

Decision Tree-Inspired Classification Algorithm for Early Detection of Tuberculosis (TB)

Completed Research Paper

Syed Fahadulla Hussainy

Lahore University of Management
Sciences
DHA, Lahore, Pakistan
sfhussainy@hotmail.com

Fareed Zaffar

Lahore University of Management
Sciences
DHA, Lahore, Pakistan
fareed.zaffar@lums.edu.pk

Muhammad Adeel Zaffar

Lahore University of Management
Sciences
DHA, Lahore
adeel.zaffar@lums.edu.pk

Aasia Khaliq

College of Earth and Environmental
Sciences, University of the Punjab
Lahore, Pakistan
aasia.khaliq.pu@gmail.com

Imran H. Khan

Department of Pathology and
Laboratory Medicine, University of
California, Davis
California, United States of America
ihkhan@uc.davis.edu

Raza Ahmad

Lahore University of Management
Sciences
DHA, Lahore, Pakistan
syed.raza@lums.edu.pk

Abstract

Pakistan has the fifth highest incidence of Tuberculosis (TB) in the world. The current tests for TB diagnosis are smear and culture tests which have accuracies of about 40% and 70% respectively. They are inefficient, complicated and relatively expensive to perform in developing countries. In this paper, we present a novel computational predictive algorithm by modifying the standard decision tree that allows us to efficiently detect TB with high accuracy. It employs supervised learning to classify the samples into patients and healthy groups based on MFI values of different antibodies with an accuracy of about 94%, outperforming the traditional classifiers including decision trees, kNN, random forests and support vector machines (SVM). Our algorithm allows simultaneous, bench-top analysis of thousands of samples per day, does not rely on highly skilled or technical staff and provides easily interpretable results. Our work strongly suggests that developing, testing and implementing automated diagnostic algorithms such as ours can be helpful in overcoming infrastructural and human-resource constraints in poorly resourced countries.

Keywords: Supervised learning, Decision Trees, Classification

Motivation

According to the World Health Organization, Pakistan is ranked amongst ‘crisis countries’ in terms of human resources available to help support the achievement of the health-related Millennium Development Goals (Health Facility Assessment, 2013). The inability of the country to recruit and retain specialized medical help is highlighted by the fact that Pakistan has about 1.13 doctors per 1000 inhabitants (Health Facility Assessment, 2013). While primary and secondary health-care is provided through a network of basic health units (BHU), rural health centers, dispensaries and district headquarter hospitals, these public-sector facilities often lack proper resources and fail to meet appropriate global healthcare standards. The decades-old, tiered healthcare delivery infrastructure has not kept up with the rising population in the country. Furthermore, there is an imbalance in the allocation of public health resources between urban and rural areas. For example, a recent health facility assessment report showed that 60% of the BHUs had less than 50% of the essential equipment needed for daily operation (Health Facility Assessment, 2013).

Ironically, it is in these areas that a substantial part of the population is at a much higher risk from chronic infectious diseases, such as *Mycobacterium Tuberculosis* (TB), where prompt diagnosis and early treatment are essential for disease control. Pulmonary TB is a highly transmittable disease that has typically been associated with poverty and low socio-economic status in third-world countries. While most patients can recover successfully from an infection, speed and accuracy of diagnosis is a critical challenge where late diagnosis can lead to increased severity of symptoms and a higher mortality rate. The countries in the sub-continent, namely Pakistan, India and Bangladesh have consistently ranked among the 22 TB High Burden Countries (HBC) (WHO, 2015). With an overall population of 185 million people, Pakistan has the fifth highest incidence of TB in the world with the recent WHO estimates suggesting 267 TB patients per 100,000 people in Pakistan (WHO, 2015). More importantly, it ranks third in the list of countries that have the most missed TB cases and the highest levels of under-diagnosis of TB.

Most of the prevalent TB diagnostic methods are either subjective, inefficient, or not cost-effective. The most common method for diagnosing TB is through a sputum smear examined under a conventional microscope (Steingart et al., 2006). While the test has high specificity, the sensitivity of the test varies from 20-70% and is often directly dependent on the availability of trained professional staff required for the careful preparation, staining and reading of each sputum sample (Steingart et al., 2006). The overall process is labour intensive and slow. Furthermore, the results require subjective reading of the stained samples. Diagnosis is typically confirmed through a culture test that requires sophisticated labs along with trained technical staff. The standard culture methods can also take up to 8 to 12 weeks to obtain results since *M. tuberculosis* is a slow-growing organism (Khan et al., 2011). Most third-world countries such as Pakistan, however, lack the basic infrastructural and human resource capacity for accurate and reliable culture tests along with the capacity to handle high-volume of cases. Clinical judgements based on symptoms, X-rays and radiological assessments over several visits to the physician are often used to supplement diagnosis and this is likely to continue until faster, cheaper and more reliable tests become available.

