Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 15 (2020) Pp-285-291

BBRC
Bioscience Biotechnology Research Communications

A Novel Machine Learning Approach for Prediction of Chronic Obstructive Pulmonary Disease

Nitin R. Chopde and Rohit Miri
Department of Computer Science and Engineering, Dr. C.V. Raman University Kota, Bilaspur (C.G.), India

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is the most serious chronic disease that begins slowly and progresses to eventual lung cancer. COPD diseases must be treated early before they become serious. Our goal is to predict COPD on the basis of extracted features from region of interest of Computed Tomography (CT) images of patients. We proposed machine learning model using supervised machine learning classifiers to predict COPD. In this paper we discussed prediction of CODP using machine learning approach. The early identification and prediction of lung diseases have become a necessity in the research, as it can facilitate the subsequent clinical management of patients. The proposed prediction models predict COPD and healthy (Non-COPD) efficiently from standard derived features set from CT images of COPD machine learning dataset. Our model used derived features set and trained model using machine learning classifier are Stochastic Gradient Descent, Logistic Regression, Multilayer Perceptron, Random Forest and XG boost applied with optimal parameter selection using distinctive approach which improves the performance of proposed Machine learning classifier. Overall scenario is novel approach for the prediction of COPD using proposed supervised machine learning algorithm.

KEY WORDS: CHRONIC OBSTRUCTIVE, PULMONARY, TOMOGRAPHY.

INTRODUCTION

COPD is common disease characterized by persistent respiratory symptoms and restricted air flow. On-going COPD is a significant reason for dreariness and mortality worldwide which situates social and financial load. Cheplygina, V.(2015),The author examines the stability of instance labels supplied by a variety of Classification systems on various data bases, three of which are medical image sets in particular for CT pulmonary photographs and uses an unmonitored measuring to determine instance stability. Cheplygina V. (2018), Gaussian texture selected and transfer learning for multicentre classification of COPD.COPD patients data tested based on knowledge graph and adaptive feature set algorithm used for feature selection and integrated model used for diagnosis. Hind J. et al.(2018), Investigate COPD cohort dataset and associative features are taken for standardization practice. These outcomes show that a normalization of training ought to be executed to guarantee the distribution of counterfeit positive outcomes is diminished and stopped and prove that as a standard practice given the subsequent data that can be furnished with its use. The proposed machine learning methods improves learning ability of model and predicts COPD or non COPD accurately.

Literature Review: Sørensen L. et al. (2009) Authors investigated the possibility of using texture metrics for on random CT samples, where the labels are based on external and with proposed texture based method can achieve 69% classification accuracy, which is much better than accuracy of area of voxels underneath threshold.

ARTICLE INFORMATION

Received 10th Oct 2020 Accepted after revision 27th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.15/50

Alharbey R. (2016), Authors proposed effective tools to provide assistance for the elderly COPD patients. A feed forward techniques back propagation algorithm is used for prediction. Sørensen L. et al. (2010) investigated in computed tomography (CT) images to estimate chronic obstructive pulmonary disease at 0.817 AUC. Cheplygina, V. et al. (2014) This is substantially better compared to integrating regional categorizations into an overall image grade and compared with standard computerized measurement measures in pulmonary CT, The researchers analysed various hypotheses in a multiple instance in the sense of COPD and received AUC of 0.742, given the overall distribution of lung tissue plates, even though there are conceptual areas of disease trends relevant to COPD. Cheplygina V. et al. (2015) Authors evaluate instance stability using unsupervised method and demonstrate that a balance performance strength adjustment can be made when comparing MIL classifiers.

