Prediction of Respiratory System Diseases Using Machine Learning Algorithms

Mr. V. Prakasham
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
vprakashamcse@gmail.com

Ms. A. Priyadharshini
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
priyakrisharul@gmail.com

A. Amirtheswari
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
amirtheswariarumugam@gmail.com

M. Jeeva
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
banu101102@gmail.com

M. Madhumitha
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
madhumitha632002@gmail.com

K. Pavithra
Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, India
pavithracse1505@gmail.com

Abstract—The network of organs and tissues which help you breathe and It includes your airways, lungs and blood vessels and the muscles which power your lungs are known as respiratory systems. These corridors work together to move oxygen throughout the body and clean out waste feasts like carbon dioxide. Respiratory conditions, including pulmonary tuberculosis(PTB), chronic obstructive pulmonary complaint(COPD), pulmonary thromboembolism(PTE), and bronchiectasis, are among the most common conditions clinically. These conditions have common symptoms similar to cough, foam salivation, gasping, and casket pain, but the treatment and follow- up of each complaint are fully different. These types of conditions are anatomized through breath rate using random forest classifier, support vector classifier and logistic regression algorithm. Then, we're assaying chronic obstructive pulmonary disease, and Asthma by breath per minute(bpm).

Keywords- pulmonary tuberculosis, chronic obstructive pulmonary diseases, pulmonary thromboembolism, bronchiectasis, casket pain, breath rate, random forest classifier, support vector machine, logistic regression, breath per minute.

I.INTRODUCTION

When the respiratory system is mentioned, people generally think of breathing, but breathing is only one of the exertions of the respiratory system. To conserve the metabolic processes, the body cells need a nonstop force of oxygen. The respiratory system works with the circulatory system to give this oxygen and to remove the waste products of metabolism. It also helps to check pH of the race. Respiration is the process of switching the oxygen and carbon dioxide between the atmosphere and the body cells. The caprice- whams impulse the stimulation of breathing process, or ventilation for every 3 to 5 seconds which moves air through a series of passages into and out of the lungs. After this, there's a trade of feasts between the lungs and the race. This is called foreign respiration. The race transports the feasts to and from the kerchief cells. Eventually, the cells exercise the oxygen for their special exertion; this is called cellular metabolism, or cellular respiration. Together, these exertions constitute respiration.

The respiratory disease of asthma is substantially caused by inheritable complaints or by polluted air. The main symptoms caused by asthma are cough, wheeze, briefness of breath and casket miserliness. It is the most common chronic inflammatory disorder that is commonly affected for children and grown-ups. In our research, we have gone through the dataset of the persons to analyze the conditions of the disease for the patient. Using some machine learning algorithms we have analyzed the disease by the environmental condition around the person such as, air condition, humidity, and temperature of the location. The chronic obstructive pulmonary disease is caused by polluted air and smoking. In our paper, we have analyzed the copd for smoking persons to find the severity of the disease. Chronic obstructive pulmonary disease is a group of respiratory conditions which causes blockage in the air pipe and affiliated breathing problems. The main causes of Chronic Obstructive pulmonary disease are frequent coughing or gasping, redundant numbness or foam, briefness of breathing, and Trouble taking in deep breath. In our paper, we have analyzed the severity of the disease for the patient using a machine learning algorithm.