Introduction

In this paper, we propose a method for efficient detection of TB. Biomarkers are naturally occurring molecules in human bodies which can be used to identify diseases and their severity. They can be effectively used for tuberculosis diagnosis as suggested by a study on plasma antibody profiles as diagnostic biomarkers for tuberculosis (Khan et al., 2011). We propose a supervised learning algorithm inspired by decision trees for efficient detection of TB using data on antigens acquired through multiplex microbead immunoassay (MMIA) technique using xMAP technology developed by Luminex (Soni et al., 2011). In our case, antigens are used as biomarkers to measure the expression of bacteria in a blood sample using multiplex technology which records the levels of expression of these antigens. Mean fluorescent intensities (MFI) is used as a measure to count the expression of antigens against the bacteria which is recorded through multiplex technology. We developed our model using antigens acquired from 356 samples containing 222 confirmed TB patients. Our model is able to classify the samples into patients and healthy groups with an accuracy of about 94%. This performance is superior compared to the accuracy of other conventional supervised algorithms applied in a similar manner on the same dataset.

Our results strongly suggest that a diagnostic algorithm such as ours can efficiently and effectively diagnose TB. The disease can also be detected much quickly since the test takes less than two hours to complete and as suggested earlier, prompt and accurate diagnosis without the need for specially trained health-care professionals can really help curb this menace. The rest of the paper is organized as follows. In the next section, we provide a review of the relevant literature. This is followed by a description of the data collection process and preliminary data analysis. Then the proposed classification algorithm is discussed at length followed by a detailed discussion of the results. Finally, the paper is concluded with a brief discussion of the implication of our results and directions for future research.

Literature Review

Computational predictive models of diseases have been developed using various learning techniques like decision trees, linear discriminant analysis (LDA), support vector machines (SVM), logistic regression and k-nearest neighbour (kNN). Ferrari and Aitken (2006) used Naive Bayes classifier to achieve a success rate of 97% in classification of human housekeeping genes. kNN was used to perform multi-label classification to determine the severity level of the disease (Zhang and Zhou 2005). Albitar et al. (2006) found the decision trees to be the best algorithm for prediction of clinical behavior in adult acute lymphoblastic leukemia with an accuracy range of 84% to 92%. Similarly, Soni et al. (2001) suggest that tree based methods outperform other classification algorithms in heart disease prediction. For classification of ovarian cancer, ensemble of decision trees known as random forests outperformed other techniques using mass spectrometry data (Wu et al., 2003).

Data mining techniques have also been used for the diagnosis of tuberculosis in the past. Asha et al. (2011) used support vector machines to classify pulmonary TB (PTB) and retroviral PTB, associated with HIV, from the data on symptoms and medical history of 700 patients. Various statistical analyses have been performed using the data on antigens specific to tuberculosis. In a study on diagnosis of tuberculosis using data on antigens extracted through multiplex technology, Khan et al. (2011) used multivariate analyses, classification algorithms such as random forests and sequential minimal optimization, and cluster analysis to detect TB.

Several studies have also modified the existing algorithms or used their combination for prediction modeling. Liao and Chin (2007) used penalized logistic regression catering for a small number of samples with microarray data. Loganantharaj (2007) proposed extensions of Naive Bayes algorithm to apply in bioinformatics increasing the classification accuracy in disease detection problems. Similarly, Francesca et al. (2006) used hierarchical Naive Bayes model for handling heterogeneity in detecting cancer with tissue microarrays. Fan et al. (2011) use a hybrid model from integration of case -based data clustering and fuzzy decision tree to predict breast cancer.

Researchers in the field of bioinformatics have employed different biomarkers as features for standard algorithms to detect and diagnose various diseases. This includes the use of SVM to detect breast cancer from a panel of three biomarkers (Li et al., 2002) and logistic regression for classification of patients from healthy subjects using standardized markers (Huang et al., 2013). In a related work, Jie-Kai et al. (2005) used five biomarkers to diagnose ovarian cancer with 96.7% accuracy using SVM. Similarly, Kim et al. (2013) combined multiple biomarkers to increase the classification accuracy for early screening of ovarian cancer using logistic regression. Decision tree-based ensembles were used to diagnose rheumatoid arthritis and inflammatory bowel diseases using proteomic biomarkers (Geurts et al., 2005). The study also suggested that the ensemble methodology used can be applied to similar problems.

