International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

COPD Classification using Machine Learning Algorithms

Siddhi Vora[1], Prof. Chintan Shah[2]

[1]*M.E. student, Dept. of Bio-medical Engineering, Government Engineering College, Gujarat, India*
[2]*Asst. Professor, Dept. of Bio-medical Engineering, Government Engineering College, Gujarat, India*

Abstract - *Several ailments compromise human wellbeing by influencing life span and its prosperity from in many ways. Among them, Lung Diseases as Chronic Obstructive Pulmonary Disease (COPD), lung cancer, pneumonia, and asthma are considered as genuine wellbeing difficulties and one noteworthy reason for death in both developed and developing nations. Specialists affirm that the prior an illness is analyzed and classified, the higher is the chances of patient cure. In this situation, Artificial Intelligence Algorithm and Expert System have been effectively used to take care of various issues in different areas including medical field, which have benefits like shortening of the diagnosis and classification period, time gain and expanded proficiency. Thusly, Artificial Neural Networks appear to have an effective utilization in COPD classification. This permits a framework that would assist the doctor with determining COPD level all the more quickly with high performance. In this paper, we utilize the Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) strategy to arrange COPD severity level. These methodologies are assessed utilizing a test dataset from Prajna Health Care medical Hospital. The exploratory outcomes demonstrated the effectiveness of these strategies.*

Key Words: Chronic Obstructive Pulmonary Disease (COPD), Machine Learning Algorithms, Support Vector Machine (SVM), K Nearest Neighbour (KNN), COPD Classification

1. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is an obstructive type of lung disease. COPD is a common outcome for subjects exhibiting suffering from emphysema or chronic bronchitis. In the condition of Emphysema the alveoli at the end of the bronchioles (smallest air passages of the lungs) are destroyed, typically from smoking. Daily cough and mucus production is the main symptom of chronic bronchitis, which lasts at least for three months a year. Shortness of breath, chest tightness, wheezing, chronic cough are the major signs of COPD disease.

COPD is now become the 4th leading cause of death worldwide and according to WHO report it is estimated to be the 3rd leading cause of death by 2030[1]. Due to environmental changes and increased use of tobacco, rate of disease progression is increased. Cure is not possible for COPD, but the properly maintained treatment and management strategies can slower the progression of COPD.

COPD patient may also experience the condition of exacerbations, during which COPD symptoms become worse. The exacerbations typically occur in response to an environmental insult encountered by the subject. A variety of biological and non-biological environmental insults may cause exacerbations in COPD patients including second hand cigarette smoke, fumes from gasoline, bacteria, and viral infections.

For healthy persons, breathing is done by a minimal effort of the human body. Respiratory system of the body is very sensitive to the various agents that can cause pulmonary diseases. The most caused respiratory diseases are asthma, chronic bronchitis, Chronic obstructive pulmonary disease (COPD) and lung cancer.

COPD is a lung disease portrayed by incessant and intermittent airflow limitation, which builds restriction of airflow [1]. The main cause of COPD is smoking [2]. The two fundamental instances of them are chronic bronchitis and emphysema. About 75% of COPD patients don't have done diagnosis, the vast majority of them are in mild stage (Stage 1), yet additionally 4% in severe (Stage 3) and 1% in very severe (Stage 4) stage of COPD [1]. The reason behind that are slow growth factor of indications as chronic cough and reduced exercise tolerance. It is estimated that, by 2020 the COPD will become the fourth leading cause of death worldwide [3].

For diagnosis of COPD, the most commonly used Pulmonary Function Tests are Impulse Oscillometry System (IOS) and Spirometry. Spirometry is more popular and use for diagnosing COPD and Asthma. Spirometry is based on measurement of forced maneuver.

The forced oscillation technique (FOT) imposes little gaseous tension bothers on the subject's regular breathing for the quantification of lung's mechanical parameters. It also estimates the respiratory impedance utilizing short beats of gaseous tension [4][5].

The best analysis is accomplished by blend of Machine Learning Algorithms and spirometry. Along these lines we acquire a general evaluation of the patient. So as to obtain patient's powerful approval of this work.

The genuine advantage of the machine learning is to automate the analysis of classification of COPD. Information that are entered to programming are identified with a catalog, indications and hazard factors drawn from two noteworthy accords for COPD, Global Initiative for perpetual Obstructive Lung Disease (GOLD). The parameters of spirometry, data about the side effects, hypersensitivities and auscultation of the patient, are incorporated into the machine learning framework, so as to assist the product with suggesting appropriate order of COPD.