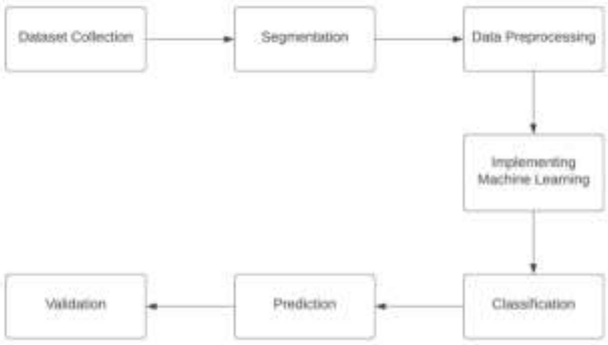
I.RELATED WORKS

In this section, we have read multitudinous papers to dissect where the machine learning algorithms are used for medical purposes.[1] An Early Detection of Asthma Using BOMLA Detector, then they have used ten machine learning classifiers videlicet, Support vector Classifier (SVC), Random Forest, Gradient Boosting Classifier, extreme Gradient Boosting, and Artificial Neural network to descry the asthma complaint. [2] In vitro Classification of Saliva Samples of COPD patients and healthy controls using Machine Learning Tools, here they have bandied COPD disease using saliva of the person through machine learning algorithms. Machine Learning algorithms for COPD patients readmission prediction: A data analytics Approach is explained in [3]. [4] Machine learning based asthma risk prediction using Iot and Smartphone Applications, In this paper they have generally used Peak Expiratory Flow Rates (PEAR) to find the inflexibility rate of asthma using external instruments. Application of Machine learning to support self- Management of Asthma with mHealth is explained in [5]. [6] Asthma , Alzheimer's and Dementia Disease Detection Based on Voice Recognition Using a Multi- Layer Perceptron Algorithm. Here they have analyzed asthma using voice recognition. [7] Artificial Intelligence Challenges in COPD management : a review. Using the review of the data which is collected from the person is reviewed by them. An Effective Random Generalized Linear Model to Predict COPD is explained in [8]. Non - invasive Diagnosis of COPD with E-nose Using XGBoost Algorithm is explained in [9]. [10] A Bioinformatics Approach for Deciphering the Pathogenic Processes of COPD. Here, they have analyzed the process of chronic obstructive pulmonary disease for the patient using bioinformation.

II.MATERIALS AND METHODOLOGY

Data collection

We've collected the dataset for asthma and COPD as a universal dataset. Through this dataset, we've trained using colorful algorithms i.eRandom Forest Classifier, Logistic Regression, and Support Vector Classifier.



.

Fig. 1. Block diagram for prediction of asthma and COPD using machine learning algorithms.

These COPD dataset contains 1000 persons with twenty four attributes: ID,AGE, PACK HISTORY, COPD SEVERITY, MWT 1, MWT 2, MWT 1 BEST, FEV 1, FEV 1 PRED, FVC, FVC PRED, CAT, HAD, SGRQ, AGE QUARTILES, COPD, GENDER, SMOKING, DIABETES, MUSCULAR, HYPERTENSION, ATRIAL FIB, IHD, BPM and for asthma, the data contains 808 persons with nine attributes: ID, GENDER, NAME, LOCATION, TEMPERATURE, HUMIDITY, TIME, AIR FRESHNESS, BPM. These sample data are collected from different periods ranging from 40 to 88. For preprocessing, the datasets are stored in a format which is grouped into two classes to know whether a case has asthma and COPD. To get the results, we've enforced violin plot, distplot, rel plot to avoid the complexity. This violin plot is used to visualize the distribution of numeric data and it also helps to visualize small data. In our project, the violin plot is used to measure the temperature and bpm for asthma and used to measure the severity of COPD by age, hypertension, and bpm.

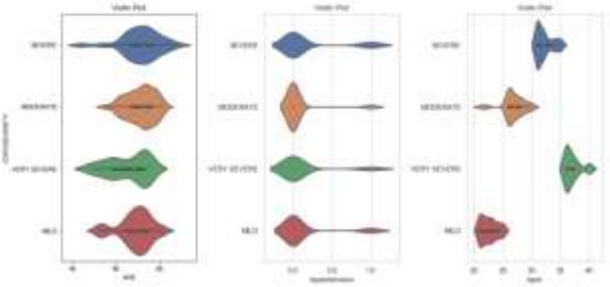
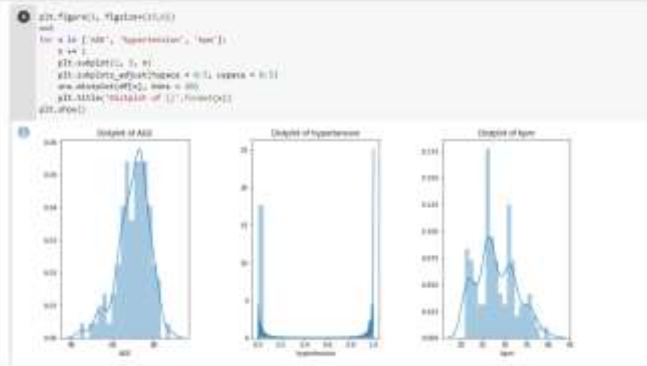


Fig. 2. Violin plot for COPD to identify the severity of the person.

