

# Diagnosing Tuberculosis With a Novel Support Vector Machine-Based Artificial Immune Recognition System

Mahmoud Reza Saybani<sup>1,\*</sup>; Shahaboddin Shamshirband<sup>2</sup>; Shahram Golzari Hormozi<sup>3</sup>; Teh Ying Wah<sup>1</sup>; Saeed Aghabozorgi<sup>1</sup>; Mohamad Amin Pourhoseingholi<sup>4</sup>; Teodora Olariu<sup>5</sup>

<sup>1</sup>Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

<sup>2</sup>Department of Computer Science, Chalous Branch, Islamic Azad University, Chalous, IR Iran

<sup>3</sup>Department of Electrical and Computer Engineering, Faculty of Engineering, University of Hormozgan, Bandar Abbas, IR Iran

<sup>4</sup>Gastroenterology and Liver Diseases Research Center, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran

<sup>5</sup>Department of Intensive Care, Faculty of Medicine, Western University of Arad, Arad, Romania

\*Corresponding Author: Mahmoud Reza Saybani, Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. E-mail: saybani@gmail.com

Received: October 13, 2014; Revised: January 28, 2015; Accepted: February 7, 2015

**Background:** Tuberculosis (TB) is a major global health problem, which has been ranked as the second leading cause of death from an infectious disease worldwide. Diagnosis based on cultured specimens is the reference standard, however results take weeks to process. Scientists are looking for early detection strategies, which remain the cornerstone of tuberculosis control. Consequently there is a need to develop an expert system that helps medical professionals to accurately and quickly diagnose the disease. Artificial Immune Recognition System (AIRS) has been used successfully for diagnosing various diseases. However, little effort has been undertaken to improve its classification accuracy.

**Objectives:** In order to increase the classification accuracy of AIRS, this study introduces a new hybrid system that incorporates a support vector machine into AIRS for diagnosing tuberculosis.

**Patients and Methods:** Patient epacris reports obtained from the Pasteur laboratory of Iran were used as the benchmark data set, with the sample size of 175 (114 positive samples for TB and 60 samples in the negative group). The strategy of this study was to ensure representativeness, thus it was important to have an adequate number of instances for both TB and non-TB cases. The classification performance was measured through 10-fold cross-validation, Root Mean Squared Error (RMSE), sensitivity and specificity, Youden's Index, and Area Under the Curve (AUC). Statistical analysis was done using the Waikato Environment for Knowledge Analysis (WEKA), a machine learning program for windows.

**Results:** With an accuracy of 100%, sensitivity of 100%, specificity of 100%, Youden's Index of 1, Area Under the Curve of 1, and RMSE of 0, the proposed method was able to successfully classify tuberculosis patients.

**Conclusions:** There have been many researches that aimed at diagnosing tuberculosis faster and more accurately. Our results described a model for diagnosing tuberculosis with 100% sensitivity and 100% specificity. This model can be used as an additional tool for experts in medicine to diagnose TBC more accurately and quickly.

**Keywords:** Artificial Intelligence; Classification; Data Mining; Support Vector Machines; Expert Systems

## 1. Background

Tuberculosis (TB) is a major global health problem, which has been ranked as the second leading cause of death from an infectious disease worldwide (1). Diagnosis based on cultured specimens is the reference standard, however results take weeks to process. The bacillus, *Mycobacterium tuberculosis*, causes this infectious disease (2), which affects mainly the lungs (pulmonary TB), however it can affect other body parts as well (extra pulmonary TB) (3). *Mycobacterium tuberculosis* invades the host in small infectious droplets in the lower respiratory tract, rather than in the upper respiratory tract (4). Resistance to *Mycobacterium* TB depends highly on the host's capability to induce an effective T-helper1 immune response (5-8). Depending on the intensity of exposure to the bacteria, the

risk of becoming infected could be high and the infection could last for a long time (9, 10). Typical symptoms of pulmonary tuberculosis are persistent coughs, fever, weight reduction and night sweats (11). Sputum smear microscopy is the most common method for diagnosing TB, yet another method for the diagnosis of TB is the use of rapid molecular tests. Countries with a better developed laboratory capacity use the culture method to diagnose TB (3).