To select the best set of features for a learning task in biological and medical domains, various algorithms have been employed, especially under the heads of filter and wrapper models (Saeys et al. 2007). Filter models are independent to the classifiers and include Chi-square, t-tests, Euclidean distance and information gain (Ben-Bassat 1982). Wrapper models, on the other hand, attach the feature selection objective with a certain classifier and then select a subset of features based on the classifier's accuracy using cross-validation (Phuong et al., 2005). Some classifiers such as decision trees and weighted naive Bayes' also have feature extraction inherent in them (Duda et al., 2012). In order to increase computational efficiency and simplification, Purkayastha et al. (2014) used forward greedy feature selection algorithm along with random forest. Additionally, linear discriminant analysis, Bayesian classifiers and k-means clustering have also been used to reduce the dimensionality of medical dataset (Ding and Li, 2003). Research in medical domain focuses on performance as well as interpretability of the predictive models. See Freitas (2013) on the interpretability of decision trees, Bayesian networks and other classifiers.

Data Collection and Preliminary Analysis

Data Summary

Data was collected from 356 samples of arriving patients at the Gulab Devi Hospital, (GDH). GDH is synonymous with the treatment and care of TB in Pakistan. It is one of the oldest hospitals in the country and was established with the view to provide treatment and care for TB patients. In our sample, 222 were confirmed as TB patients who were diagnosed at GDH. Culture, Smear and X-Ray tests of all the samples and follow-ups of the patients were performed in three different hospitals to confirm the TB incidence of the sample. From the blood samples of the individuals, the mean fluorescent intensities (MFI) values of 32 antigens, which are suspected to be specific to TB (Khan et al. 2011), were recorded. As discussed earlier, MFI is used as a measure to count the expression of antigens against the bacteria which is recorded through multiplex technology (Soni et al. 2011). The table in the Appendix shows the abbreviations assigned to each of the antigens so that they are easily referred to in the analysis. Table 1 gives the summary statistics of the MFI values of antigens by TB status, patient and healthy.

Antigen		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
Healthy	Mean	35.26	33.85	36.25	35.80	42.94	62.41	47.83	32.30	70.75	67.13	83.20	22.78	301.05	217.46	29.42	39.65
	Std. Dev	34.22	64.64	80.74	28.37	55.15	87.60	107.33	27.29	87.94	56.14	70.70	24.64	1136.45	178.85	30.31	36.13
Patients	Mean	2451.00	2124.99	875.80	1137.57	768.79	307.55	111.56	893.66	395.97	411.35	277.91	129.39	517.38	453.74	255.04	148.55
	Std. Dev	4176.97	3994.88	2454.83	2093.46	1990.98	824.74	291.24	1728.52	392.65	599.18	300.58	296.94	1115.96	470.84	1052.81	678.70

Antigen		A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32
Healthy	Mean	65.09	1674.44	37.04	66.29	43.64	79.40	54.59	66.49	160.68	49.35	245.84	55.82	22.94	462.79	100.89	191.17
	Std. Dev	87.04	3142.90	30.21	86.47	54.14	205.68	56.25	89.84	380.59	76.51	835.23	135.33	19.93	589.24	119.69	513.00
Patients	Mean	344.39	1910.14	222.93	146.98	112.10	149.89	165.22	125.04	683.50	228.71	282.55	366.53	360.95	795.25	750.58	293.28
	Std. Dev	959.76	2920.41	972.58	289.66	141.50	381.23	274.17	228.58	1300.35	515.55	580.97	1470.22	872.46	1002.21	1695.41	548.05

Table 1. Mean and standard deviation of MFI values of 32 antigens for TB (patients) and non-TB (health) subjects

It can be seen from the table that almost all the antigens have a higher average MFI value with patients' samples than with healthy samples which was expected since the antigen expresses higher MFI values when there is greater number of bacteria against that antigen (Khan et al. 2011). Box and whiskers plots were plotted to compare log₂ MFI values of each antigen between the two groups, patients and healthy (Figure 1). It is evident from the figure that in some cases, such as antigens A13 and A18, the log₂ MFI values of patients and healthy have similar distribution and hence are not discriminating between the two groups. Therefore, detailed analysis for differential expression was required.

To systematically identify features (antigens) that are discriminative for both healthy and TB patient groups, conventional learning techniques from the literature were applied. This allowed us to benchmark the results against our dataset. We then proceed to our proposed classification algorithm and compared the results.