Randomly sampled Region of Interest together with a simple Machine learning classifier and uses the averaging rule is a good starting point for distinguishing COPD from non COPD scans, achieving at most 79.0. Koppad, S. H. and Kumar, A. (2016) their at most aim is to develop a health care expert for the identification of COPD by using decision tree algorithm. Cheplygina V. et al. (2018), discovered that surface based measure was basically better at isolating between subjects with and without COPD than were the two most fundamental quantitative extents of COPD in the composition, which rely upon thickness. Yang G. et al. (2018) Investigator proposed a method for predicting daily COPD exacerbation risk has been developed using the trend pattern features obtained from longitudinal physiological measurement data. Fang Y. et al. (2019), Designed COPD model based on knowledge graph and find out results by using DSA-SVM. Jain P., et al. (2019)

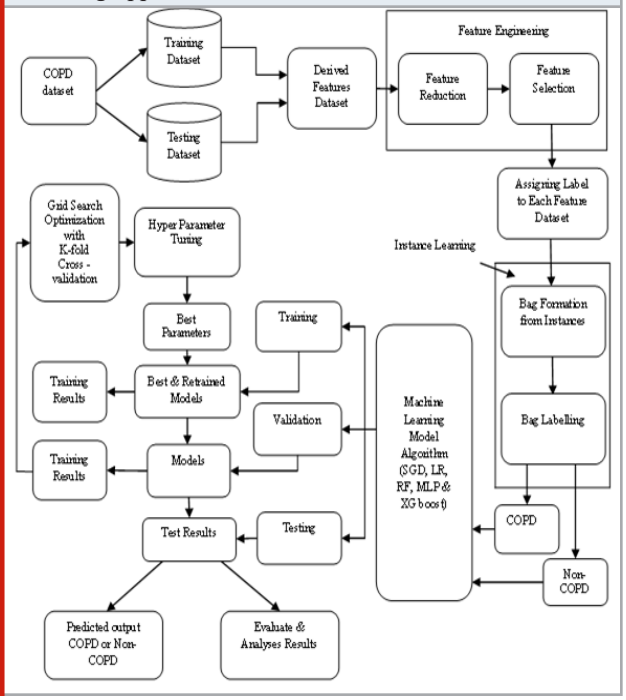
Authors applied classification techniques to improve classification performance and assessment time by reducing imbalance from used COPD dataset which has high dimension. Wang, Q. et al. (2020) proposed Gaussian process classification with gradient boosting decision tree feature transformation model to improve accuracy. Islam M. A. et al. (2018), Lung sound signals data are captured and classification performed using multichannel classifies approach which based on multiclass classification. Ajina K.A. et al., (2017) Author designed model for detection of COPD at early stages new method implemented using Naïve Bayes classifier and Euclidean distance algorithm. Matheson M. et al. (2016) Data driven predictive algorithms for predicting the occurrence of an exacerbation of COPD of a patient in the near future were trained, validated and compared and best accuracy achieved using a PNN classifier was 85.5%. As per review most of COPD prediction model suffered from prediction accuracy and performance in terms of AUC. We are aiming to optimize prediction ability of the proposed COPD prediction model.

MATERIAL AND METHODS

Dataset Acquisition: (COPD Machine Learning Datasets, 2020) COPD dataset contains derived features from CT images of patients scanned at the National Jewish Center in Denver, Colorado. This COPD dataset include training Dataset (DLCST) and testing Dataset (Frederikshavn).In given dataset two features are derived (extracted) gss and kdei. GSS is Gaussian scale space features or histograms of intensity values after filtering the image using eight filters, four scales and histograms of 10 bins. Botev, Z., et al. (2010), KDEI is a histogram intensity values in the ROI, estimated with Kernel Density Estimation. The feature size 320 for GSS and 256 for KDEI. In model learning final feature length for GSS is 30000×320 and 30000×256 for KDEI. The feature length of combined feature vector is 30000 ×576. The label of each ROI, determined by the subject's diagnosis: COPD (1) or healthy (Normal) (0). COPD diagnosis is determined according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria (FEV1/FVC < 0.7). COPD is a heterogeneous disease with various clinical presentations.

Feature Pre-processing and Feature Reduction: In feature pre-processing Calculate normalized value for all features set and scale all features data into specific range. Paoletti M. et al., (2009), Principle Component Analysis (PCA) allow us to reduce dataset into smaller number of dimensions with minimal loss of information. PCA used to make a classifier system more effective. PCA method is used before classifying used for dimensionality reduction of COPD disease dataset. In proposed prediction model performance of proposed Machine learning algorithms increased with feature reduction using PCA method.