By using matplot library, distplot is used to know the variations of the data. Here, distplot is used to analyze the age, hypertension, and bpm for COPD, and for asthma we have used temperature and bpm to analyze the data.



```
plt.figure(1, figsize=(15,6))
n=0
for x in ['AGE', 'hypertension', 'bpm']:
    n += 1
    plt.subplot(1, 3, n)
    plt.subplots_adjust(hspace = 0.5, wspace = 0.5)
    sns.distplot(df[x], bins = 20)
    plt.title('Distplot of {}'.format(x))
plt.show()
```

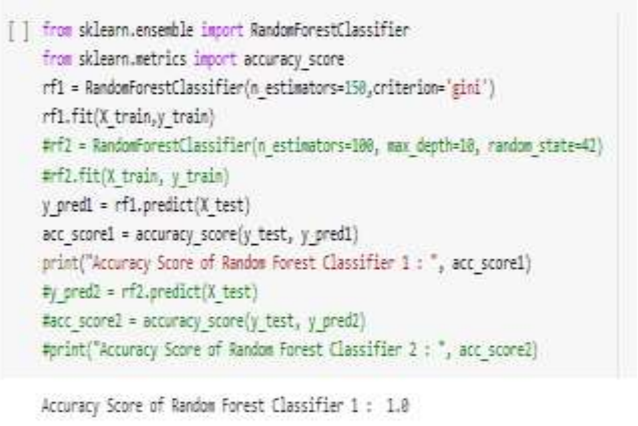
Fig. 3. Displot graph for COPD to identify the variation of severity of the patients.

Classification Models

Machine learning is used in this paper. Through machine learning, we have implemented three main algorithms namely; Random forest Classifier, Logistic Regression, and Support Vector Machine.

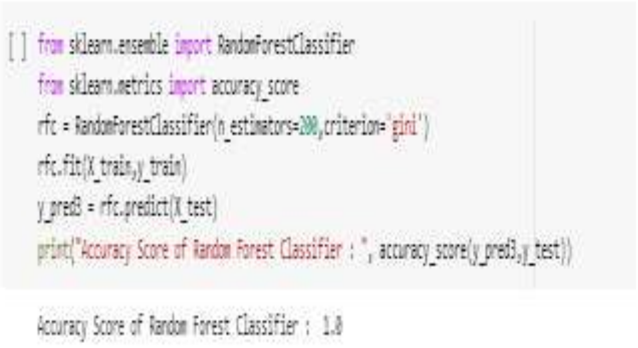
1.Random Forest Algorithm

Random forest algorithm is one of the machine learning algorithms which belongs to supervised learning. Supervised learning is defined as the process of training the data by known input and output. This algorithm is used to classify the data and helps for regression problems in Machine Learning. Random forest algorithm is a group of decision trees where the best decision is made by the method called Voting. It helps to check the multiple conditions and makes the data into less complexity. In our paper we have implemented this algorithm for testing and training the data to get the accuracy. The accuracy rate of random forest classifier 1.0.



```
[ ] from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import accuracy_score
    rf1 = RandomForestClassifier(n_estimators=150,criterion='gini')
    rf1.fit(X_train,y_train)
    #rf2 = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)
    #rf2.fit(X_train, y_train)
    y_pred1 = rf1.predict(X_test)
    acc_score1 = accuracy_score(y_test, y_pred1)
    print("Accuracy Score of Random Forest Classifier 1 : ", acc_score1)
    #y_pred2 = rf2.predict(X_test)
    #acc_score2 = accuracy_score(y_test, y_pred2)
    #print("Accuracy Score of Random Forest Classifier 2 : ", acc_score2)

    Accuracy Score of Random Forest Classifier 1 :  1.0
```

Fig. 4. The accuracy score of RFC for COPD.