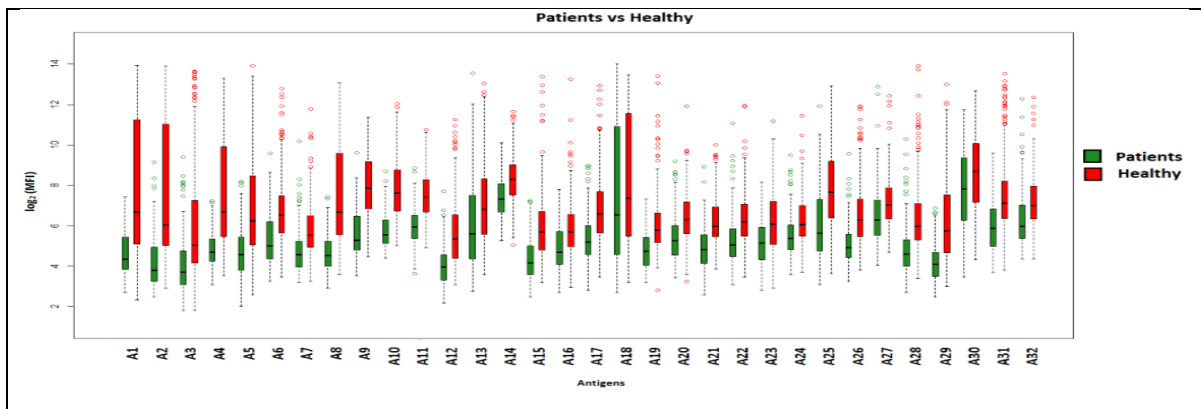


Figure 1. Box and whisker representation of each antigen's log2 MFI values for TB (patients) and non-TB (health) subjects

Feature Selection

Feature selection was done through a technique known as differential expression analysis. It is a procedure to assess how differently an antigen is expressed in a sample group compared to another group. This is done through fold changes and their statistical significance. Fold change is the number of times a quantitative measure is greater in one group relative to the other group. We determined a linear model fit for each antigen by using the limma package in R and calculated the relative fold changes (increase or decrease) of the levels of antigens between the healthy controls and patient samples according to the procedure described in (Khan et al., 2008). Fold changes were calculated by fitting a linear model to the data, and the statistical significance was estimated by a two-step process. First, an analytical data set that encompassed all the antigens for which a significantly positive signal was detected in at least one comparison between healthy and patient samples was obtained. This was followed by differential expression across the multiple comparisons detected by an F-test combined with separate F-tests for each antibody. For the selection of useful antibodies, we adjusted p-values by using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Five antigens that had MFI values which were not statistically different in patients and healthy at 1% level were discarded from further analysis (Table 2).

Anitgen	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
Patients vs Healthy	69.51**	62.78**	24.16**	31.78**	17.90**	4.93**	2.33**	27.67**	5.60**	6.13**	3.34**	5.68**	1.72*	2.09**	8.67**	3.75**
Anitgen	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32
Patients vs Healthy	5.29**	1.14	6.02**	2.22**	2.57**	1.89*	3.03**	1.88**	4.25**	4.63**	1.15	6.57**	15.74**	1.72**	7.44**	1.53*

Table 2. Fold change analysis by TB status (significance at $p < 1\%$; * significance at $p < 5\%$)**

Conventional Classification Techniques

Data on the 27 antigens obtained after feature selections for 356 samples was used to develop TB classification models using conventional techniques discussed in the literature. Six techniques discussed in the literature review section were selected: decision trees, k-nearest neighbor (kNN), logistic regression, Naïve Bayes, random forest and support vector machines (SVM). In order to obtain robust results and to avoid over fitting, three-fold cross validation technique was employed. The standard accuracy metrics (precision, recall and F1) were calculated for each of the three folds. The results revealed that decision trees and random forests performed the best followed by SVM, logistic regression, kNN and Naïve Bayes. Table 3 summarizes the results obtained through these techniques.