Figure 1: Proposed COPD Perdition model using Machine Learning approach



Discrete Feature Selection and Hybrid Feature Selection Strategy: In discrete feature selection method we selected GSS as separate feature and not selecting other feature. We analyse effect of GSS feature on proposed ML classifier. The GSS feature is histogram intensity values from ROI of CT images of dataset. In this feature selection process algorithm performance are measured. In phase of hybrid feature selection both features are selected and combined features GSS and KDIE to improve accuracy of classifier. The proposed strategy of mixed feature selection optimizes the performance of classifiers.

In this case of features selection ML classifier model trained with large set of features which effects on predicting capability of classifier as compared. In bag labelling this bag formation instance learning performed and the bags formed from instances of dataset. This process happened before bag labeling. As per given dataset labelling for given data is COPD or Non COPD(Healthy) which is used to labelling data. The proposed ML model filters out data based on these labels.

Machine Learning Approach: The proposed approach is supervised machine learning algorithm and these learning algorithms make use of labeled data. We applied five supervised methods of machine learning (ML). Stochastic Gradient Descent Algorithm (SGD) is famous for its performance, which is mostly linear with the learning rate and simple implementation. This algorithm process that concerns with Random probability for i in range (m) and it is a slope plunge technique improved by the pace of union. Second we investigated COPD prediction model using Logistic Regression(LR) classifier which is most utilized ML algorithm to calculations for twofold arrangement and make use of given set of independent variable, it predict dependent variable by giving values 0 and 1, it predict values which lies in rage 0 to1.Third machine learning method is multilayer perceptron algorithm(MLP) consist of input, output and hidden layer of activating nodes learned through back propagation, during learning happened by switching weight after that data element processed.

Proposed Ensemble Machine Learning Method

Random Forest (RF): RF are very promising algorithm for its performance and takes less training time as compared to other algorithms.It predicts output with high accuracy, even for the large dataset it runs efficiently. We optimize RF algorithm by optimal selection of hyperparameter which improves its testing accuracy.RF proceeds with selecting random data points (K) from the training data and Build Decision tree (DT) related with the particular sample data points. Then Choose numbers say N for decision trees that to build. Repeating these procedure and finally for new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes. Considering B is bagging then Predications can be made by taking the majority vote in the case of classification trees using.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

XG Boost–

The learning ensemble increases the score of our computer model relative to simple models. Ichikawa D. et al. (2016), XG boosts an algorithm which constructs an arbitrarily differentiable loss function of the model, in a way similar to other boosting methods but which is more popular than a lot of approaches. In XG boost Loss function gives $l(y_i,y_i)$ our purpose to minimize following objectives

$$L(\emptyset) = \sum_{i=1}^{I} l(\hat{Y}_i, Y_i)$$

In case of the XG boost parameter prioritization, select a reasonably high rate of learning. A learning rate of 0.1 usually functions but varies from 0.5 to 0.3 for multiple issues. For this learning rate, decide the best number of trees in order to improve testing accuracy.

The proposed Novel Machine learning approach for Prediction of Chronic Obstructive Pulmonary Disease-
Input: Training and Testing instance set S, a vector of feature values and the class i.e.
label value
Feature Set F(i) = {GSS(i), KDEI(i)}
Label Set L(i) = {COPD(1), NORMAL(0)}
Initialization

Step1: Collect and Prepare feature data and label data from raw dataset values from

COPD Machine Learning Dataset.

Preprocessing Phase

Step2: For each feature data
Calculate the normalized value of all features set.
Scale the all feature data into specific range.
Perform Feature reduction using PCA

Parameter Hyper tuning Phase

Step3: Define the model for SGD, LR, RF, MLP and XG Boost.

