Tuberculosis and Lung Cancer Prediction using Machine Learning Methods and Over-Sampling Technique

Amani YAHYAOUI
Department of Software Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
amani.yahyaoui@izu.edu.tr

Amir Karaj
Department of Software Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
amirkaraj02@gmail.com

Merve Hamzaoğlu
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
mervehamzaogluu@gmail.com

Akhtar Jamil
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-2592-1039

Nejat YUMUŞAK
Department of Computer Engineering
SAKARYA University
Sakarya, Turkey
nyumusak@sakarya.edu.tr

***Abstract*—With the continuous advancement of technology, people and machines can complement their specific skills to achieve effective results. In this sense, with the inevitable increase in the number of diseases that threaten human health, Decision Support Systems (DSS) are widely used in the medical field to help doctors making better clinical decisions. Among these diseases, such as tuberculosis and lung cancer are considered potentially serious infectious and are among the top 10 causes of death in the world. This paper presents a medical DSS for tuberculosis and lung cancer diagnosis by using machine learning algorithms, such as the Support Vector Machines (SVM) and Artificial Neural Network (ANN). Moreover, Borderline Synthetic Minority Over-Sampling Technique (Borderline - SMOTE) was also employed to increase the number of minor sample size. The experimental dataset used is taken from Diyarbakir chest diseases hospital. The obtained results proved the efficiency of the proposed system in helping doctors making the right decision and improving the quality of health care.**

Keywords—Decision Support Systems (DSS), chest diseases, tuberculosis, lung cancer, machine learning, Support Vector Machines, Artificial Neural Network.

I. INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence which is able to create intelligent machines that think like humans and make some human tasks and decisions [1]. The machine learning techniques are used in many research area such as industry, medicine, banking, statistics to name just a few.

Lungs are vital organs in the human body and vulnerable by many diseases, namely tuberculosis and lung cancer. In fact, Tuberculosis is a contagious infectious disease that, primarily attacks the lungs and can damage other organs such as the brain [1]. This disease is caused by bacteria called Mycobacterium tuberculosis that transmits the disease from person to person by air [1]. According to the World Health Organization, tuberculosis is one of the ten leading causes of death worldwide. Indeed, it caused the deaths of 1.3 million people in 2017 [2]. Tuberculosis develops slowly and can be manifested by several symptoms like fever with night sweats, cough sometimes with a few bloodstreams, shortness of breath, pains in the chest, a state of fatigue, loss of appetite, weight loss, headache, presence of large ganglia [3].

Another disease that attack lings is the lung cancer, which is the most dangerous of all cancer types [4] that initially starts in the lungs and, then can spread throughout the human body. It consists of uncontrolled cell division in the lungs. Doctors confirm that smoking is the principal cause of lung cancer but also passive tobacco exposure can also cause lung cancer for nonsmokers [4]. According to the World Health Organization last statistics, lung cancer caused the deaths of 2.09 million people in 2018 [5]. Unlike other cancers, the lung cancer symptoms appear when the disease is in an advanced stage. Among these symptoms, the appetite loss, the voice changes, the frequent chest infections such as bronchitis or pneumonia, the breath shortness, unexplained headaches, the weight loss, the wheezing [6] Lung cancer treatments depend on three main factors, namely the tumor location, its stage and the person health state. Generally, surgery, radiation and chemotherapy are the main lung cancer treatment [6]. In medicine, doctors confirm that the earlier the disease is discovered, the greater the recovery chances. The different tuberculosis and lung cancer symptoms mentioned in this first section help doctors to make the correct diagnosis of these diseases and to start the necessary treatment. Despite this, the mortality rate in worldwide of patients suffering from chest disease is continually increasing. This situation can be explained by the possible diagnosis errors, doctor's competence, the continuous appearance of new complicated diseases or the lack of means that can help doctors to make the right decisions.

As a solution, in today's technologically advanced world, researchers are focusing on the use of artificial intelligence techniques to help doctors making the right decision in medical diagnosis on the right time. Among these techniques, ANN and SVM are used in the present paper due to their popularity. To increase the samples, Borderline-SMOTE was also used to generate new samples. The main

objective was to remove the imbalance issue and increase the performance of the classifier.