The main objective of this research article is to automates the process of COPD classification with high performance and faster response.

2. MATERIALS AND METHODS

In this section, a comparative study between the SVM, KNN, DT and the RF is performed to classify COPD.

2.1 Support Vector Machine (SVM)

SVM is one AI procedure for solving regression and classification issues, presented by Vapnik and Cortes in 1990 for classification of binary dataset[6][7][8][9]. Today, it is utilized in a various research zones, for example, medicinal conclusion [10], speaker recognition , face recognition [11], and so on.

Truth be told, as appeared in Figure 1, SVM comprises in building the most extreme edge hyperplane that ideally isolates two classes of a datasets D=where and the class name of . Nonetheless, numerous hyperplanes can isolate binary classes. In this manner, Support Vector Machine utilizes a preparation stage to locate the ideal hyperplane called Optimal Separating Hyperplane (OSH) which satisfies, where w is a n-dimensional vector and b is a "Bias" term that isolates classes and expands the margin between that classes [11].

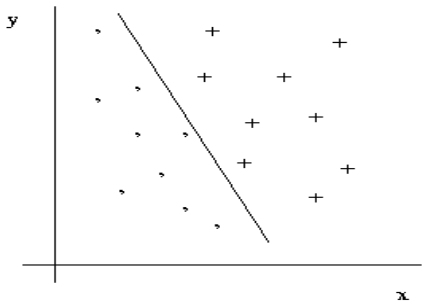

Figure 1 SVM for Linear Kernel

Henceforth, in the above figure, the ideal H hyperplane confirms that the H1 and H2 hyperplanes are parallel to H and go through the nearest points to it which are known as Support Vectors (SV) [21]. Consequently, SVMs pick the ideal hyperplane that augments the margin between the two classes which is the separation somewhere in the range of H1 and H2. Hyperplane associated with the maximized margin, is characterized by [2/||w||]. Scientifically, the ideal hyperplane associated with a most extreme margin can be found by limiting equation (1) under the constraint of Equation (2) utilizing Quadratic Programming enhancement issue as it is clarified as pursues [12].

$$Min_{w,b}\ 1/2\left||w|\right|^2 \rightarrow (1)$$
$$y_i(w^T x_i + b) \geq 1, i = a, \dots, m \rightarrow (2)$$

The above issue is illuminated by utilizing the strategy for Lagrangian multipliers communicated in Equation 3 as follows

$$Q\ (w, b, \propto) = \frac{1}{2} w^T\ w - \sum_{i=1}^{m} \propto_i \{y_i(w^T x_i + \ b) - 1\}$$
$$\downarrow (3)$$

Where αi are the Lagrange multiplier. Optimum of objective function Q is gotten by limiting it with respect to b and w and by augmenting objective function as for αi utilizing Equation 4.

$$\frac{\partial Q}{\partial w} = 0 \quad , \quad \frac{\partial Q}{\partial b} = 0 \rightarrow (4)$$

Using Equation 4 we deduce:

$$\begin{cases} w = \sum_{i=1}^{m} \propto_i y_i x_i \\ \sum_{i=1}^{m} \propto_i y_i = 0 \end{cases} \rightarrow (5)$$

Thus, substituting Equations 4 and 5 in Equation 3, the accompanying double issue to maximize is acquired.

$$\begin{cases} \text{Maximizing } Q(\propto) = \sum_{i=1}^{m} \propto_i - \sum_{i=1}^{m}\sum_{j=1}^{m} \propto_i \propto_j y_i y_j x_j \\ \text{Subjected to } \sum_{i=1}^{m} \propto_i y_i = 0 \\ \qquad\qquad \propto_i \geq 0 \end{cases} \qquad (6)$$

The issue of classification of another unknown sample x is illuminated then by the below mentioned Equation 7.

$$\begin{cases} H(x) = \sum_S \propto_i y_i x^T x_i + b \\ b = y_i - w^T x_i \end{cases} \rightarrow (7)$$

Thus, the decision of class can be done as follows:
- If H (x)=0 then x is not classifiable
- If H (x) <0 then x= ε -1
- If H (x)>0 then x= ε +1

Here the linear separation of information isn't possible such as the case in Figure 2, the thought is to delineate non-direct space in another direct higherdimensional space, where the linear separation of training set is

possible with a kernel capacity communicated as follows in Equation 8.