Step4: Define the range of possible value for all hyper parameters of ML algorithms.
SGD: { 'alpha', 'max_iter', 'loss', 'penalty', 'n_jobs'}
LR: {'C', 'random_state', 'penalty', 'n_jobs'}
RF:{'n_estimators','max_features''criterion''max_depth','min_samples_split',' max_leaf_node s','random_state','min_samples_leaf'}
MLP:{ 'hidden_layer_sizes', 'max_iter' , 'activation', 'solver', 'alpha', 'learning_rate'}
XGboost:{'min_child_weight','objective','gamma','subsample','colsample_bytree',

'n_estimators','learning_rate','max_depth'}

Step5: Sampling of hyper parameters values using Grid Search CV Function.

Step6: Evaluate and find the best score among all hyper parameters value.

Step7: Validate the model using K-Fold Validation Learning Method.

Training Phase
Step8: Initialize the parameter tuned for ML model of SGD, LR, RF, MLP and XG boost.

Step9: Initialize the feature data and label data for training dataset.

Step10: Train the model for respective ML algorithms.

Step11: Validate the model performance using K-fold cross validation method.

Step12: If validation successful then saves the trained model TMsgd, TMlr,TMrf,TMmlp
and TMxgboost and if not the repeat from step 8.

Testing Phase
Step13: Initialize the feature data for testing dataset.

Step14: Load the trained model of ML algorithms.

Step15: Predict the results whether its COPD (1) or Normal (0).

Step16: Plot Confusion matrix between Actual Label Data and Predicted Label Data to
Check system accuracy.

Evaluation Phase
Step 17: Evaluate performance of classification model based on ROC, Confusion Matrix

Parameters based TP, FP, TN and FN.

Hyper parameter Optimization and Gird Search optimization: The choosing of acceptable hyperparameters for an algorithm is a problem for the priority tuning of hyperparameters. A hyperparameter is used to track the learning process. It strengthens model parameters to reliably train and analyze. Various thresholds, weights or speeds of learning may be used for generalizing different data patterns in the same form. Grid search lets you essentially pick the problem optimization parameter choices so that the test and error solution is automatic. We used Grid search method for optimization problems which supports to achieve the highest model accuracy. K-fold learning validate according to number of fold in order to optimize machine learning model.

RESULT AND DISCUSSION

We trained model using discrete features as GSS features and applied SGD, LR, MLP, RF and XG boost algorithm. The validate model using cross validation and finally performed testing to evaluate model. In second phase proposed classifier trained with KDEI features set and ML classifier tuned with optimal parameter. The confusion matrix contributes performance of algorithm and prediction ability of classifier. The confusion matrix for hybrid feature selection with Random Forest and XG boost method shown in Table-1.

Table 1. Confusion matrix using hybrid feature selection with RF and XG boost algorithm

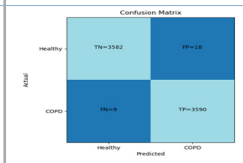
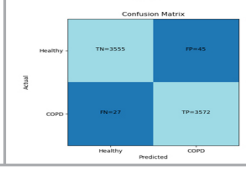
Machine Learning Algorithm	Confusion matrix	Discussion
Random Forest		In case of resultant confusion matrix, the correctly classified instances for Random Forest algorithm are 3582 which shows improved prediction ability for given sampling data. RF gives high prediction accuracy.
XG boost		The confusion matrix of XG boost is more precise in terms of prediction. XG boost True Positive 3572 and True Negative 3555 as shown in which correct classification of XG boost.

Table 2. Classification performance of RF Classifier with hybrid features set

RF	Precision	Recall	f1-score
Healthy	1.00	0.99	1.00
COPD	1.00	1.00	1.00

Table 3. Classification performance of XG boost Classifier with hybrid features set

XG boost	Precision	Recall	f1-score
Healthy	0.99	0.99	0.99
COPD	0.99	0.99	0.99

RF classifier with mixed (hybrid) feature selection method gives precision 1.00 for COPD and 1.00 for healthy and F1 score 1.00 for Healthy and 1.00 for COPD shown in Table-2. XG boost classifier with mixed (hybrid) feature selection method gives precision 0.99 for COPD and 0.99 for healthy and F1 score 0.99 for Healthy and 0.99 for COPD where recall is 0.99 as shown in Table- 3.