Modern computer technologies have assisted medical doctors to store vast amounts of records. It is obviously impossible for an unassisted human to process this huge amount of data. On the other hand if available medical data (such as patients' laboratory data and knowledge and experience of experts) are not in an easily compre-

hensible format, they become less useful. The human immune system performs like a de-centralized learning system; it learns ways to identify patterns and then uses memory cells to recognize previously identified patterns. These very significant computational features have motivated researchers to create Artificial Immune Systems (AIS). Artificial immune systems were built primarily for feature extraction or data clustering, Watkins (12) introduced a supervised learning algorithm named the Artificial Immune Recognition System (AIRS) and focused on classification problems (13). This was the first version of AIRS that was later dubbed as AIRS1. Performance of AIRS1 when applied on various benchmark classification problems was comparable to other highly-regarded supervised learning techniques for the same benchmarks (14). Moreover, AIRS1 has shown significant achievements on a wide range of classification problems (13), and in medical decision making, classification plays a significant role. Cuevas et al. argued that bio-inspired computing has proven to be useful for classification problems (15).

Brownlee decided to refine the process of AIRS by reducing the complexity of the AIRS approach while maintaining the accuracy of results, and introduced AIRS2 (16). Watkins et al. showed that AIRS2 is simpler and computationally more efficient than AIRS1 and declared AIRS2 as the standard AIRS implementation (14). Performance of AIRS has been the subject of numerous researches (17, 18) and AIRS has been found to have many successful applications in diagnosing various diseases (19-21). Analysis of our literature review revealed that relatively a few efforts have been made to improve the classification accuracy of AIRS; the majority of published articles have applied AIRS to solve a particular classification problem.

The AIRS uses k-Nearest Neighbor (kNN) as a classifier, and in machine learning, kNN was developed in a way that it identifies patterns of data without demanding for an exact match to any cases or stored patterns, which relatively speaking, means low accuracy. In addition, selecting k for kNN affects the performance of kNN, if k is chosen to be too small, the result might be sensitive to noise in data, and if k is too large, then the neighborhood may include points from other classes (22).

## 2. Objectives

This study made the following modification to the process of AIRS2: instead of using kNN as a classifier, this research used the Support Vector Machine (SVM), which is a very robust classifier with many applications. The effects of these changes were analyzed by applying the new model to a real tuberculosis dataset. Obtained classification accuracy was then compared with other classifiers using the same dataset. This new hybrid SVM-based AIRS2 system is dubbed as SAIRS2.

## 3. Patients and Methods

This cross-sectional study was carried out during the

year 2014; it used SVM, which has a root in the advanced statistical learning theory. It is one of the most accurate and robust algorithms among all well-known algorithms; it is insensitive to the number of dimensions, and it needs only a dozen instances for training (22). The SVM is able to maximize the predictive accuracy of a model without over-fitting the training instances. It tries to find an optimal hyper plane in a multi-dimensional space, which separates cases of different classes (23). An ideal hyper plane can be constructed through an iterative training algorithm. In a two-class learning task, for example, SVM makes sure that the best hyper plane is found through maximizing the margin between the two classes. The margin is the shortest distance between the closest data points to a point on the hyper plane (22). In cases when training data cannot be separated without error, SVM needs to separate the training set with a minimal number of errors, which are also known as classification error (24).

The SVM tries to exclude training errors from the training set. The error-free part of the dataset can be separated without errors. Support Vector Machine models use a cost parameter C, in order to allow some flexibility in separating the classes. The cost parameter C creates a soft margin that allows some misclassifications (25); increasing C improves the classification accuracy for the training data; however, this may also lead to over-fitting, because it increases the cost of misclassifying points as well (26). The problem of finding the optimal hyper plane is an optimization problem and Cortes et al. used the saddle point of Lagrangian function to solve this problem (27). This classical optimization problem is solved by standard quadratic programming techniques and programs (27). The complexity of calculations does not depend on the dimension of the input space but on the number of support vectors, which is a small subset of the training vector (28); they assist model interpretation (29). Similarity and dissimilarity of data objects are quantified by kernels, which can be constructed for a broad range of data objects (29).

There are a number of kernels that can be used in SVM models. These include linear, polynomial, radial basis function (RBF) and sigmoid. This research uses RBF, which is the most popular choice of kernel used in SVMs. This is chiefly due to their finite and localized responses across the full range of the real x-axis (30).

The RBF is defined in Equation 1:

$$1) \quad K(X_i, Y_i) = e^{-\gamma|X_i - Y_i|^2}$$

Where  $\gamma$  is a kernel parameter. Increasing the value of  $\gamma$  improves the classification accuracy, yet this can also lead to over fitting (27, 31, 32).