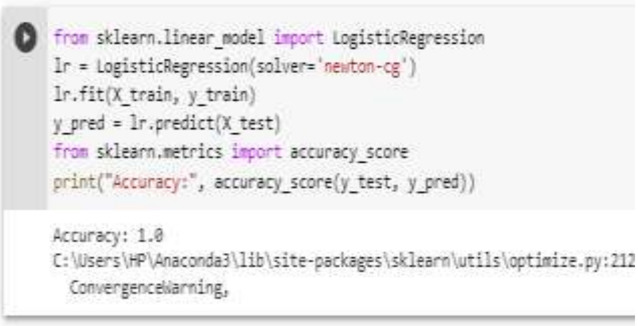


```
[ ] from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import accuracy_score
    rfc = RandomForestClassifier(n_estimators=200,criterion='gini')
    rfc.fit(X_train,y_train)
    y_pred3 = rfc.predict(X_test)
    print("Accuracy Score of Random Forest Classifier : ", accuracy_score(y_pred3,y_test))

    Accuracy Score of Random Forest Classifier :  1.0
```

Fig. 5. The accuracy score of RFC for Asthma.

2. Logistic regression

Logistic Regression is an algorithm which is based on supervised learning technique and it is also a machine learning algorithm. The resultant output is based on the relationship between the input and data. The logistic regression prediction value lies between 0 and 1. Logistic regression is used to visualize the training data using graphs. The accuracy score of Logistic regression for asthma is 1.0.

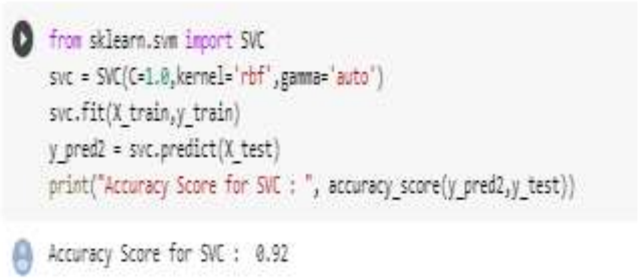


```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(solver='newton-cg')
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
from sklearn.metrics import accuracy_score
print("Accuracy:", accuracy_score(y_test, y_pred))

Accuracy: 1.0
C:\Users\HP\Anaconda3\lib\site-packages\sklearn\utils\optimize.py:212
  ConvergenceWarning,
```

Fig. 6. Accuracy score for COPD using logistic regression

3. Support Vector Machine

Support vector machine or support vector classifier is a machine learning algorithm which is also the supervised learning algorithm. It is used for the classification of data and for regression problems. Support vector machine is used to differentiate the similarities of the data. It helps to map the data which can be categorized to the similar data in the given dataset.



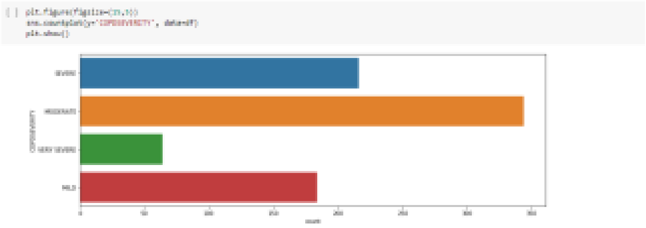
```
from sklearn.svm import SVC
svc = SVC(C=1.0,kernel='rbf',gamma='auto')
svc.fit(X_train,y_train)
y_pred2 = svc.predict(X_test)
print("Accuracy Score for SVC : ", accuracy_score(y_pred2,y_test))

Accuracy Score for SVC :  0.92
```

Fig. 7. The accuracy score of Support vector Classifier for Asthma.

III.RESULTS AND DISCUSSIONS

In our study, We have implemented machine learning algorithms like, Random forest algorithm, Support vector classifier, and Logistic regression. The analytical tool in our project is Google Cloud Platform(GCP) where the codes are implemented to test and train the data. We have collected the specific data of breath per minute, hypertension, and age for chronic obstructive pulmonary disease and asthma to predict the severity of the diseases for patients. In the COPD dataset we have twenty four attributes where the data are collected to research the person's severity of COPD. The main reason to collect the data of the person who is smoking is by knowing their pack history. If a person smokes, the severity range will increase and make the patient suffocate while they are breathing.