In case of GSS feature the performance of proposed model are improved and testing accuracy increased as we improved learning of algorithm. In case hybrid feature selection outstanding performance of both classifier as shown in Table-6.

Table 4. Performance of Proposed Machine learning algorithm using discrete feature selection

Feature Selection Strategy	Machine Learning Algorithm	Sensitivity	Specificity	AUC	Accuracy
Discrete features Selection GSS	Random Forest	0.9874	0.9975	0.9997	99.24%
	XG Boost	0.9874	0.9825	0.9987	98.49%

Table 5. Performance of Proposed Machine learning algorithm using discrete feature selection

Feature Selection Strategy	Machine Learning Algorithm	Sensitivity	Specificity	AUC	Accuracy
Discrete features Selection KDEI	Random Forest	0.9874	0.985	0.9996	98.62%
	XG Boost	0.9624	0.99	0.9949	97.62%

Table 6. Performance of Proposed Machine learning algorithm using hybrid feature selection

Feature Selection Strategy	Machine Learning Algorithm	Sensitivity	Specificity	AUC	Accuracy
Hybrid features Selection GSS and KDEI	Random Forest	0.9974	0.995	0.9999	99.62%
	XG Boost	0.9924	0.9875	0.9992	98.99%

Figure 2: ROC shows predction of COPD and Healthy using proposed Random forest classifier by using hybird feature selection

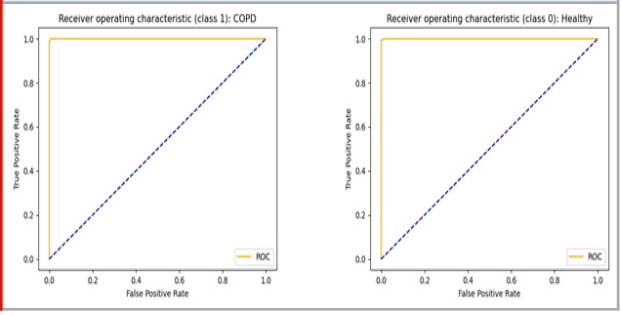


Figure 3: ROC shows predction of COPD and Healthy using proposed XG boost classifier by Hybrid feature selection

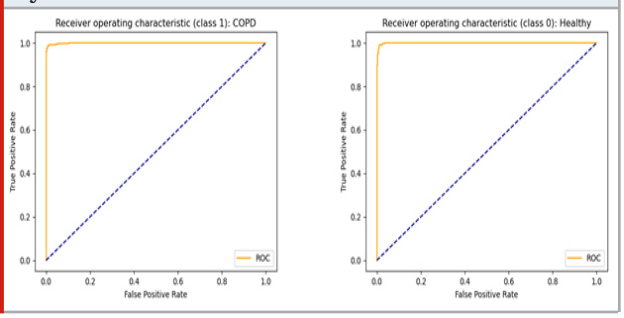


Figure 4: Accuracy comparison of proposed COPD prediction model using machine learning approach with previous model

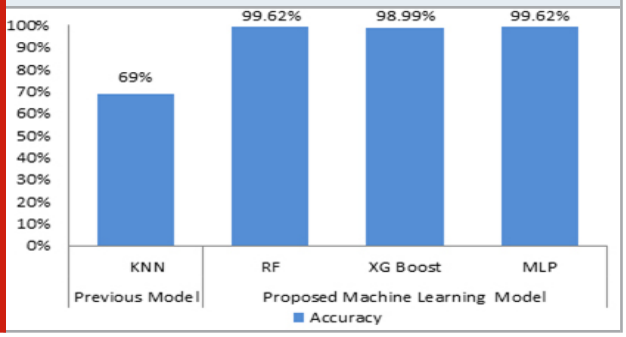


Figure 5: Comparative AUC of Proposed Model and Previous Model using Discrete feature selection GSS and KDEI