### 3.1. Pseudo Code of SAIRS2

Box 1 illustrates the pseudo code for the new hybrid

algorithm. SAIRS2 uses SVM and the general concept of AIRS2. It creates a random base called memory cell pool (M) and maintains the pool of cells, which are prepared through exposing the system to a one-shot iteration of the training data. As long as the memory cell is inadequately stimulated for a given input pattern, candidate memory cells are prepared. Most stimulated memory cells undergo a process of cloning and mutation. Then the algorithm deals with competition for resources in the development process of a candidate memory cell. The clones compete with each other based on the amount of resources each cell is using and its stimulation value. Competition for resources is necessary to control the size of the Artificial Recognition Ball (ARB) pool, as well as to promote such ARBs that have greater affinity (stimulation) for the antigen that the model is being trained on. The goal was to develop a memory cell that is most successful in accurately classifying a given antigen. Then the potential candidate memory cell is

introduced into the set of already established memory cells, for training. If the memory cell candidate is more stimulated by the training antigen, than the memory cell match, then the memory cell candidate will be added to the set of memory cells. If the affinity between memory cell candidate and memory cell match is less than a threshold, then memory cell candidate replaces memory cell match in the pool of memory cells. The above process repeats until all antigens have been introduced to the system.

The proposed algorithm prepares the content of M for the SVM by reformatting the content to the format used by the SVM. When training for all antigens is completed, this algorithm presents the formatted M to the SVM classifier for classification.

This research used Weka LibSVM (WLSVM) (33), which is a wrapper of Lib SVM (32) inside the Waikato Environment for Knowledge Analysis (WEKA) (34) to address the classification problem.

---

#### Box 1. Pseudo Code for SAIRS2

##### I. Initialize and normalize dataset

##### II. Seed the memory cell pool (M), if desired.

##### III. For each training example (antigen) do the following:

1. If M is empty, add antigen to M.
2. Select the memory cell (mc) in M of the same classification having the highest affinity to antigen.
3. Clone mc in proportion to its affinity to antigen.
4. Mutate each clone and add to the B-cell pool (ARB).
5. Allocate resources to ARB. Remove the weak cells (population control of ARB).
6. Calculate the average stimulation of ARB to antigen and check for termination. If the termination condition is satisfied, go to step 9.
7. Clone and mutate a random selection of B-cells in ARB based upon their stimulation.
8. Loop back to step 5.
9. Select the B-cell in ARB with the highest affinity to antigen (candidate). If candidate has a higher affinity to antigen than mc, add candidate to M. If mc and candidate are sufficiently similar, then remove mc from M. Return M:prepare content of M for SVM

##### IV. Transfer M to format: run SVM as classifier

##### V. Perform SVM classification using M.

---

### 3.2. Real Data Set

This study used a tuberculosis data set, taken from the Pasteur laboratory of Amol, Iran. The data was collected during the year 2014, and ethical approval was obtained from the ethics committee. All equipment at the institute was regularly calibrated, and experts measured the clinical variables. For this study, representativeness was very vital therefore it was important to have adequate numbers of instances for both TB and non-TB cases. The initial data contained some missing values; after consulting with the experts, instances with missing values were either replaced by average values of the people of the same sex and age group, or the entire instance was excluded from data analysis. The dataset consisted of 175 samples. Each sample consisted of twenty-one

features including: sex, Cluster of Differentiation 4 (CD4) counts, Human Immunodeficiency Virus (HIV), Purified Protein Derivative (PPD), chest pain, weight loss, coughs, night sweats, fever, shortness of breath, hemoglobin concentration, platelet count, total White Blood Cell (WBC) count, neutrophil count, lymphocyte count, erythrocyte sedimentation rate, alanine aminotransferase level, alkaline phosphatase level, lactate dehydrogenase concentration, and the status of TB (positive or negative). Overall, 114 of the samples belonged to the positive class (people who had TB), and remaining 60 samples belonged to the negative class. Table 1 illustrates average values, and standard deviation of numerical features of the data set.