Classification Technique	Accuracy	Weighted F1	Patients			Healthy		
			Precision	Recall	F1	Precision	Recall	F1
k-Nearest Neighbor	79.49%	79.79%	74.32%	91.44%	81.87%	87.97%	67.41%	76.21%
Naïve Bayes	80.32%	80.58%	72.83%	94.08%	82.03%	92.47%	67.56%	78.05%
Logistic Regression	83.15%	83.35%	80.58%	91.45%	85.58%	87.46%	73.22%	79.59%
Support Vector Machine	83.42%	83.62%	79.28%	93.40%	85.52%	90.05%	72.82%	80.29%
Decision Tree	84.82%	84.79%	88.33%	87.55%	87.93%	78.74%	80.07%	79.38%
Random Forest	84.83%	84.79%	87.86%	88.04%	87.89%	79.38%	79.64%	79.37%

Table 3. Performance of conventional classification techniques on our dataset

The best weighted F1 measure is about 85% for decision trees and random forests. This is low in the context of previous literature in this domain reported earlier. Furthermore, it is low for diagnostic tests of diseases like TB, keeping in view their prevalence and severity.

Proposed Classification Algorithm

Since decision trees and random forests had the highest accuracy and weighted F1, we modified the conventional algorithm to improve the classification model. A standard decision tree is grown by splitting a node at each step that gives the maximum information gain calculated using the precision/recall of the children of that node. In our proposed algorithm, this selection criterion is modified to evaluate true positives of patients and their precision. The following steps summarize the process for one node of the tree:

Step 1: For each feature (antigen) in the training set, determine the values that discriminate the samples with highest patient precision.

Step 2: Among the values determined in Step 1, select one which has the greatest number of true positives and branch on that value of the node. That value will be called the cutoff for that antigen.

Step 3: From Step 2, the right child ($\geq$ cutoff value) will be very close to a pure set (100% precision) which need not be split anymore. If the right child is not a pure set, then stop; otherwise use the dataset of the left child ($<$ cutoff value) as input to Step 1.

Our algorithm uses count of true positives to decide which node to split into branches such that there is highest patient precision at that point. Intuitively, the modification aims at focusing on capturing true positives and avoiding false positives. False negatives are more tolerable, since further and more comprehensive testing for these subjects will either confirm or rule out the incidence of disease. False positives are not well tolerated, since it will place that healthy individual on unnecessary medication such as isoniazid prophylaxis, a usual anti-tuberculosis drug, which causes side effects and may lead to nerve dysfunction, liver inflammation, and other reactions (Marcos et al., 2010). Choosing a node on the count of true positives amounts to choosing the antigen that most frequently predicts disease incidence in the data set, while splitting a node on precision helps to avoid false positives. This makes sense in the problem scenario of interest as field testing for infectious disease in rural areas is a first-cut process and we want to capture as many positive cases as possible.

Table 4 contains sample data to illustrate the proposed algorithm. Suppose we have ten samples in the training set and three features in the feature set. Six samples belong to TB patients (labeled as Group A) and four samples belong to healthy subjects (labeled as Group B).

Group	Feature 1	Feature 2	Feature 3
A	100	500	300
A	120	400	350
A	115	450	320
A	125	430	400
A	130	350	250
A	75	470	300
B	80	375	175
B	70	325	125
B	65	200	220
B	110	250	315

Table 4. Sample data to illustrate the proposed algorithm

Using each feature for classification purposes, true positives, false positives and precision values are calculated. The results for the sample data set are shown in Figure 2.

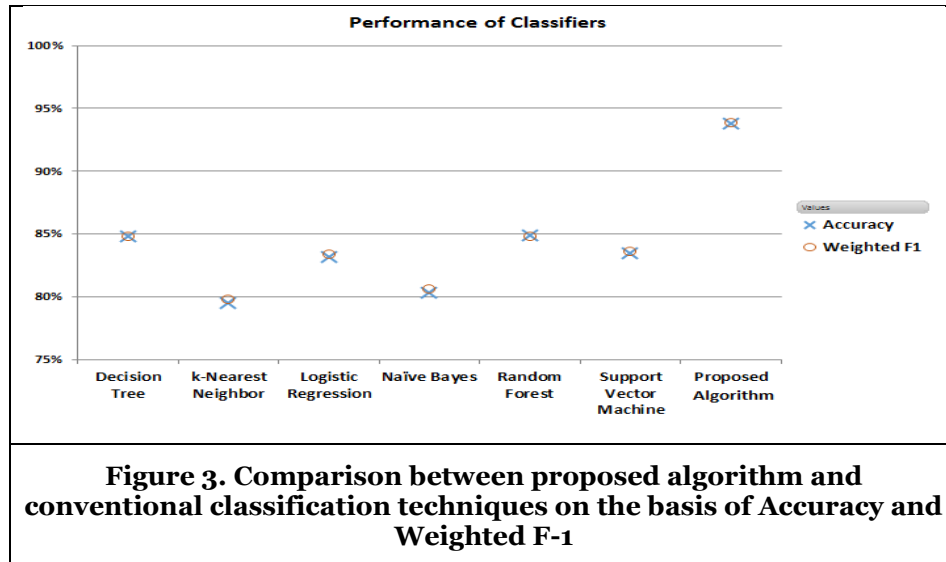
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Group	Feature 1	TP	FP	Precision	Feature 2	TP	FP	Precision	Feature 3	TP	Precision	Recall
2	A	100	5	1	83%	500	1	0	100%	300	5	1	83%
3	A	120	3	0	100%	400	5	0	100%	350	2	0	100%
4	A	115	4	0	100%	450	3	0	100%	320	3	0	100%
5	A	125	2	0	100%	430	4	0	100%	400	1	0	100%
6	A	130	1	0	100%	350	6	1	86%	250	6	1	86%
7	A	75	6	2	75%	470	2	0	100%	300	5	1	83%
8	B	80	5	2	71%	375	5	1	83%	175	6	3	67%
9	B	70	6	3	67%	325	6	2	75%	125	6	4	60%
10	B	65	6	4	60%	200	6	4	60%	220	6	2	75%
11	B	110	4	1	80%	250	6	3	67%	315	3	1	75%