This paper is organized as fellow: Section 2 presents some previous works that focus on lung cancer and tuberculosis diagnosis by using some of machine learning techniques. Section 3 describes the machine learning algorithms used in this paper. The dataset, the proposed methods, the application developed and the obtained results used are detailed in Section 4. Finally, the conclusions are summarized in section 5.

II. Related Work

In the literature, many researchers have proposed various methods for lung diseases diagnosis. For instance, Udayakumar E. et. al [7] have designed an automated approach for predicting tuberculosis by using the SVM method. The authors employed two datasets have been used from Shenzhen Hospital, China and from Tuberculosis clinic of Montgomery County (USA). As result, the SVM has shown good performance by giving 82% as classification accuracy [7].

In addition, Rehana Rajan and K. G. Satheesh Kumar [8] have proposed a hybrid classification system composed by two classification methods, which are the SVM and the Multi-layer Perceptron (MLP) to identify tuberculosis. The results have shown that the hybrid system is able to identify tuberculosis with an accuracy rate of 83.42% for the SVM method and with an accuracy rate of 74.61% for the MLP method [8].

P.JohnVivek, and Swathika.S.R [9] have presented a method for tuberculosis identification by using the most used classifiers in disease identification, which are the K-Nearest Neighbors KNN , the Binary-SVM and the Multi-class SVM. The Multi-class SVM gave the best results 10 with an accuracy of 92%, followed by the Binary SVM with 89% and 75% by the KNN [9].

Gayatri S Mahajan and Dr. S. R. Ganorkar [10] have proposed a new method based on SVM for the tuberculosis detection in chest radiography. In this research paper, authors have included image processing field to extract feature from chest images, have applied the SVM method to classify the extracted features, than checked whether the patient is affected with tuberculosis or not. In this research paper, the authors have shown that the proposed system can efficiently verify the presence or not of tuberculosis with an accuracy of 88% [10].

Wafaa Alakwaa, Mohammad Nassef and Amr Badr have suggested a computer- aided diagnosis (CAD) system for lung cancer diagnosis by using Neural Network method [11]. The dataset used in this research was taken from Kaggles Data Science Bowl (DSB) and composed with 1397 patients computed tomography (CT) scan. The classification accuracy from this research has reached 86 %[11].

Moreover, in [12], Raviprakash S. Shriwas proposed a system for lung cancer prediction in its earliest stage by using Artificial Neural Network ML technique. The performance of the proposed system was very good and has reached an accuracy of 96%.

In addition, in [13], Abdelwadood M. Mesleh have proposed a Computer-Aided Design (CAD) hybrid system that allow to detect the lung cancer by using three different algorithms which are Multi-Layer (ML), Neural Networks (NNs) and the Independent Component Analysis (ICA). The performance of the proposed system has achieved an accuracy of 91%. Based on this literature review, it can be said that SVM and ANN algorithms have good performance in the medical field and specially in lung cancer and tuberculosis diagnosis.

III. Materials and Method

A. Data Set

The dataset used in our research was taken from Diyarbakir chest diseases hospital. The dataset 27 include 250 real record from patients distributed as follows: 100 patient suffering from 28 tuberculosis, 100 patient suffering from lung cancer and 50 healthy patients. The dataset include the most 38 relevant attributes that can help doctors to easily identify the disease. For some attributes, the maximum and the minimum values are given to indicates the interval that can be accepted for non-infected patient. For example, the value of White Blood Cell (WBC) must be between 4-11. For other attributes, the value can be 0 which indicate the non-existence of the attribute and 1 mean the existence of the attribute, for 36 example the smoking addiction attributes is 0 or 1.

B. SVM

SVM is basically a binary classification method, developed by Cortes and Vapnik in 1995 [14]. The main idea of SVM is to classify the membership of an entry into one of two classes separated by a hyperplane (Fig. 1). If the data are linearly separable, the hyperplane separates the two classes by maximizing the minimum distance between the data and the hyperplane.

