are not statistically significant whereas SVM and Naive Bayes have statistical significant differences between them. We would like to improve the dataset by including even the laboratory details.

## VI. Acknowledgement

Our thanks to KIMS Hospital, Bangalore for providing the valuable real Tuberculosis data and principal Dr. Sudharshan for giving permission to collect data from the Hospital.

## References

[1] Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, Tuberculosis disease diagnosis using Artificial Neural networks , DOI 10.1007/s10916-008-9241-x *online, Journal of Medical Systems, Springer,* 2008.

[2] M. Sebban, I. Mokrousov, N. Rastogi and C. Sola, A data-mining approach to spacer oligo nucleotide typing of Mycobacterium tuberculosis, *Bioinformatics, oxford university press, Vol 18, issue 2,* pp 235-243 2002.

[3] Jin Huang and Charles X. Ling , Using AUC and Accuracy in Evaluating Learning Algorithms, *IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 3,* March 2005.

[4] Tuan D. Pham, *Senior Member, IEEE,* Honghui Wang, Xiaobo Zhou, Dominik Beck, Miriam Brandl, Gerard Hoehn, Joseph Azok, Marie-Luise Brennan, Stanley L. Hazen, King Li, and Stephen T. C. Wong, Computational Prediction Models for Early Detection of Risk of Cardiovascular Events Using Mass Spectrometry Data , *IEEE Transactions On Information Technology In Biomedicine, Vol. 12, No. 5,* September 2008.

[5] A.-L. Boulesteix, C. Strobl, T. Augustin and M. Daumer, Evaluating Microarray-based Classifiers: An Overview, *Cancer Informatics 2008:6.* pp77–97

[6] Janez Demsar , Statistical Comparisons of Classifiers over Multiple Data Sets, *The Journal of Machine Learning Research, Volume 7, 12/1/2006.*

[7] Polina Golland and Bruce Fischl, Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies, ,*IPMI* pp330-341,2003.

[8] Tim W. Nattkempera, Bert Arnricha, Oliver Lichtea, Wiebke Timma, Andreas Degenhardb, Linda Pointonc, Carmel Hayesc,, Martin O. Leachc ,Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods, pp 129-139 ,*Artificial Intelligence in Medicine, Elsevier Science Publishers ,Volume 34 , Issue 2* , June 2005.

[9] T.S. Subashini , V. Ramalingam, S. Palanivel ,Breast mass classification based on cytological patterns using RBFNN and SVM, pp 5284-5290, *Expert Systems with Applications, Elsevier, Volume 36, Issue 3, Part 1,* April 2009.

[10] Nicandro Cruz-Ramırez , Hector-Gabriel Acosta-Mesa , Humberto Carrillo-Calvet , Rocıo-Erandi Barrientos-Martınez ,Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks, pp 1331–1342, *Applied Soft Computing, Elsevier, volume 9,issue 4,*September 2009.

[11] Tzung-I Tang,Gang Zheng ,Yalou Huang ,Guangfu Shu ,A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis, pp. 102-108, *IEMS ,Vol. 4, issue 1,* June 2005.

[12] Tsirogiannis, G.L. Frossyniotis, D. Stoitsis, J. Golemati, S. Stafylopatis, A. Nikita, K.S., Classification of medical data with a robust multi-level combination scheme, pp 2483- 2487, *Proc. IEEE International Joint Conference on Neural Networks, 25-29 July 2004, volume 3.*

[13] Fayyad UM, Uthurusamy R. ,Data mining and knowledge discovery in databases (editorial),pp 24–6,*Commun ACM 1996;39(11*).

[14] Fayyad UM, Piatetsky-Shapiro G, Smyth P.,The KDD process for extracting useful knowledge from volumes of data, pp27–41,*Commun ACM 1996;39(11).*

[15] R.E. Abdel-Aal,"Improved classification of medical data using abductive network committees trained on different feature subsets, pp 141-153, *Computer Methods and Programs in Biomedicine, Volume 80, Issue 2,* 2005.

