

Detection of Different stages of COPD Patients Using Machine Learning Techniques

Ali Hussain*, Ikromjanov Kobiljon Komil Ugli*, Beom Su Kim*, Minji Kim*, Harin Ryu* Satyabrata Aich*, Hee-Cheol Kim**

*Institute of Digital Anti-Aging Healthcare, Inje University, South Korea

** Department of Computer Engineering/Institute of Digital Anti-Aging Healthcare/u-HARC, Inje University, South Korea

alihussain.dream@gmail.com, kobiljonikromjanov@gmail.com, beomsu91@naver.com, mj0616@hanmail.net, Christine-ryu@hanmail.net, satyabrataaich@gmail.com, heeki@inje.ac.kr

Abstract: In recent years, there are an increase in the mortality rate due to Chronic obstructive pulmonary disease (COPD) patients and it is estimated that it will increase in the coming years. Traditional methods take a long time to identify these diseases because a lot of clinical tests has to be performed for getting the confirmation, however with the advent of intelligent techniques as well as looking at the potential of the powerful techniques for predicting other critical diseases, it is believed that it would help to detect the chronic diseases at an early stage in a precise manner. In this paper, an attempt has been made to detect COPD patients and at the same time, it could distinguish the stages such as the early stage of chronic obstructive pulmonary disease patients (ESCP) and the Advanced stage of COPD patients (ASCP). We have used the Recursive Feature Elimination, Cross-Validated (RFECV) feature selection method to select features and then we consult the doctors, those are expert in the field to recommend the features among the features selected using RFECV method. The features selected using the doctor recommendation called features reduction with doctor recommendation (FRDR). After selecting two groups of features, we have used different machine learning algorithms to compare the performance of the algorithms as well as the importance of the features. It was found that the features selected using the RFECV method could able to provide accuracy of 96%, whereas the features selected with doctor recommendation name FRDR could able to provide accuracy of 90%. Although there is a difference of result in both the methods but overall, both set of features produces a good result. So, it is recommended that this approach would help distinguish different stages in real-life situations.

Keywords: Machine Learning, Chronic Obstructive Pulmonary Disease (COPD), Classification, Features Selection, Performance comparison.

1. INTRODUCTION

The percentage of the affected people with Chronic Obstructive Pulmonary disease is increasing compared to the past year, and it is expected that it will increase at a higher rate because the growth rate of our population is increasing day by day. The death from Chronic Obstructive Pulmonary Disease (COPD) is increasing if we do not take urgent action to reduce

it then it will become the third leading cause of death in the world by 2030 [1]. 3 million people were died because of COPD in 2012, and this number is equivalent to 6% of the total death in that year [2].

According to the world economic Forum (WEF) in 2030, the cost associated with chronic diseases could reach 47 trillion globally. Diagnosis of exacerbation of chronic obstructive pulmonary disease (COPD) is considered to be a big challenge and the most difficult part of the medical field. In this paper, we have proposed a method that can distinguish the early stage of COPD(ESCP) and the advanced stage of COPD patients (ASCP). There is no cure only therapy can slow its infection, reduced its complexity, and improve quality of life [3], and avoid exacerbation. This minor contribution seeks to predict exacerbation. Ordinarily, the exacerbation is defined as a damaged lung function, acute event, or abrupt worsening of COPD symptoms likely to cause death [4]. Moreover, exacerbation may contribute to damage to the quality of life [5]. Hence, timely detection of exacerbation can reduce its effects, facilitate lung recovery [6]. Looking at the past work as well as the need for proper methods for detecting COPD, in this paper, we have attempted to detect and distinguish different stages of COPD.