However, the majority of classification tasks are not linearly separable. To apply the SVM on such tasks, the data is first transformed into a higher dimensional space in which the data is supposed be linearly separable by using a kernel function. There are various kernel function, such as the radial basic kernel, polynomial kernel, linear kernel etc. Among these kernel functions radial basis and polynomial kernel are widely used. For more detail about the SVM refer to [15] and [16].

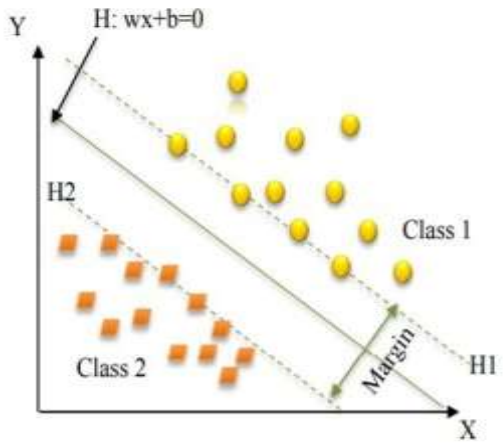

Fig. 1. Hyperplane construction using SVM method for classification [17]

C. ANN

Motivated from the working of the human brain, the ANN algorithms tries to mimic the similar behavior of the human brain. The first mathematical modeling tests of the human brain by the notion of formal neurons was carried out by W. M. McCulloch and W. Pit in 1943[18]. This concept was then developed into a full-fledged neural network method known as perceptron system in 1957 by Franck Rosenblatt, which consisted of a set of artificial neurons arranged in layers and interconnected by synaptic weight [19]. Such architecture of the neurons learns the patterns in the data similar to the way the human brain works.

As shown in Fig. 2, the perceptron is organized in three layers: the input, hidden and output layers. The input layer is a set of neurons carrying the input signal. The hidden layer(s) tries to extract the hidden relations between the variables. The number hidden layers and the number of neurons on these layers can be varied. The output layer consists of the neurons according to the number of the output classes for the target problem.

The working of the neural network system is based on two phases: the training phase and the operation phase [19]. In the first phase, based on the knowledge extracted from the learning data, the synaptic weights of each neuron will be adapted to solve a particular problem. Once the training phase is completed, the ANN system produces the results based on knowledge gained from the training phase.

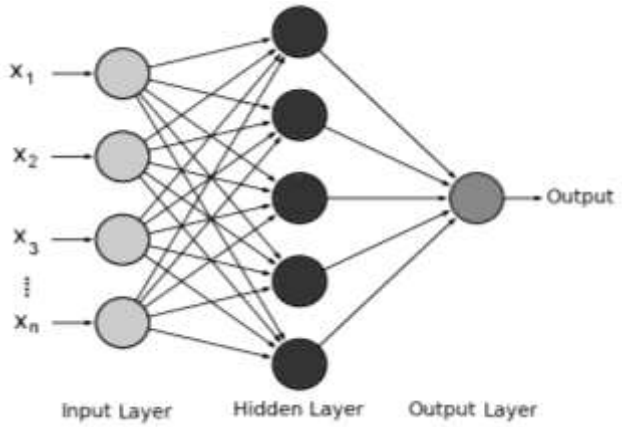

Fig. 2. Multilayer Perceptron organization in ANN [18]

D. Classification

In our study, two classifiers which are the SVM and the ANN have been investigated. The overall workflow of the proposed method is shown in Fig. 3. Any supervised classifier requires training and testing data. Therefore, first we divided the data into test and train datasets. In fact, the train data represent two-thirds of the total data, and the test data represent one-third of the total data. Furthermore, analysis was performed to replace any missing attributes or invalid values. In such case, the attributes were replaced with the mean value of all the attributes in each set. Moreover, the data is labelled as shown in Table I.

Finally, both classifiers were trained using the training data. For each classifier, we employed K-fold cross-validation to use all the available data for training. The trained models were saved and then feed then with the test data. They model produced labels for all three possible classes which were then compared with real ground truth for evaluations.