### 3.3. Applied Parameters

Both AIRS2 and SAIRS2 come with user-defined parameters that allow users to control and fine tune these classifiers; their functions are explained below:

**Artificial Recognition Ball (ARB):** represents a number of identical B-Cells and is also a mechanism to reduce duplication and dictate survival within the population. **Affinity Threshold Scalar (ATS):** is a value between 0 and 1 that when multiplied by the affinity threshold, it provides a cut-off value for memory cell replacement during the training phase. **Clonal rate:** is an integer value that is used to determine the number of mutated clones a given ARB is allowed to attempt to produce. **Hyper-mutation rate:** is an integer value that is used to determine the number of mutated clones that a given memory cell is allowed to inject into the cell population. **Seed cell:** repre-

sents an antibody that is drawn from the training set and is used to initialize memory cell and ARB populations at the beginning of training. **Stimulation value:** is the value returned by the stimulation function; this function is used to measure the response of an ARB to an antigen or to another ARB. The function returns a value between 0 and 1. **Total resources:** is a parameter, which sets limitations on the number of ARBs allowed in the system. Each ARB is allocated a number of resources based on its stimulation value and the clonal rate (12).

Table 2 shows parameter settings and their values that were used to run AIRS2 and SAIRS2 algorithms. Here,  $k$  is a user defined variable that is used for the KNN algorithm;  $\gamma$ ,  $C$  and cash memory size are also user defined variables that are used to adjust the SVM.

**Table 1.** Average Values of Numerical Properties

Feature	Average Value	Standard Deviation
Alanine aminotransferase, U/L	140.04	117.62
Albumin concentration, g/dL	3.17	2.82
Alkaline phosphatase, U/L	493.44	495.33
Erythrocyte sedimentation, mm/h	50.36	40.36
Hemoglobin concentration, g/dL	11.33	2.82
Lactate dehydrogenase concentration, U/L	1091.17	780.15
Lymphocyte count, lymphocytes $\times 10^9/L$	1.66	1.25
Neutrophil count, neutrophils $\times 10^9/L$	8.62	6.08
Platelet count, platelets $\times 10^9/L$	394.57	201.84
Total WBC count, cells $\times 10^9/L$	9.84	5.49

**Table 2.** Parameters Used for AIRS2 and SAIRS2 Algorithms

Parameters	AIRS2	SAIRS2
Affinity threshold scalar ATS	0.2	0.2
Clonal rate	10.0	10.0
Hyper-mutation rate	2.0	2.0
Seed cell	1	1
Stimulation value	0.9	0.5
Total resources	150	150
$k$ of kNN	3	n/a
SVM Type	n/a	V-SVC
Kernel function	n/a	RBF
$\gamma$	n/a	1
$C^a$	n/a	7
Cash memory size	n/a	100 MB

<sup>a</sup>  $C$ , is the parameter for the soft margin cost function, which controls the influence of each individual support vector.

This study used the Waikato Environment for Knowledge Analysis (WEKA); a machine learning software (34) to develop the SAIRS2 classifier. In order to evaluate SAIRS2's performance, this paper chose classifiers within the WEKA and applied them against the same TB data set. Chosen classifiers were: a) AIRS2, b) Multi-Layer Perceptron (MLP), c) Naïve Bayes, d) J48 (C4.5), e) k-Nearest Neighbor (KNN), f) Radial Basis Function Network (RBF Network), g) Learning Vector Quantization (LVQ), h) Hierarchical LVQ, i) SPEGASOS, j) Random Forest, k) Ridor, l) LibLinear, m) RandomTree, n) Multipass LVQ, o) CLONALG-CSCA: Clonal Selection Classification Algorithm (CSCA) v1.0, p) CLONALG v1.0: Clonal selection algorithm, q) SMO, r) ZeroR and s) LibSVM (34).

### 3.4. Restrictions of the Research

This study ran the algorithm only on one computer with Intel core 7 CPU and 8 GB RAM. It is of interest to verify the performance with other computing resources, since the model uses the computer memory to keep the prepared memory cells. A further restriction was that we could not find real data for TB from other medical centers to verify the model.

### 3.5. Performance and Evaluation Measurements

In order to evaluate the performance of the proposed algorithm and compare it with that of benchmark algorithms, this study used classification accuracy, n-fold cross validation, sensitivity, specificity and root mean-squared error.

### 3.6. Classification Accuracy

The accuracy was calculated by dividing the number of truly classified instances by the number of instances in the test phase. The classification accuracy for the dataset was measured according to Equations 2 and 3 (12):

$$2) \quad \text{Accuracy}(T) = \frac{\sum_{i=1}^T \text{asset}(t_i)}{T}, t_i \in T$$

$$3) \quad \text{asset}(t) = \begin{cases} 1, & \text{if } \text{classify}(t) = t.c \\ 0, & \text{otherwise} \end{cases}$$

Where T is the set of data items to be classified (the test set),  $t \in T$ , t.c is the class of the item t, and classify(t) returns the classification of t by AIRS.