The data collected from the Inje University Paik Hospital. Total of 2900 patients which includes early-stage patients as well as advanced-stage patients. The COPD data sets are divided into early-stage COPD patients (ESCP) and Advanced Stage COPD Patients (ASCP). Features selection is a crucial part of any medical field. the data set contains a certain amount of redundant or irrelevant features that can reduce our classifier performance. We have used the Recursive Feature Elimination, Cross-Validated (RFECV) feature selection method to select features and then we consult the doctors, those are expert in the field to recommend the features among the features selected using RFECV method. The features selected using the doctor recommendation called features reduction with doctor recommendation (FRDR). After selecting two groups of features, we have used different machine learning algorithms to compare the performance of the algorithms as well as the importance of the features.

The paper structure is arranged as follows: Section II describes the related work. Section III presents the

methodology used for this research work. Section IV describes the result and discussion of the present work. Section V presents the conclusion and future work.

2. RELATED WORKS

Several researchers in the past attempted and tried to use different machine learning algorithms for analysing the condition of COPD patients, and they used different feature selection methods to point to the utmost important features. Some related works are mentioned below.

Siddhi and Chintan used SVM (Support Vector Machine) and KNN (K-Nearest Neighbour) for classification. The perfect kernel choice is not yet completely solved the issue. They observed that using different kernels but the linear kernel in SVM good for classification problems in the medical field. They have found accuracy 96.97% with SVM and KNN accuracy 92.30% [7] and also find here in the medical field SVM is good for classification problems [8] SVM works well the classification by creating a multidimensional hyperplane that optimally split the data into two different categories. These models are similar to neural networks. The sigmoid kernel function used by the SVM model is a neural network with a two-layer perceptron [9].

In Chronic Obstructive Pulmonary disease, Random Forest classifier performs well for classification problems. They observe that in the medical field better perform, and they have got 97.7% accuracy with random forest classifier [10].

Jin and Zhang found Linear discriminant analysis (LDA) showed the best precision of 62% concerning another classifier [11]. LDA is used because of its simplicity and it is a good classification method for labelled data [12].

A decision tree that uses variables commonly gathered by physicians can provide a hurried assessment of the rigor of the disease. Decision and regression tree based on simply predict mortality in patients with stable chronic obstructive pulmonary disease (COPD) [13] Prognosis of the acute exacerbation of chronic obstructive pulmonary disease used C5.0 decision tree classifier to predict the acute exacerbation of COPD patients. They got 80.3% accuracy with 28 features [14].

XGBoost decision tree algorithm for classification shows good performance with two different datasets with high accuracy, sensitivity, specificity, precision, and they observed it a suitable model for this work, and also provide the best performance on unseen data with accuracy and precision, sensitivity, specificity, and precision of the XGBoost classifier a good model for COPD classifications using edge devices [15,16].

3. METHODOLOGY

In this paper, we have collected data from Inje University Paik Hospital, Busan, S. Korea. This data includes two groups, one belongs to (ESCP) and the other is (ASCP). Early Stage of chronic obstructive pulmonary disease (ESCP) patients belong to one class, and Advanced Stage of obstructive pulmonary disease (ASCP) patients belong to the second class. Due to the imbalanced dataset, we need to balance the numbers of classes,

for this purpose we used (SMOTE) algorithm for the over-sampling technique for balanced the training dataset.

Xia et al. have also used the synthetic minority over-sampling technique (SMOTE) algorithm was used to upsample the data in a training dataset. They used SMOTE algorithm for COPD patients to balance the dataset [17]. We have used RFECV with a Random forest algorithm to select relevant features. Originally the data contained 144 features. By using RFECV feature selection techniques the feature set has been reduced to 71 features shown in fig 1, and the Doctor recommended (FRDR) 23 features. Then we have used different classification models on these two different feature sets such as feature selection using RFECV method and feature selection with doctor recommendation (FRDR). Finally, a performance comparison has been made between the RFECV set of features as well as the FRDR features. we have compared the performance based on accuracy, precision, recall, f1-score.