Figure 2. Classification of sample data set using each feature.

The maximum precision at every value across the features is highlighted yellow in Figure 2. As discussed earlier, under the proposed algorithm the tree will branch on that feature (and value) where the count of true positives is the maximum (green highlighted cell in Figure 2). Therefore, in this case the tree will branch on Feature 2 with a value of 400. The samples that have Feature 2 value ≥ 400 will be predicted as Group A. The remaining samples will be the input for the next iteration. The iterations would stop when maximum precision % at any iteration comes out to be less than a particular value (e.g. 100%).

Results and Discussion

Our proposed modification resulted in far superior results compared to other classification algorithms. Our algorithm gave an accuracy of 93.82% and a weighted F1 of 93.6%. Figure 3 compares its performance with standard classification algorithms.

**Figure 3. Comparison between proposed algorithm and conventional classification techniques on the basis of Accuracy and Weighted F-1**

Furthermore, the tree produced a reduced set of 15 antigens by recording the antigens used in the Step 2 of proposed algorithm. These antigens facilitated in the best classification performance in comparison to the other algorithms. The cutoffs obtained for these antigens are shown in Table 5. These cutoffs can now be used to diagnose TB if the data on MFI values are recorded for each of the antigens in the multiplex panel.

A1	A2	A3	A4	A5
174	568	684	151	294
A6	A9	A11	A12	A15
774	335	467	220	158
A19	A21	A23	A26	A28
160	600	400	320	1253

Table 5. MFI cutoffs determined by our proposed algorithm for the selected 15 antigens

For classification purposes, if any of the MFI values of the 15 antigens for a new subject is greater than the cutoffs in Table 5, the subject will be diagnosed as a TB patient.

Conclusion and Future Work

In this paper, we proposed a decision tree inspired classification algorithm that provides accurate and robust results in comparison to the standard decision tree and other conventional classification algorithms. Our results are encouraging not just in the context of early detection of an infectious disease such as TB, but also in the context of developing countries where prolonged neglect of the public healthcare sector has created many challenges. In a third world country like Pakistan where TB is endemic, it is also important that a healthcare center is capable to process as many cases in a day for the early diagnosis. With this test that this paper enables, hundreds of samples can be processed in a day within one facility. Also, the cost of this diagnostic aid is considerably lower not just to the patient but also to the overall healthcare system is compared to the conventional diagnosis techniques which have high throughput. We believe our promising results further support the development of similar approaches towards fast and efficient detection of harmful diseases through the multiplex technology. As part of future research, the generalizability of this algorithm can be explored in the diagnosis of other diseases where biomarkers have higher values correlated to its incidence. Furthermore, we are in the process of conducting further research into the correlation between the selected MFI values of the antigens and severity of the disease. This may further help in early detection, ongoing treatment and care of TB patients.