### 3.7. N-Fold Cross Validation

In this approach, the instances were randomly divided to equal stratified subsets. Each instance is placed in one subset. At each iteration, n-1 subsets were merged to form the training set and the classification accuracy of the algorithm was measured on the remaining subset. This process was repeated n times, choosing a different sub-

set as the test set each time. Therefore, all data instances were used n-1 times for training and once for testing. The final predictive accuracy was computed over all folds in the usual manner by dividing the number of correct classifications taken over all folds by the number of data instances in all folds. This study used 10-fold cross-validation for evaluation purposes.

### 3.8. Sensitivity and Specificity

Sensitivity refers to the proportion of people with disease who have a positive test result, and specificity refers to the portion of people without disease who have a negative test result (35).

### 3.9. Root Mean-Squared Error

Root mean-squared error is calculated by Equation 4:

$$4) \quad \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$$

Where  $p_1, \dots, p_n$  are predicted values on the test instances, and  $a_1, \dots, a_n$  are actual values (35).

### 3.10. Area Under the Curve (AUC)

Receiver Operating Characteristic (ROC) originates from the Signal Detection Theory (SDT) that was developed during the Second World War for the analysis of radar images. Radar operators needed to know whether the blip on the screen represented a friendly ship, an enemy, or just noise. Signal Detection Theory measures the ability of radar receiver operators to distinct these three important signs. This ability of the operators was called the Receiver Operating Characteristic (36). Receiver Operating Characteristic is a useful visualizing technique that is used for assessing data mining schemes; it is used to illustrate the trade-off between true hit rate and false alarm over a noisy channel (35). Graphs of ROC are able to provide a richer measure of classification performance than error cost, accuracy or error rate (37). Curves of ROC illustrate the performance of a classifier regardless of the error costs or class distribution; the vertical axis is used to plot the true positive rate and the horizontal axis represents the true negative rate (35). Equations 5 and 6 show the formula for True Positive Rate (TP Rate) and False Positive Rate (FP Rate) (35):

$$5) \quad \text{TP Rate} = 100 * \left( \frac{\text{TP}}{\text{TP} + \text{FN}} \right)$$

$$6) \quad \text{FP Rate} = 100 * \left( \frac{\text{FP}}{\text{FP} + \text{TN}} \right)$$

Where TP is the number of true positives, FN is the num-



ber of false negatives, FP is the number of false positives and TN is the number of true negatives. TP and TN are correct classifications; FP is the number of outcomes when the prediction is positive when it is actually negative, and FN is the number of outcomes when the prediction is negative when it is actually positive (35).

The area under the ROC curve (AUC) is a single quantity that summarizes ROC curves. This area denotes the probability that a classifier assigns a higher score to the positive instance than a randomly chosen negative sample (38). The greater the area, the better the model (35). A rough guide for classifying the accuracy of a test is the common academic grading scale (36): Excellent (A) = 0.90 to 1; Good (B) = 0.80 to 0.90; Fair (C) = 0.70 to 0.80; Poor (D) = 0.60 to 0.70; Fail (F) = 0.50 to 0.60.

This means for example, if the area under ROC falls within 0.9 and 1 the classifier performs excellent and if it falls within 0.5 and 0.6, it fails.

### 3.11. Youden's Index J

One way to summarize the performance of a diagnostic test is to use the Youden's index J, which is a value in the range of zero and one. While a value of one indicates that the test is perfect without any false positives or false negatives, a value of zero indicates that the test is worthless. This index is calculated through Equation 7.

$$7) \quad J = \text{Sensitivity} + \text{Specificity} - 1$$

J is also equivalent to the AUC (39). This index value is considered by some researchers to be optimal in that sense that both specificity and sensitivity are given the same weight to their errors, thus maximizes the overall rate of correct classification in the absence of a loss function (40).

## 4. Results

### 4.1. Comparing Classification Accuracy of SAIRS2 With Other Classifiers

Classification accuracies, sensitivities, specificities, Youden's Index (J Index), Area Under ROC Curve (AUC), and root mean-squared errors (RMSE) obtained with SAIRS2 and the above mentioned classifiers are illustrated in Table 3. This study used 10-fold cross-validation for evaluation purposes. As it can be seen, the performance of SAIRS2 is comparable to the top classifiers used in this study. It can be concluded that SAIRS2 with a classification accuracy of 100%, a sensitivity of 100%, a specificity of 100%, J Index of 1, AUC of 1, and a RMSE of 0 can be used to classify tuberculosis. SAIRS2 ranked excellent (A) among other classifiers in this study.