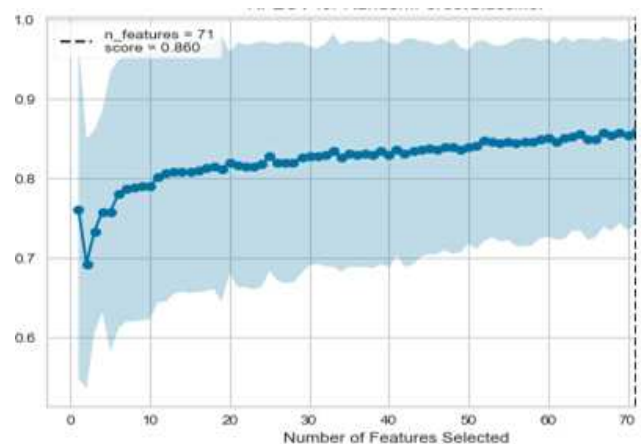


Figure 1. Selected features obtained using RFECV method

3.1. Feature Selection Algorithm and Classifiers

It is one kind of wrapper method that eliminates the weak and unnecessary features, and removing those features have a minimum effect on the performance and retains the strong and independent features for improving the performance of the model. It's an iterative process and keeps updated on the importance of the features in each iteration [18].

The random forest also belongs to the decision tree family. It is known due to high classification accuracy, random forest nature is non-parametric suitable for various types of data analysis and perform classification, survival analysis, regression, etc [19].

Ploner et al. have used gradient boosting classifier for stability selection on health insurance claims data to identify disease trajectories in chronic obstructive pulmonary disease. This approach for automatic detection of disease trajectories in claim data and could help diagnose diseases early identify unknown risk factors and optimize treatment plans [20].

Chang et al. proposed an approach that can predict hypertension outcomes using machine learning techniques with different feature selection techniques such as RFE and RFECV. They have done the combination of RFECV feature selection

techniques with different classifiers such as SVM C4.5, RF, XGBoost, and it was found that the combination would be able to provide good performance [21].

3.2. Performance Measure Metrics

The parameter used to compare the performance and validation of the classifier are as follows: accuracy, Precision, recall, f1-score. The precision is the ratio of the number of true positives to the total number of truly positive and false positive. The recall is the ratio of the number of true positives to the total number of true positive and false negative. F1-score is defined as the weighted average of precision and recall, therefore this score takes both false positive and false negative. Accuracy is defined as the ratio of the number of correct predictions made to the total prediction made and the ratio is multiplied by 100 to make it in term percentage.

4. RESULTS AND DISCUSSION

Python programming language is used for the development of the code. The feature selection technique as well as doctor's recommendation has been used for the reduction of features. REFCV technique has been used for feature reduction since this technique is quite popular nowadays and also doctor recommended features are used for real-time performance check. The ratio of train data to test data is 80:20. The features from two groups fed into different classifiers such as RF, SVM, Decision Tree, Gradient Boost, k-NN, XGBoost, LDA. The performance metrics consists of Accuracy, precision, recall, and F1-Score has been compared using different classifiers with two feature sets such as REFCV feature set as well as FRDR feature set and the results are shown in Figure 2,3,4 and 5. In this paper we have found that SVM performed well on both the feature sets. To validate our result, we have compared our result with few states of-art models, which performed well on healthcare data. The state of art models are as follows: Aich et al., performed similar procedure on Parkinson disease voice datasets combining different classifiers such as SVM, RF, boosted 5.0 and few others with two different feature sets and found that SVM performed better among other classifiers [22]. Aich et al., performed similar procedure on Parkinson's disease gait dataset combining different classifiers with different feature sets and found that SVM performed better among other classifiers [23].

4.1. Comparison of Accuracy

Figure 2 shows that SVM performed better among other classifiers with both the feature sets. SVM with REFCV feature set could provide 96% accuracy and SVM with FRDR set could provide 90%.

4.2. Comparison of Precision

Figure 3 shows that SVM performed better among other classifiers with both the feature sets. SVM with REFCV feature set could provide precision of 96% and SVM with FRDR set could provide precision of 89%.