$$k(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j) \rightarrow (8)$$

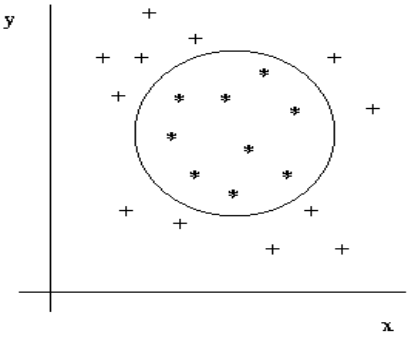

Figure 2 SVM for Non-linear case

Without a doubt, the calculation of transformation space is performed utilizing a "mapping capacity" $F = (\emptyset(x)|x \in X)$ which is certainly characterized by SVM portion's decision. The new space is classified "Feature space"(Figure 3) [12].

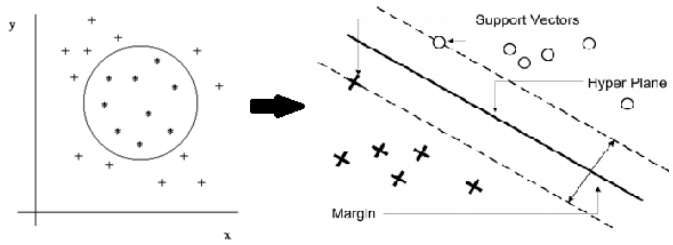

Figure 3 Mapping Space Concept

Numerous part capacities have been utilized while applying SVM. Among the most used capacities, we will specify polynomial, sigmoid , direct, RBF kernels [12].

2.2 COPD Classification using SVM

The decision of kernel of SVM is as yet a not fully solved issue and shows a confinement of execution of SVM. For the most part, to pick the correct piece, analysts need to do the tests. In our investigation, the linear kernel was utilized, the other kernel functions will be tried and performed in our subsequent work. The SVM structure utilized in our examination is appeared in Figure 4. Consequently, the information will be first separated into two gatherings of train and test utilizing the cross-approval as an information apportioning technique. The technique of cross-approval utilized is the hold-out strategy; it comprises of partitioning the information into two examples: the first is for training and the second is for the test. The model is based on the training test and approved on the test. Subsequently, amid the arrangement task, SVM starts via preparing the train information given with their classes' factors, to develop the reference display. The test information are then arranged by anticipating their classes dependent on the model previously acquired. As an outcome, patients with class esteems "1" has a place with the individual influenced by Class – 1 (Mild), while patients with class esteems "2" has a place with the individual influenced by Class – 2 (Moderate). Also, class esteems "3" has a place with the patients influenced by Class – 3 (Severe), while class esteems "4" has a place with the patients influenced by Class – 4 (Very Severe).

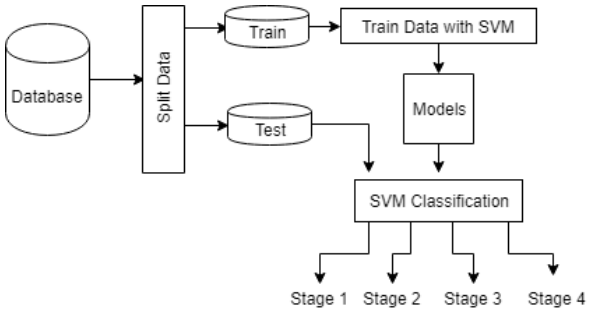

Figure 4 COPD Classification system using SVM

2.3 K Nearest Neighbour (KNN)

The KNN calculation utilizes classification of neighboring pixel as the predictive estimation of the new test. On the basis of distance traveled nearer neighbors are calculated. The KNN strategy calculation is basic, works dependent on the most short distance from the sample used in training to decide the KNN.

Subsequent to gathering KNN, at that point most of KNNs were taken to be anticipated from the test. The information for the KNN calculation comprises of some multi-variate qualities.

$$d_1 = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \rightarrow (9)$$

Where:X_1= sample data
X_2 = data uji
I = variable data
P = dimension of data

For the classification of Y, Xi will be utilized. Information from K-NN can be on any size scale, from ordinal to ostensible [13]. The benefits of K-NN have a few preferences, that it is flexible to preparing effective and noisy information if training dataset is vast. While the drawbacks of KNN are:

- KNN needs to decide the estimation of the parameter K (number of closest neighbors).
- Training dependent on distance isn't clear about what sort of distance to utilize and which credits ought to be utilized to get the best outcomes.