**Table 3.** Comparing Classification Accuracy of SAIRS2 With Other Classifiers of Tuberculosis<sup>a</sup>

Method	Classification Accuracy, %	Sensitivity, %	Specificity, %	J Index	AUC	RMSE
SAIRS2	100.00	100.00	100.00	1	1	0
AIRS2	100.00	100.00	100.00	1	1	0
MLP	100.00	100.00	100.00	1	1	0.003
Naive Bayes	100.00	100.00	100.00	1	1	0
J48	100.00	100.00	100.00	1	1	0
KNN, K=7	100.00	100.00	100.00	1	1	0
RBF Classifier	100.00	100.00	100.00	1	1	0.054
Hierarchal LVQ	94.83	92.11	100.00	0.96	0.96	0.227
Spegasos	99.82	100.00	99.50	1	0.99	0.006
Random Forest	99.43	100.00	100.00	1	0.98	0.025
Ridor	99.43	99.11	100.00	1	0.99	0.024
Lib Linear	98.33	97.44	100.00	1	0.96	0.067
Random Tree	98.21	98.7	97.30	0.98	0.98	0.066
LVQ with kNN	97.13	92.11	100.00	0.96	0.96	0.227
Multi pass LVQ	95.98	93.86	100.00	0.97	0.96	0.2
LVQ	94.83	92.11	100.00	0.96	0.96	0.227
CLOANALG-CSCA	93.68	90.35	100.00	0.95	0.95	0.251
CLONALG	92.52	92.27	93.00	0.93	0.92	0.217
SMO	89.66	91.59	98.31	0.89	1	0.221
Zero R	65.49	100.00	0.00	0.5	0.5	0.475
Lib SVM	65.49	100.00	0.00	0.5	0.5	0.227

<sup>a</sup> Abbreviations: AUC, Area Under ROC Curve; CLONALG, The CLONal selection ALGORITHM; CSCA, The Clonal Selection Classification Algorithm; kNN, k-Nearest Neighbor; LVQ, Learning Vector Quantization; MLP, Multi-Layer Perceptron; SMO, Sequential Minimal Optimization; RMSE, root mean-squared errors.

## 5. Discussion

There are different methods for determining whether or not someone has TB, such as culturing, sputum smear examination and molecular tests, however these methods take a long time to process results, need special equipment and are expensive. In special cases such as diagnosing TB in children, smear and culture method have been reported to have less than 50% accuracy and researchers are increasingly looking for new ways of rapid and accurate diagnosis (41). Furthermore, researchers have become more interested in studying new aspects of tuberculosis identification, including the use of artificial intelligence (21). Early and accurate identification of TB helps prevent, treat and control this disease. Timely and accurate diagnosis leads to timely treatment, which in turn reduces the risk of transmission to others. In addition smear microscopy has shown low sensitivity and versatility (42).

One of the most important decision-making tools in medicine is classification (43), and data classification problems are common in many fields including medicine (19). Huang et al. indicated that for saving patients' lives, classification accuracy of medical data sets has to be optimal (20).

Researchers have used only a few AIRS-based classifiers for diagnosing various diseases, and all of them have reported that their methods improved classification accuracy. Because of the need for more accurate classifiers, this study modified AIRS2 in such a way that the kNN was replaced by SVM. A real medical dataset, the tuberculosis dataset from the Pasteur laboratory of Iran, was used in this study.

For the same data set used in this study, other researchers obtained a classification accuracy of 99.14% with sensitivity and specificity of 87% and 86.12%, respectively; their model was named Fuzzy AIRS (21).

Currently many people are infected by and die from tuberculosis, which has impacts not only on their life but also on their family members and the society as a whole. There have been many researches on diagnosis of tuberculosis, yet we still need to do more to diagnose this disease faster and more accurately. Applications of computer science in medicine has entered a period of rapid discovery, because the technology allows us to find patterns in historical data and gain knowledge that can help experts diagnose diseases such as tuberculosis faster and more accurately.

Our study described a model for diagnosing tuberculosis, and revealed that SAIRS2 with an accuracy of 100%, sensitivity of 100% and specificity of 100%, Youden's Index equal to one, AUC equal to one, and RMSE equal to zero for TB classification is very well comparable with other top techniques evaluated in this study. Therefore it can be concluded that SAIRS2 is a highly sensitive and specific tool that can be used as an additional tool by experts in medicine to diagnose tuberculosis more accurately and quickly. SAIRS2 is an improvement of AIRS2 in terms

of classification accuracy for the tuberculosis data set. SAIRS2 is open to other application areas.