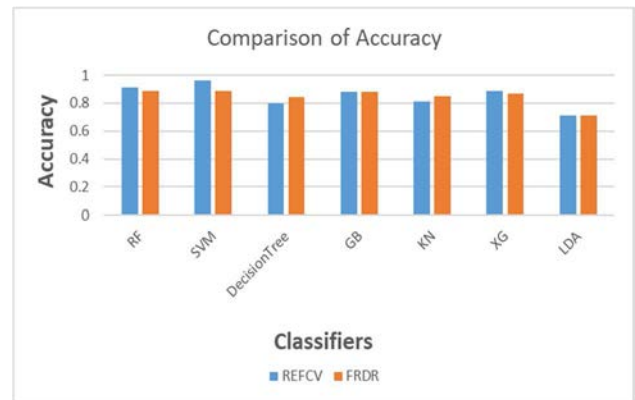


Figure 2. Accuracy of different classifiers with REFCV and FRDR feature sets

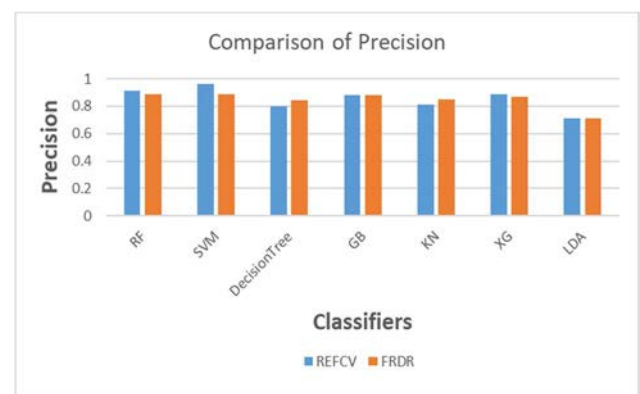


Figure 3. Precision of different classifiers with REFCV and FRDR feature sets

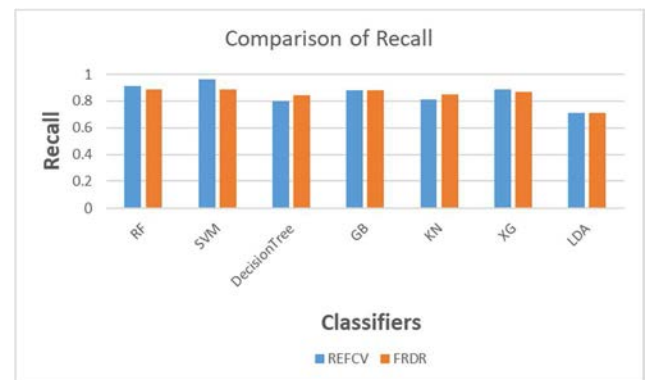


Figure 4. Recall of different classifiers with REFCV and FRDR feature sets

4.3. Comparison of Recall

Figure 4 shows that SVM performed better among other classifiers with both the feature sets. SVM with REFCV feature

set could provide recall of 96% and SVM with FRDR set could provide recall of 89%.

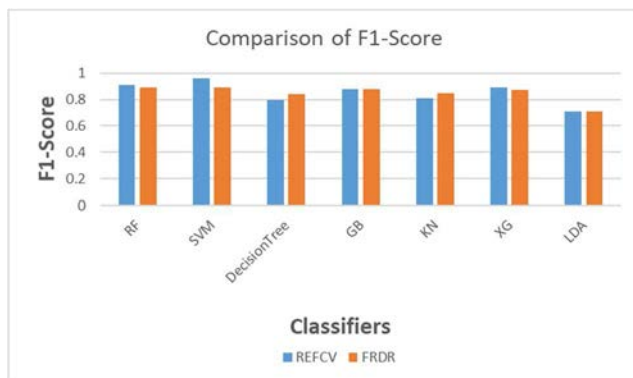


Figure 5. F1-Score of different classifiers with REFCV and FRDR feature sets

4.4. Comparison of F1-score

Figure 5 shows that SVM performed better among other classifiers with both the feature sets. SVM with REFCV feature set could provide F1-score of 0.96 and SVM with FRDR set could provide precision of 0.89.