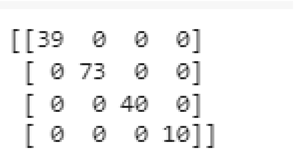
Through this data we have implemented plot diagrams (distplot, relplot, violin plot) and barchart diagram to know the patients severity, age, pack history ranges.



```
[ ] plt.figure(figsize=(10,5))
    sns.countplot(y='COPDSEVERITY', data=df)
    plt.show()
```

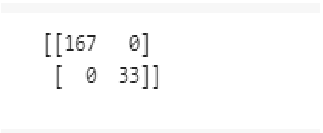
Fig. 8. Severity of COPD using barplot.

Using this data we have implemented the elbow method and find the label. These labels help to find the centers. These cluster centers are used to draw the 3D diagrams for the data. In Data pre-processing, standard scalar method is used to standardize the value to predict the accuracy. Then, these standard scalar values are fit to train the data.RFC, SVM, and Logistic regression algorithm are ensembled by voting classifier. Voting classifier is a model of machine learning which helps to predict the majority of votes by combining two algorithms to find the accuracy as output. The voting classifier is used for finding the accuracy for combination of two algorithms: Random forest classifier and Logistic regression for finding the accuracy value of COPD data. In the asthma dataset, voting classifier is ensemble random forest classifier and Support vector classifier to calculate the accuracy score. This ensemble model is used to fit and train the data, predict the testing data, and helps to calculate the accuracy score and confusion matrix. Confusion matrix is used to predict the TruePositive, TrueNegative, FalsePositive and FalseNegative of the matrix.



```
[[39  0  0  0]
 [ 0 73  0  0]
 [ 0  0 40  0]
 [ 0  0  0 10]]
```

Fig. 9. Confusion matrix for COPD.



```
[[167   0]
 [  0  33]]
```

Fig. 10. Confusion matrix for Asthma.

IV. CONCLUSION AND FUTURE SCOPE

The result of the project is to predict the disease of COPD and asthma using a universal dataset. The Receiver Operating Characteristic is used to show the graphical representation of all classification model classes. It helps to show the threshold of classification models which shows the True Positive rate and False Positive rate.

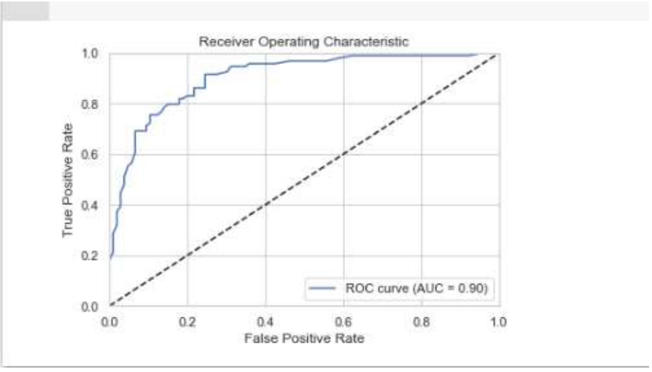


Fig. 11. The ROC curve for asthma.

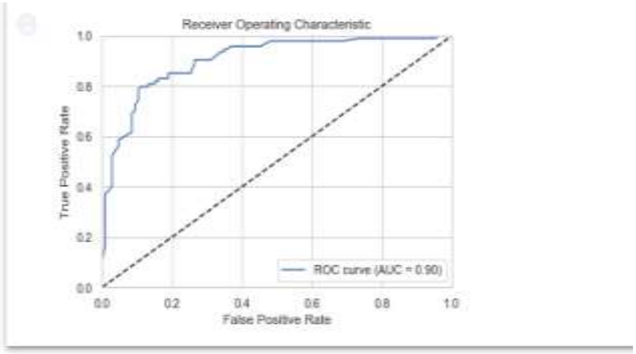


Fig. 12. The ROC curve for COPD.