[16] Frawley W, Piatetsky-Shapiro G, Matheus C., Knowledge discovery in databases: an overview, *Knowledge discovery in databases. Menlo Park, CA: The AAAI Press, 1991.*

[17] Seppo Puuronen, Vagan Terziyan and Alexander Logvinovsky, Mining Several Data Bases With an Ensemble of Classifiers, in proc**.,** PP: 882 – 891, *Proc. 10th International Conference on Database and Expert Systems Applications, Vol.1677***,***1999.*

[18] Lei Chen and Mohamed S. Kamel ,New Design of Multiple Classifier System and its Application to the time series data, pp: 385-391, *proc. IEEE International Conference on Systems, Man and Cybernetics*, 2007.

[19] Kai Jiang, Haixia Chen, Senmiao Yuan ,Classification for Incomplete Data Using Classifier Ensembles, *Neural Networks and Brain*, 2005.

[20] Seppo Puuronen and Vagan Terziyan , A Similarity Evaluation Technique for Data Mining with an Ensemble of classifiers,pp:163-174, *Proc.Cooperative Information Agents III, Third International Workshop, CIA* 1999.

[21] HIV Sentinel Surveillance and HIV Estimation, 2006. New Delhi, India: National AIDS Control Organization, Ministry of Health and Family Welfare, Government of India. http://www.nacoonline.org/Quick_Links/HIV_Data/ Accessed 06 February, 2008.

[22] Fatemeh Hosseinkhah, Hassan Ashktorab, Ranjit Veen, M. Mehdi Owrang ,Applying Data Mining Techniques to Medical Databases, *Proc. International Conference on Information Resources Management (CONF-IRM)* (CONF-IRM 2008.

[23] Statnikov A, Aliferis C, Tsamardinos I, Hardin D, Levy S: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis**.** PP 631-643, Bioinformatics,21(5) 2005.

[24] F. Provost and P. Domingos, Tree Induction for Probability-Based Ranking, pp.199-215, *Machine Learning, vol.52, no.3*, 2003.

[25] F. Provost and T. Fawcett, Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distribution, pp. 43-48, *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining*,1997.

[26] F. Provost, T. Fawcett, and R. Kohavi, The Case Against Accuracy Estimation for Comparing Induction Algorithms, pp. 445-453, *Proc. 15th Int'l Conf. Machine Learning,*1998.

[27] Swanson DR, Smalheiser NR. Undiscovered public knowledge: a ten-year update. In: Simoudis E, et al., editors. pp 295–298, *Proc. the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon,* August 1996.

[28] Tsumoto S, Tanaka H. Automated discovery of medical expert system rules from clinical databases based on rough sets. In: Simoudis E, et al., editors. PP 63–69, *Proc. Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon* 1996.

[29] Thomas M. Cover and Peter E. Hart, Nearest neighbor pattern classification, pp. 21-27, *IEEE Transactions on Information Theory, Vol. 13, issue 1*, 1967.

[30] Rish, Irina.(2001) , An empirical study of the naïve Bayes classifier, workshop on empirical methods in artificial intelligence, *IJCAI 2001*.

[31] J.R. QUINLAN , *Induction of Decision Trees , pp 81-106 ,Machine Learning 1*(Kluwer Academic Publishers, Boston, 1986),

[32] R. J. Quinlan, Bagging, boosting, and c4.5, pp.725-730 , *Proc. 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference*. Portland, Oregon, AAAI Press / The MIT Press, Vol. 1, 1996).

[33] Corinna Cortes and V. Vapnik , Support-Vector Networks, *Machine Learning, 20,* 1995.http://www.springerlink.com/content/k238jx04hm87j80g/

[34] X. Wu, V. Kumar, Ross, J. Ghosh, Q. Yang, H. Motoda, G.Mclachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand and D. Steinberg, Top 10 algorithms in data mining, pp. 1-37, Knowledge and Information Systems, vol. 14, no.1, January 2008.

[35] J. Han and M. Kamber. *Data mining: concepts and techniques*: (Morgan Kaufmann Pub, 2006).

[36] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (*Second Edition, Morgan Kaufmann Pub, 2005).

[37] Asha.T, S. Natarajan, K.N.B.Murthy, Diagnosis of Tuberculosis using Ensemble Methods, *Proc. of 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, China, Volume 8, pages 409-412*, July 2010.

[38] Asha.T, S. Natarajan, K.N.B.Murthy, Effective Classification Algorithms to Predict the Accuracy of Tuberculosis- A Machine Learning Approach, *IJCSIS, Volume 9, No.7, ISSN 1947-5500*, July 2011.