5. CONCLUSION AND FUTURE WORK

This paper explores different machine learning classifiers with different feature sets for comparing the performance. Especially this paper proposed a result which is very much practical because we took the features based on the doctor recommendation and compared with widely used feature section techniques. We could say that we have done validation of the result and that could be useful for real life implementation. The result shows that SVM classifier with REFCV provided accuracy of 96% and SVM classifier with doctor recommended feature sets provided accuracy of 90%. From the result it is recommended that this model could be tested in real time to know the effectiveness and in the future work we will collect more data and add more features and try deep learning models to improve the performance metrics.

ACKNOWLEDGMENT

This research was supported by the Ministry of Trade, Industry, and Energy (MOTIE), KOREA, through the Education Program for Creative and Industrial Convergence (Grant Number N0000717)

REFERENCES

- [1] <https://www.who.int/respiratory/copd/burden/en/> extracted on 18/11/2020.
- [2] [https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-\(copd\)](https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)) extracted on 18/11/2020.
- [3] Saleh, L., Mcheick, H., Ajami, H., Mili, H. and Dargham, J., 2017, August. Comparison of Machine Learning Algorithms to Increase Prediction Accuracy of COPD Domain. In International Conference on Smart Homes and Health Telematics (pp. 247-254). Springer, Cham.
- [4] Lareau, S., Moseson, E. and Slatore, C.G., 2014. Patient information series: Exacerbation of COPD. American journal of respiratory and critical care medicine, 189(6).
- [5] Maiolo, C., Mohamed, E.I., Fiorani, C.M. and De Lorenzo, A., 2003. Home telemonitoring for patients with severe respiratory illness: the Italian experience. Journal of Telemedicine and Telecare, 9(2), pp.67-71.
- [6] Wilkinson, T.M., Donaldson, G.C., Hurst, J.R., Seemungal, T.A. and Wedzicha, J.A., 2004. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine, 169(12), pp.1298-1303.
- [7] Vora, S. and Shah, C., 2019. COPD Classification using Machine Learning Algorithms.
- [8] Aich, S., Choi, K.W., Pradhan, P.M., Park, J. and Kim, H.C., 2018. A Performance Comparison Based on Machine Learning Approaches to Distinguish Parkinson's Disease from Alzheimer Disease Using Spatiotemporal Gait signals. Advanced Science Letters, 24(3), pp.2058-2062.
- [9] Huan, N.J. and Palaniappan, R., 2004. Neural network classification of autoregressive features from electroencephalogram signals for brain-computer interface design. Journal of neural engineering, 1(3), p.142..
- [10] Spathis, D. and Vlamos, P., 2019. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health informatics journal, 25(3), pp.811-827.
- [11] Jin, Y., Zhang, T., Cao, Z., Zhao, N., Chen, C., Wang, D., Lei, K.C., Leng, D. and Zhang, X.D., 2018, December. Prediction Indicators for Acute Exacerbations of Chronic Obstructive Pulmonary Disease By Combining Non-linear analyses and Machine. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2515-2521). IEEE..
- [12] Fergani, B., 2012, May. Evaluating C-SVM, CRF and LDA classification for daily activity recognition. In 2012 International Conference on Multimedia Computing and Systems (pp. 272-277). IEEE.
- [13] Esteban, C., Arostegui, I., Moraza, J., Aburto, M., Quintana, J.M., Pérez-Izquierdo, J., Aizpiri, S. and Capelastegui, A., 2011. Development of a decision tree to assess the severity and prognosis of stable COPD. European Respiratory Journal, 38(6), pp.1294-1300 .
- [14] Peng, J., Chen, C., Zhou, M., Xie, X., Zhou, Y. and Luo, C.H., 2020. A Machine-learning Approach to forecast Aggravation Risk in patients with Acute exacerbation of chronic obstructive pulmonary Disease with clinical indicators. Scientific reports, 10(1), pp.1-9.
- [15] Zarrin, P.S., Roeckendorf, N. and Wenger, C., 2020. In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools. IEEE Access, 8, pp.168053-168060.
- [16] Zhu, W., Zeng, N. and Wang, N., 2010. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, 19, p.67.
- [17] Xia, J., Sun, L., Xu, S., Xiang, Q., Zhao, J., Xiong, W., Xu, Y. and Chu, S., 2020. A Model Using Support Vector Machines Recursive Feature Elimination (SVM-RFE) Algorithm to Classify Whether COPD Patients Have Been Continuously Managed According to GOLD Guidelines. International Journal of Chronic Obstructive Pulmonary Disease, 15, p.2779.
- [18] Granitto, P.M., Furlanello, C., Biasioli, F. and Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83(2), pp.83-90.
- [19] Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, pp.93-104.
- [20] Ploner, T., Heß, S., Grum, M., Drewe-Boss, P. and Walker, J., 2020. Using gradient boosting with stability selection on health insurance claims data to identify disease trajectories in chronic obstructive pulmonary disease. Statistical Methods in Medical Research, 29(12), pp.3684-3694.