References

- Albitar, M., Potts, S. J., Giles, F. J., O'Brien, S., Keating, M., Thomas, D., Clarke, C., Jilani, I., Aguilar, C., Estey, E., and Kantarjian, H. 2006. "Proteomic-based prediction of clinical behavior in adult acute lymphoblastic leukemia." *Cancer* (106:7), pp. 1587-94.
- Asha., T., Natarajan, S. and Murthy, K. N. B. 2011. "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification" *Journal of Computing*, (3:4)
- Bellman, R. 1957. *Dynamic Programming*, Princeton University Press
- Ben-Bassat, M. 1982. "Pattern recognition and reduction of dimensionality", in P. R. Krishnaiah and L. N. Kanal, (Eds), *Handbook of Statistics II*, pp. 773–791. North-Holland, Amsterdam.
- Benjamini, Y. and Hochberg, Y. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society. Series B (Methodological)*, (57:1), pp. 289-300
- Demichelis, F., Magni, P., Piergiorgi, P. and Rubin, M. A. 2006. "A hierarchical Naïve Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays", *BMC Bioinformatics*, 7, pp. 514-526
- Ding, C. and Li, T, 2003. "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, 3, pp. 1157-1182
- Duda, P., Hart, P. and Stork, D. 2012. *Pattern Classification*. Wiley, New York.
- Fan, C. F., Lin, P. C. and Hsieh, J. C. 2011. "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification", *Applied Soft Computing*, (11:1), pp. 632-644

- Ferrari, L. and Aitken, S. 2006. "Mining housekeeping genes with a Naive Bayes classifier", *BMC Geneomics*, (7:1), 277
- Freitas, A. A. 2013. "Comprehensible Classification Models: A Position Paper", *ACM SIGKDD Explorations Newsletter*, (15:1), pp. 1-10
- Geurts, P., Fillet, M., De Seny, D., Meuwis M. A., Malaise, M., Merville, M. P. and Wehenkel, L. 2005. "Proteomic mass spectra classification using decision tree based ensemble methods." *Bioinformatics*, (21:15), pp. 3138-3145.
- Hall, M. 1999. "Correlation-based Feature Selection for Machine Learning" PhD thesis, Department of Computer Science, Waikato University, New Zealand, 1999.
- Health Facility Assessment – Pakistan National Report, TR, May 2013
- Hu, J. and Yan, C. 2008. "Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information", *BMC bioinformatics* 9, pp. 297.
- Huang, Y., Pepe, M. S. and Feng, Z. 2013. "Logistic Regression Analysis with Standardized Markers", *The Annals of Applied Statistics*, (7:3), pp. 1640–1662
- Jie-Kai, Y., Shu, Z., Yong, T. and Li, L. 2005. "An integrated approach utilizing proteomics and bioinformatics to detect ovarian cancer", *Journal of Zhejiang University Science B*, (6:4), pp. 227-231.
- Khan, I.H., Ravindran, R., Yee, J., Ziman, M., Lewinsohn, D. M., Gennaro, M. L., Flynn, J. L., Goulding, C. W., DeRiemer, K., Lerche, N. W., and Luciw, P. A. 2008. "Profiling antibodies to *Mycobacterium tuberculosis* by multiplex microbead suspension arrays for serodiagnosis of tuberculosis." *Clinical and Vaccine Immunology*, (15:3), pp. 433-438.
- Khan, I.H., Ravindran, R., Krishnan, V. V., Awan, I. N., Rizvi, S. K., Saqib, M. A., Shahzad, M. I., Tahseen, S., Ireton, G., Goulding, C. W., Felgner, P., DeRiemer, K., Khanum, A. and Luciw, P. A. 2011, "Plasma antibody profiles as diagnostic biomarkers for tuberculosis", *Clinical and Vaccine Immunology*, (18:12), pp. 2148-2153.
- Kim, Y. S., Jang, M. K., Park, C. Y., Song, H. J. and Kim, D. J. 2013. "Exploring Multiple Biomarker Combination by Logistic Regression for Early Screening of Ovarian Cancer", *International Journal of Bio-Science and Bio-Technology*, (5:2), pp.67-75
- Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. and Chan D. W. 2002. "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer." *Clin Chem*, (48:8), pp. 1296-1304.
- Liao, J.G. and Chin, K. V. 2007. "Logistic regression for disease classification using microarray data: model selection in a large p and small n case", *Bioinformatics*, (23:15), pp. 1945-1951
- Loganathanaraj, R. 2007. "Extensions of Naive Bayes and Their Applications to Bioinformatics", *Bioinformatics Research and Applications: Lecture Notes in Computer Science*, 4463, pp. 282-292
- Marcos Abdo Arbex, Marília de Castro Lima Varella, Hélio Ribeiro de Siqueira, Fernando Augusto Fiúza de Mello. 2010 "Antituberculosis drugs: Drug interactions, adverse effects, and use in special situations. Part 1: First-line drugs" *J Bras Pneumol*. 2010;36(5):626-640
- Mani, A., Ravindran, R., Mannepalli, S., Vang, D., Luciw, P. A., Hogarth, M., Khan, I. H. and Krishnan V. V. 2015 "Data Mining Strategies to Improve Multiplex Microbead Immunoassay Tolerance in a Mouse Model of Infectious Diseases", *PLoS, ONE*, (10:1)
- Matthews, B.W. 1975 "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, pp. 442–451.
- Phuong, T. M., Lin, Z. and Altman, R. B. 2005. "Choosing SNPs using feature selection" *Bioinformatics*, (1:1)
- Purkayastha, P., Rallapalli, A., Murthy, N. L. B., Malapati, A., Yogeeswari, P. and Sriram, D. 2014. "Effect of Feature Selection on Kinase Classification Models", *Computational Intelligence in Medical Informatics*, pp. 81-86, 2014