Although, it might be reasonable to say that revisions of AIRS2 provide great accuracy, this cannot be blindly accepted as a characteristic of SAIRS2. As with all learning algorithms, the domain to be learned has a greater influence on the performance of the algorithm than any other factor, and authors recommend further studies and simulations of TB cases.

## Acknowledgements

The authors appreciate the Pasteur laboratory for providing the data.

## Authors' Contributions

Study concept, design, analysis and interpretation of data: Mahmoud Reza Saybani, Shahaboddin Shamshirband, Shahram Golzari Hormozi, Teh Ying Wah and Saeed Aghabozorgi. Drafting of the manuscript: Mahmoud Reza Saybani, Shahaboddin Shamshirband, Mohammad Amin Pourhoseingholi. Critical revision of the manuscript for important intellectual content: Mahmoud Reza Saybani, Shahaboddin Shamshirband and Mohammad Amin Pourhoseingholi.

## References

- Jeyaseelan L, Williams D, Tibrewal S, Ali SA, Hassan M, Vemulapalli K. Tuberculosis of the Cuboid: A Case Report and Review of the Literature. *J Foot Ankle Surg.* 2014.
- Bird L. Infectious disease: The tuberculosis signature. *Nat Rev Immunol.* 2010;**10**(10):677.
- WHO. World Health Organization publications on tuberculosis Global tuberculosis report 2013. In: WHO, editor. Geneva: 2013.
- Kugelberg E. Immune evasion: Mycobacteria hide from TLRs. *Nat Rev Immunol.* 2014;**14**(2):62-3.
- Allie N, Grivennikov SI, Keeton R, Hsu NJ, Bourigault ML, Court N, et al. Prominent role for T cell-derived tumour necrosis factor for sustained control of Mycobacterium tuberculosis infection. *Sci Rep.* 2013;**3**:1809.
- Silva-Gomes S, Appelberg R, Larsen R, Soares MP, Gomes MS. Heme catabolism by heme oxygenase-1 confers host resistance to Mycobacterium infection. *Infect Immun.* 2013;**81**(7):2536-45.
- Carow B, Reuschl AK, Gavier-Widen D, Jenkins BJ, Ernst M, Yoshimura A, et al. Critical and independent role for SOCS3 in either myeloid or T cells in resistance to Mycobacterium tuberculosis. *PLoS Pathog.* 2013;**9**(7).
- Jayaraman P, Sada-Ovalle I, Nishimura T, Anderson AC, Kuchroo VK, Remold HG, et al. IL-1beta promotes antimicrobial immunity in macrophages by regulating TNFR signaling and caspase-3 activation. *J Immunol.* 2013;**190**(8):4196-204.
- Chiang CY, Van Weezenbeek C, Mori T, Enarson DA. Challenges to the global control of tuberculosis. *Respirology.* 2013;**18**(4):596-604.
- Lienhardt C, Glaziou P, Uplekar M, Lonnroth K, Getahun H, Raviglione M. Global tuberculosis control: lessons learnt and future prospects. *Nat Rev Microbiol.* 2012;**10**(6):407-16.
- Phillips M, Basa-Dalay V, Bothamley G, Cataneo RN, Lam PK, Natividad MP, et al. Breath biomarkers of active pulmonary tuberculosis. *Tuberculosis (Edinb).* 2010;**90**(2):145-51.
- Watkins AB. . AIRS: A resource limited artificial immune classifier. Starkville: Mississippi State University; 2001.
- Brownlee J. Artificial immune recognition system (airs)-a re-