TABLE-I CLASSES AND LABELS

Case	Abbreviation	Label
Tuberculosis	TB	1
Lung Cancer	LC	2
Normal	NR	3

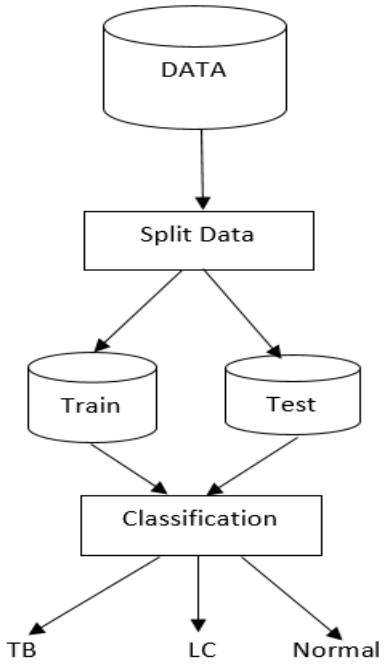

Fig. 3. Tuberculosis and lung cancer diagnosis using ANN and SVM

IV. RESULTS

Several experiments were performed to evaluate the performance of the used classifiers. The following sections provide a detailed description.

A. Experimental Setup

Since, the supervised classifiers depend on several parameters, obtaining the optimal values for these parameters is crucial as they effect the classification accuracy of the classifier. Therefore, before applying the model, the optimal parameters for each classifier were obtained by applying an exhaustive grid search. The parameters that produced highest validation accuracy were used for final classification. In this study, SVM with radial basis function was used. It requires tuning of two most important parameters, the cost (C) and gamma (γ). The optimal values for these parameters were obtained by searching in a predefined range: $C \in \{ e^{-5} - e^{5} \}$ while $\gamma \in \{e^{-3} - e^{0}\}$. The final values obtained for C was 275 while γ was 0.15. Similarly, for ANN both momentum and learning rate were fine-tuned. Learning rate was empirically set to 0.1 and momentum was set to 0.7. Moreover, the maximum number of iterations were 1000 and two hidden layers each with 200 neurons were used.

The method were conducted on Intel® Xeon® CPU E3-1231 with 2.40GHz processing power and 32GB RAM, using with Matlab © Environment.

B. Evaluation

Performance of the proposed methods were evaluated using the overall accuracy, precision, recall and f-measure metrics. These metrics were derived using True Positive (TP) which is the correct predictions, True Negative (TN) which is correctly predicted for wrong event, False Positive (FP) is incorrectly predicted and False Negative (FN) correctly predicted for wrong event. These measures were obtained using following formulas:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The obtained results for SVM and ANN classifiers are summarized in Table II and Table III respectively. As shown that both classifiers produced good classification accuracy for each class. The classification results obtained by using the ANN model is relatively higher than the SVM for each class. The classification results for tuberculosis disease by using ANN is 97.68% while the SVM is 94.11%. The classification results for lung cancer disease by using ANN is 96.51% while the SVM presented 83.37%. Similarly, for normal cases both ANN and SVM produced same result (97.68%).

SVM produced the lowest accuracy for the LC class which achieved highest accuracy for normal patient class. While ANN produced consistently similar results for all three classes. In terms of processing time, SVM took 20 seconds for training while ANN took 68 minutes for training. The testing time was same for both and was negligible therefore, we did not account for the testing time.

TABLE- II CLASSIFICATION ACCURACY (%) FOR SVM CLASSIFIER

Class	OA	PR	RC	FM
TB	94.11	94.65	91.36	92.97
LC	83.37	85.36	84.3	84.82
NR	97.68	98.36	96.36	97.34
Mean	91.72	92.79	90.73	91.71

*OA: Overall Accuracy, PR: Precision, RC: Recall, FM: f-score

TABLE- III CLASSIFICATION ACCURACY (%) FOR ANN CLASSIFIER

Class	OA	PR	RC	FM
TB	97.68	98.36	95.77	97.04
LC	96.51	94.65	95.05	94.84
NR	97.68	99.02	96.36	97.67
Mean	97.29	97.34	95.72	96.52

*OA: Overall Accuracy, PR: Precision, RC: Recall, FM: f-score

V. CONCLUSION

In this paper, a comparative analysis was performed using SVM and ANN classifiers for detection and diagnosis of tuberculosis and lung cancer diseases. The experimental results indicated that both techniques were effective to detect these diseases with high accuracy. These can be incorporated in a decision make system that can help the doctors to diagnosis tuberculosis and lung cancer with higher accuracy.