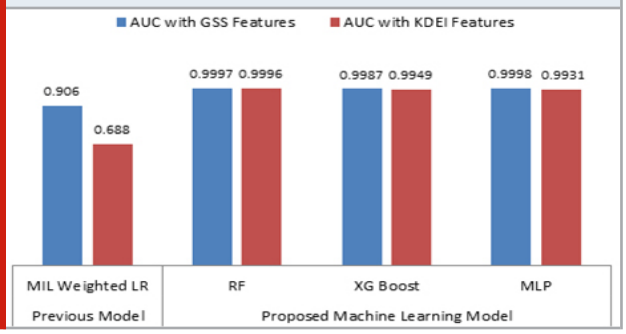
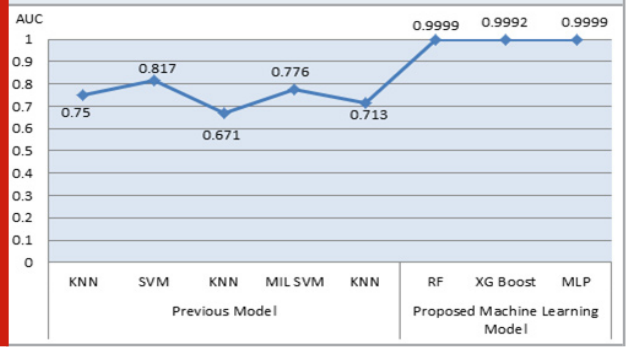


Figure 6: Comparative AUC of Proposed machine learning Model and Previous Model using hybrid features selection



ROC graph plotted TPR against FRP, if predicted and actual value is true then it is TP and if actual value is negative then is false positive. ROC of RF and XG boost model shows best performance at all threshold as shown in Figure-2 and Figure-3. We optimize proposed machine learning algorithm Random forest, XG boost and Chopde, N. R. and Miri, R., (2020) Multilayer perceptron which results in high accuracy as compared to previous model based on Sørensen, L. et al.(2009), KNN method, which shown in Figures-4.

The peak point of AUC 0.9999 for RF, 0.9992 for XG boost and Chopde, N. R. and Miri, R., (2020) 0.9999 for MLP are our proposed model AUC which are outstanding as compared to AUC of previous prediction models based on Sørensen, L. et al. (2009), KNN method, Sørensen, L. et al. (2010), SVM Method and Cheplygina, V. et al. (2014) MIL SVM method, Sørensen, L. et al. (2012) KNN as shown in Figure-6. COPD prediction model performance shows that our proposed machine learning algorithms are best in terms in many measures like AUC and accuracy.

CONCLUSION

In the proposed COPD prediction model we analyses performance of proposed machine learning classifier using derived features set GSS, KDEI from COPD dataset. It is found that feature selection effects on machine learning classifier performance. Our proposed COPD prediction model trained with variety of features set of COPD dataset and our novel approach optimize the performance of proposed machine learning classifier and gives remarkable AUC in case of MLP, RF and XG boost. The proposed prediction model worthy to diagnosis of patient is COPD or healthy it gives promising results. The accuracy of proposed Multilayer perceptron classifier gives 99.62% with AUC of 0.9999.

The ensemble method Random Forest classifier gives 99.62% accuracy with AUC of 0.9999 and XG boost reported 98.99% accuracy and AUC of 0.9992 which gives outstanding performance as compared to other classifier when hybrid features set are used. The resultant measures are superior like prediction accuracy and AUC with approach of machine learning classifier than previously reported research on same dataset. The novel machine learning approach for proposed COPD prediction model is very helpful for predicting chronic obstructive pulmonary disease patients at early stages and to assist patients by the expert. Our implemented system is reducing burden of healthcare system by contributing efficient COPD prediction model.

REFERENCES

Ajina, K. A., Prasad, A. S., Haseen, K. J., & Sivadasan, E. T. (2017). Application to predict chances for occurring COPD from symptoms. In IEEE International Conference on Intelligent Sustainable Systems (ICISS), pp.773–775.