- Ravindran, R., Krishnan, V. V., Dhawan, R., Wunderlich, M. L., Lerche, N. W., Flynn, J. L. Luciw, P. A. and Khan, I. H. 2014. "Plasma antibody profiles in non-human primate tuberculosis", *Journal of Medical Primatology*, (43:2), pp. 59-71.
- Saeys, Y., Inza, I. and Larranaga, P. 2007. "A review of feature selection algorithms in bioinformatics", *Bioinformatics*, (23:19), pp. 2507-2517
- Seera, M. and Lim C. P. 2014. "A hybrid intelligent system for medical data classification", *Expert Systems with Applications*, (41:5), pp. 2239-2249
- Soni, J., Ansari, U., Sharma, D. and Soni, S. 2011. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer Applications IJCA*, (17:8), pp. 43-48
- Steingart, K. R., Henry, M., Ng, V., Hopewell, P. C., Ramsay, A., Cunningham, J., Urbanczik, R., Perkins, M., Aziz, M. A. and Pai M. 2006. "Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review", *The Lancet Infectious Diseases*, (6:9), pp. 570-581.
- World Health Organization, "List of 57 countries facing Human Resources for Health crisis", *World Health Report 2006*
- World Health Organization, 2015. "Global Tuberculosis Report, 2015"
- Wu, B., T. Abbott, Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. and Zhao, H. 2003. "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data", *Bioinformatics*, (19:13), pp. 1636-1643
- Zhang, M. L. and Zhou M. H. 2005. "A k-nearest neighbor based algorithm for multi-label classification", *Proceedings of the 1st International IEEE Conference on Granular Computing*

Appendix

Abbreviation	Description
A1	150 Rv3881 12.5ug pool-10/20/11
A2	148 Rv0934-P38 4ug pool-10/20/11
A3	111 Rv2031c-HSPX 25ug pool-10/20/11
A4	175 Ag85b 6.25ug pool-10/20/11
A5	124 Rv1860-MPT32 25ug pool-10/20/11
A6	178 CFP10 25ug pool-10/20/11
A7	177 ESAT6 6.25ug pool-10/20/11
A8	158 Ag85a 6.25ug pool-10/20/11
A9	113 H37Rv MEM 6.25ug pool-10/20/11
A10	181 HN878 MEM 6.25ug pool-10/20/11
A11	194 CDC1551 MEM 10ug pool-10/20/11
A12	151Bfrb-pak 25ug pool-10/20/11
A13	108 GroES-Pak 25ug pool-10/20/11
A14	110 Rv3507 100ug pool-10/20/11
A15	112 Rv1926c 25ug pool-10/20/11
A16	116 CFP10-ESAT6 25ug pool-10/20/11
A17	117 Rv2878c 25ug pool-10/20/11
A18	120 Rv1099 25ug pool-10/20/11
A19	123 Rv3619 25ug pool-10/20/11
A20	126 Rv1677 100ug pool-10/20/11
A21	128 Rv2220 6.25ug pool-10/20/11
A22	129 Rv2032 25ug pool-10/20/11
A23	131 Rv1984-CFP21 25ug pool-10/20/11
A24	137 Rv3873 25ug pool-10/20/11
A25	139 Rv0054 12.5ug pool-10/20/11
A26	140 Rv3841 25ug pool-10/20/11
A27	142 1566c 25ug pool-10/20/11
A28	143 Rv2875 25ug pool-10/20/11
A29	145 Ag85c-10/20/11
A30	146 Rv1009 6.25ug pool-10/20/11
A31	149 Rv1980-MPT64 25ug pool-10/20/11
A32	152 Rv0831 25ug pool-10/20/11

Table A1: Original antigen names and corresponding tags used in analysis