Our results are in line with the state-of-the-art method for lungs cancer and tuberculosis detection. However, in future, we would like to further enhance the accuracy of the proposed method by integrating a deep learning-based approach such as convolutional neural networks.

REFERENCES

[1] A. T, S. Natarajan, and K. N. B. Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classificatio,"
no. December 2014, 2011.

[2] World health Organization, global tuberculosis report. 2018.

[3] A. Yahiaoui, O. Er, and N. Yumusak, "A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines, " Biomed. Res., vol. 28, no. 9, pp. 4208–4212, 2017.

[4] M. A. Hussain, T. M. Ansari, P. S. Gawas, and N. N. Chowdhury, "Lung Cancer Detection Using Artificial Neural Network & Fuzzy Clustering,
" Ijarcce, vol. 4, no. 3, pp. 360–363, 2015.

[5] A. Yahyaoui and N. Yumuşak, "Decision support system based on the support
vector machines and the adaptive support vector machines algorithm for solving chest
disease 46 diagnosis problems," Biomed. Res., vol. 29, no. 7, pp. 1474–1480, 2018.

[6] U. E., S. S., and V. P., "TB screening using SVM and CBC techniques," Curr.
Pediatr. 48 Res., vol. 21, no. 2, pp. 338–342, 2017.

[7] K. Patel, Brijeshkumar; Chavda, "Hybrid SVM for Automatic Detection of 50
Tuberculosis," Int. J. Adv. Res. 1Computer Sci. Manag. Stud., vol. 3, no. 11, pp. 44–53,
2013.

[8] M. P. JohnVivek and S. S.R, "Accurate TB manifestation using multi class SVM
54 classifier," Iarjset, vol. 2, no. 1, pp. 37–44, 2015.

[9] G. S. Mahajan, "Detection of Tuberculosis Using Chest Cardiograph," Int. J. Innov. Res.
Sci. Eng. Technol., vol. 6, no. 5, pp. 9858–9865, 2017

[10] W. Alakwaa, M. Nassef, and A. Badr, "Lung Cancer Detection and Classification with
3D Convolutional Neural Network (3D-CNN)," Int. J. Adv. Comput. Sci. Appl., vol. 8,
no. 8, 2017.

[11] R. S. Shriwas and A. D. Dikondawar, "Lung Cancer Detection and Prediction By Using 1
Neural," IPASJ Int. J. Electron. Commun., vol. 3, no. 1, pp. 17–21, 2015.

[12] A. M. Mesleh, "Lung cancer detection using multi-layer neural networks with 3 independent component analysis: A comparative study of training algorithms," Jordan
4 J. Biol. Sci., vol. 10, no. 4, pp. 239–249, 2017.

[13] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and Applications," in 7 Advanced Course on Artificial Intelligence, Springer,Berlin, Heidelberg., 2001, pp. 249–257.

[14] A. Kowalczyk, Support Vector Machines succınctly. 2017.

[15] C. Z. Deng, Naiyang, Yingjie Tian, Support vector machines: optimization based theory, algorithms, and extensions. 2012.

[16] L. H. Lee, R. Rajkumar, L. H. Lo, C. H. Wan, and D. Isa, "Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach," Expert Syst. Appl., vol. 40, no. 6, pp. 1925–1934, 2013.

[17] R. M. Cesar and L. da Fontoura Costa, An introduction to neural networks. CRC press, 2014.

[18] K. Suzuki, Artificial Neural Networks – architectures and applications, no. August. 21 2013.