- view and analysis. 2005.
14. Watkins A, Boggess LC. . A new classifier based on resource limited artificial immune systems. *Proceedings of Congress on Evolutionary Computation, Part of the 2002 IEEE World Congress on Computational Intelligence held in Honolulu*. May 12-17 2002 HI, USA. IEEE: 2002..
  15. Cuevas E, Osuna-Enciso V, Zaldivar D, Perez-Cisneros M, Sossa H. Multithreshold segmentation based on artificial immune systems. *Math Probl Engin*. 2012;**2012**(20).
  16. Brownlee J. *Clever algorithms: nature-inspired programming recipes*. Melbourne: Lulu Enterprises; 2011.
  17. Golzari S, Doraisamy S, Sulaiman MN, Udzir NI, editors. Improving the accuracy of AIRS by incorporating real world tournament selection in resource competition phase.; *Evolutionary Computation, 2009. CEC'09. IEEE Congress on.; 2009; Trondheim*. IEEE; pp. 3040-4.
  18. Le X, Mo-Yuen C, editors. Distribution fault diagnosis using a hybrid algorithm of fuzzy classification and artificial immune systems.; *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*.; 2008; Pittsburgh. pp. 1-6.
  19. Tunc T. A new hybrid method logistic regression and feedforward neural network for lung cancer data. *Math Probl Engin*. 2012;**2012**(1).
  20. Huang YH, Ko YC, Lu HC. An Optimal Classification Method for Biological and Medical Data. *Math Probl Engin*. 2012;**2012**(17).
  21. Shamshirband S, Hessam S, Javidnia H, Amiribesheli M, Vahdat S, Petkovic D, et al. Tuberculosis disease diagnosis using artificial immune recognition system. *Int J Med Sci*. 2014;**11**(5):508-14.
  22. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl Inf Syst*. 2008;**14**(1):1-37.
  23. Rahimi AA, Bahmani MK, Masoudnejad A, Saadati M, Khosravi A, Shafi'khani A, et al. Application of near-infrared spectroscopy and support vector machine in detection of HIV-1 infection. *Trauma Mon*. 2010;**15**(1):23-9.
  24. Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, et al. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS One*. 2011;**6**(10).
  25. Koutsouleris N, Gaser C, Bottlender R, Davatzikos C, Decker P, Jager M, et al. Use of neuroanatomical pattern regression to predict the structural brain dynamics of vulnerability and transition to psychosis. *Schizophr Res*. 2010;**123**(2-3):175-87.
  26. DTREG.. *Support Vector Machines*. 2012. Available from: <http://www.dtreg.com/svm.htm>.
  27. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;**20**(3):273-97.
  28. Auria L, Moro RA. Support vector machines (SVM) as a technique for solvency analysis. *Ger Inst Econ Res*. 2008;**1**(1).
  29. Campbell C, Ying Y. *Learning with Support Vector Machines*. Bristol: Morgan & Claypool; 2011.
  30. Statsoft.. *Support Vector Machines*. Melbourne: Statsoft; 2013. Available from: <http://www.statsoft.com/textbook/support-vector-machines/>.
  31. Statnikov A. Theory and methods. Theory and methods. In: Hackensack NJ, editor. A gentle introduction to support vector machines in biomedicine. Vol. 1. Singapore: World Scientific; 2011.
  32. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;**2**(3):27.
  33. EL-Manzalawy Y, Honavar V. *WLSVM: Integrating LibSVM into Weka Environment, 2005*. 2005. Available from: <http://www.cs.iastate.edu>.
  34. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explor newsl*. 2009;**11**(1):10-8.
  35. Witten IH, Frank E. *Data Mining: Practical machine learning tools and techniques*. 3rd ed San Francisco: Morgan Kaufmann; 2011.
  36. Tape TG. *Interpreting Diagnosis Test: University of Nebraska Medical Center*. University of Nebraska Medical Center: Nebraska; 2013. Available from: <http://gim.unmc.edu/dxtests/roc3.htm>.
  37. Fawcett T. An introduction to ROC analysis. *Pattern recogn lett*. 2006;**27**(8):861-74.
  38. Berrar DP. An Empirical Evaluation of Ranking Measures With Respect to Robustness to Noise. *J Artif Intell Res*. 2014;**49**(1):241-67.
  39. Powers D. Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation. *J Mach Learn Technol*. 2011;**2**(1):37-63.
  40. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*. 2005;**16**(1):73-81.
  41. Graham SM, Ahmed T, Amanullah F, Browning R, Cardenas V, Casenghi M, et al. Evaluation of tuberculosis diagnostics in children: 1. Proposed clinical case definitions for classification of intrathoracic tuberculosis disease. Consensus from an expert panel. *J Infect Dis*. 2012;**205** Suppl 2:S199-208.
  42. Amin I, Idrees M, Awan Z, Shahid M, Afzal S, Hussain A. PCR could be a method of choice for identification of both pulmonary and extra-pulmonary tuberculosis. *BMC Res Notes*. 2011;**4**:332.
  43. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. *Expert Syst Appl*. 2008;**35**(1):82-9.