Alharbey, R., (2016). Predictive Analytics Dashboard for Monitoring Patients in Advanced Stages of COPD. 49th Hawaii International Conference on System Sciences (HICSS), pp.3455–3461.

Botev, Z., Grotowski, J., and Kroese, D. (2010). Kernel density estimation via diffusion. The Annals of Statistics Journal, 38(5), pp.2916-2957.

Cheplygina V., Sørensen, L.,Tax, D. M. J, Pedersen, J. H., Loog, M., and de Bruijne,M.,(2014).Classification of COPD with multiple instance learning. In Proceedings International Conference on Pattern Recognition (ICPR), pp.1508–1513.

Cheplygina, V., Pena, I., Pedersen, J., Lynch, D., Sorensen,

L., and de Bruijne, M. (2018). Transfer Learning for Multicenter Classification of Chronic Obstructive Pulmonary Disease. IEEE Journal of Biomedical and Health Informatics, 22(5), pp.1-11.

Cheplygina, V., Sørensen, L., Tax, D., de Bruijne, M., and Loog, M. (2015). Label Stability in Multiple Instance Learning. Journal of Computer Science, 9349(6), pp.539-546.

Chopde, N. R. and Miri, R., (2020).Predictive Model for Chronic Obstructive Pulmonary Disease using Machine Learning Classifier. Solid State Technology Journal, 63(4), pp.7791 -7802.

COPD Machine Learning Datasets. http://bigr.nl/research/projects/copd. (Browsing date: 3rd Jan 2020).

Fang, Y., Wang, H., Wang, L., Di, R., & Song, Y. (2019). Diagnosis of COPD Based on a Knowledge Graph and Integrated Model. IEEE Access, 7, pp. 46004-46013.

Hind, J., Hussain, A., Al-jumeily, D., Montañez, C., Chalmers, C., & Lisboa, P. (2018). Robust Interpretation of Genomic Data in Chronic Obstructive Pulmonary Disease (COPD). In International Conference on Developments in eSystems Engineering (DeSE), pp. 12-17.

Ichikawa, D., Saito, T., Ujita, W., and Oyama, H. (2016). How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach, Journal of Biomedical Informatics, 64, pp.20–24.

Jain, P., Agarwal, A., and Behara, R. (2019). An Approach to Supervised Classification of Highly Imbalanced and High Dimensionality COPD Readmission Data on HPCC. IEEE International Systems Conference (SysCon), pp.1–7.

Koppad, S. H., and Kumar, A. (2016). Application of big data analytics in healthcare system to predict COPD. International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp.1–5.

Matheson, M., Bowatte, G., Perret, J., Lowe, A., Senaratna, C., Hall, G., de Klerk, N., Keogh, L., McDonald, C., Waidyatillake, N. T., Sly, P., Jarvis, D., Abramson, Lodge, C., Dharmaje, S. (2018). Prediction models for the development of COPD: a systematic review. International Journal of COPD, 13, pp. 1927-1935.

Paoletti, M., Camiciottoli, G., Meoni, E., Bigazzi, F., Cestelli, L., Pistolesi, M., & Marchesi, C. (2009). Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. Journal of Biomedical Informatics, 42(6), pp.1013–1021

Sørensen, L., Lo, P., Ashraf, H., Sporring, J., Nielsen, M., & de Bruijne, M. (2009). Learning COPD Sensitive Filters in Pulmonary CT. Medical Image Computing and Computer-Assisted Intervention – MICCAI, 5762(8), pp.699–706

Sørensen, L., Loog, M., Lo, P., Ashraf, H., Dirksen, A., Duin, R. P. W., & de Bruijne, M. (2010). Image Dissimilarity-Based Quantification of Lung Disease from CT. Medical Image Computing and Computer-Assisted Intervention – MICCAI, 6361(6), pp. 37–44

Yang, G., Kong, C., and Xu, Q. (2018). A Home Rehabilitation Comprehensive Care System for Patients with COPD Based on Comprehensive Care Pathway. In IEEE Fourth International Conference on Big Data Computing Service and Applications (Big Data Service), pp.161–168