- [21] Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S. and Zhou, S., 2019. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics*, 9(4), p.178.
- [22] Aich, S., Kim, H.C., Hui, K.L., Al-Absi, A.A. and Sain, M., 2019, February. A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease. In 2019 21st International Conference on Advanced Communication Technology (ICACTION) (pp. 1116-1121). IEEE.
- [23] Aich, S., Pradhan, P.M., Park, J. and Kim, H.C., 2018. A machine learning approach to distinguish Parkinson's disease (PD) patient's with shuffling gait from older adults based on gait signals using 3D motion analysis. *International Journal of Engineering & Technology*, 7(3.29), pp.153-156.



Satyabrata Aich, is working as a Lecture at the Institute of Digital Anti-Aging Healthcare, Inje University. He has published many research papers in journals and conferences in the realms of Machine learning, Text mining, and Artificial intelligence. His research interests are natural language processing, Machine learning, supply chain management, Text mining, Bio informatics, and Blockchain.



Hee-Cheol Kim, received his BSc at Department of Mathematics, MSc at Department of Computer Science in SoGang University in Korea, and PhD at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is Professor at Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include Machine learning, Text mining, Bio Informatics.



Ali Hussain is a master student at the Institute of Digital Anti-Aging Healthcare, Inje University. He has received his Bachelor degree in computer engineering from Pakistan. He has worked in many real-life projects related to artificial intelligence. His research interests are artificial intelligence, Machine learning, big data, machine vision.



Ikromjanov Kobiljon Komil Ugli is a master student at the Institute of Digital Anti-Aging Healthcare, Inje University. He has received his Bachelor degree in computer engineering from South Korea. He has worked in many real-life projects related to artificial intelligence. His research interests are artificial intelligence, Machine learning, big data, machine vision, Internet of Things.



Beomsu Kim is a master student at the Institute of Digital Anti-Aging Healthcare, Inje University. He has worked in many real-life projects related to artificial intelligence. His research interests are artificial intelligence, Machine learning, big data, text mining



Minji Kim is a master student at the Institute of Digital Anti-Aging Healthcare, Inje University. She has worked in many real-life projects related to artificial intelligence. Her research interests are artificial intelligence, Machine learning, clinical research, healthcare data analytics



Harin Ryu is a master student at the Institute of Digital Anti-Aging Healthcare, Inje University. She has worked in many real-life projects related to artificial intelligence. Her research interests are artificial intelligence, Machine learning, clinical research, healthcare big data, and data analysis