The search for similitude between the new case and the old case is accomplished by comparing the signs entered by the user as per the signs in the knowledge base. This recovery procedure will utilize the K-Nearest Neighbor technique.

K-Nearest Neighbor (K-NN) known as one of the most straightforward nonparametric classifiers however in the high-dimensional precision settings, KNN is influenced by obstruction highlights. In this examination, K-Nearest Neighbor is imperative as another way to deal with multi-class classification in high-dimensional issues [14].

The K-Nearest Neighbor calculation (k-closest neighbor or K-NN) is a calculation for arranging objects dependent on learning information that is nearest to the object. An uncommon situation where grouping is anticipated dependent on the nearest learning information (at the end of the day, k = 1) is known as a nearest neighbor Algorithm [15].

The reason for this algorithm is to characterize new objects dependent on the qualities and training of sample. Grouping does not utilize any model to coordinated and just dependent on memory. Given a test point, various K objects (preparing focuses) would discovered nearest to the test point. Characterization utilizes the most votes among groupings of K objects. The KNN calculation utilizes contiguousness order as the prescient estimation of the new test. Close or far neighbors generally determined dependent on Eucledian separate. The KNN strategy calculation is extremely basic, taking a shot at the most limited separation from the test to the preparation test to decide the KNN.

3. CONCLUSIONS

According to the results of testing on prediction of classification of severity level of COPD using the Support Vector Machine and K-Nearest algorithm, conclusions are obtained as follows:

1. The after effects of the forecasts completed by the Support Vector Machine calculation delivered a high accuracy of 96.97% and along these lines had the capacity to classify precisely COPD severity level.

2. Based on the classification result of COPD severity level using the algorithm of K-Nearest Neighbour, obtained accuracy is 92.30%, which is low compared to the accuracy obtained using SVM algorithm.

ACKNOWLEDGEMENT

I heartily thanks to Prof. Chintan Shah for his support in preparation of this review article.

REFERENCES

[1] S. Anakal and P. Sandhya, "Clinical Decision Support System for Chronic Obstructive Pulmonary Disease using Machine Learning Techniques," *Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech.*, pp. 0–4, 2017.

[2] K. N. M. Rao, "Diagnosis and Management of Chronic Cough due to Extrapulmonary Etiologies," vol. 25, no. 5, pp. 437–443, 2014.

[3] "WORLD HEALTH STATISTICS." WORLD HEALTH ORGANIZATION, 2018.

[4] K. Karuppanan, A. S. Vairasundaram, and M. Sigamani, "COPD prognosis under biologically inspired neural network," *Proc. - 2012 Int. Conf. Adv. Comput. Commun. ICACC 2012*, no. 1, pp. 22–26, 2012.

[5] Andrew Bremer and A. Dias, "Low-cost, Open-source Spirometer," 2009.

[6] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.

[7] S. Kosinov and I. Titov, "Large margin multiple hyperplane classification for content-based multimedia retrieval," no. 4, pp. 1–3.

[8] E. Zarrouk and Y. Benayed, "Hybrid SVM / HMM Model for the Arab Phonemes Recognition," vol. 13, no. 5, pp. 574–582, 2016.

[9] S. S. Mehta and N. S. Lingayat, "Support Vector Machine for Cardiac Beat Detection in Single Lead Electrocardiogram," no. May, 2007.

[10] S. Bhatia, P. Prakash, and G. N. Pillai, "SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features," 2008.

[11] P. Rico, "Training Support Vector Machines : an Application to Face Detection," no. June, 1997.

[12] E. Sup *et al.*, "Thèse Utilisation des méthodes Support Vector Machine (SVM) dans l ' analyse des bases de données," 2012.

[13] P. Ociepka, K. Herbus, O. N. Kuzyakov, I. N. Glukhikh, and A. E. Sidorova, "A Case-Based Reasoning Method with Rank Aggregation A Case-Based Reasoning Method with Rank Aggregation," 2018.

[14] H. R. Shahraki, S. Pourahmad, and N. Zare, "*K* Important Neighbors : A Novel Approach to Binary Classification in High Dimensional Data," vol. 2017, 2017.

[15] K. Klinis and K. M. Rahim, "PENERAPAN ALGORITMA k-NN (nearest Neighbor) UNTUK DETEKSI PENYAKIT (KANKER SERVIKS) Novita Mariana, Rara Sriartati Redjeki, Jeffri Alfa Razaq Abstrak," vol. 7, no. 1, pp. 26–34, 2015.