This implementation of the project is developed into a machine which is used to detect Asthma and COPD by the person's breathing and heat maps. The waves are captured by the machine and analyze the disease of the person while crossing through the machine. This machine will scan and analyze the breath rate of the person and heat maps of the person to analyze whether they have asthma or chronic obstructive pulmonary disease.

REFERENCES

[1] Md. Abdul Awai, Md. Shahadat Hossain, Kumar Debjit, Nafiz Ahmed, Rajan Dev Nath, G.M Monsur Habib, Md Salauddin Khan, Md. Akhtarul Islam, M. A. Parvez Mahmud, vol. 9, "An Early Detection of Asthma Using BOMLA Detector", 2021 IEEE Access.

[2] Pouya Soltani Zarrin, Niels Roeckendorf, Christian Wenger, "In Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools", 2020, vol. 8, IEEE Access.

[3] Israa Mohamed, Mostafa M. Fouda, Khalid M. Hosny, "Machine Learning algorithms for COPD patients readmission Prediction: A Data Analytics Approach",vol. 10, 2022, IEEE Access.

[4] Gautam S. Bhat, Nikhil Shankar, Dohyeong Kim, Dae Jin Song, Sungchul Seo, Issa M. S. Panahi, Lakshman Tamil, "Machine Learning - Based Asthma Risk Prediction Using Iot and Smartphone Applications" , IEEE Access, vol . 9, 2021.

[5] Kevin C. H. Tsang, Hilary Pinock, Andrew M. Wilson, Syed Ahmar Shah, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society“Application of Machine learning to support self- Management of Asthma with mHealth”

[6] D. C Shubhangi, A.K Pratibha, 2021“Asthma , Alzheimer’s and Dementia Disease Detection Based on Voice Recognition Using Multi- Layer Perception Algorithm”.

[7] Lemana Spahic Becirovic, Amar Deumic, Lejla Gurbeta Pokvic,Almir Badanjevic, 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering“Artificial Intelligence Challenges in COPD management : a review”.

[8] Linah Saraireh, Mhd Saeed Sharif, Muna Alsallal,“An Effective Random Generalised Linear Model to Predict COPD”, 2020.

[9] V A Binson, Sania Thomas, G K Ragesh, Ajay Kumar, “Non - invasive Diagnosis of COPD with E-nose Using XGBoost Algorithm”, 2021.

[10] Aditya Saxena, “A Bioinformatics Approach for Deciphering the Pathogenic Processes of COPD”, 2021 5th International Conference on Information system and computer Networks.

[11] Stavros Nousias, Aris S. Lalos, Gerasimos Arvanitis, Konstatinos Moustakas, Triantafillos Tsirelis, Dimitrios Kikidis, Konstantious Votis,Dimitrious Tzovaras, “An mHealth System for Monitoring Medication Adherence in Obstructive Respiratory Diseases Using Content Based Audio Classification”, vol. 6.,2018.

[12] Md. Nazmul Islam Shuzan, Moajjem Hossain Chowdhury, Md. Shafeyet Hossain, Muhammad E. H. Chowdhury, Mamun Bin Ibne Reaz, Mahammad Monir Uddin, Amith Khandakar, Zaid Bin Mahbub, Sawal Hamid Md. Ali, “A Novel Non - Invasive Estimation of Respiration Rate From Motion Corrupted Photoplethysmography Signal Using Machine Learning Model”, vol. 9., 2021.

[13] Li Lou, Xinzhu Yu, Zhilin Yong, Chunyang Li, Yonghong Gu, “Design Comorbidity Portfolios to Improve Treatment Cost Prediction Of Asthma Using Machine Learning”, vol. 6, Issue: 6, 2020.

[14] Mouzzam Husain, Andrew Simpkin, Claire Gibbons, Tanya Talkar, Daniel Low, Paolo Bonato, Satrajit S. Ghosh, Thomas Quatieri, Derek T. O’Keeffe, “Artificial Intelligence for Detecting COVID - 19 With the Aid of Human Cough, Breathing and Speech Signals: Scoping Review”,vol. 3., 2022.

[15] Yuzhen Chen, Menghan Hu, Chunjun Hua, Guangtao Zhai, Jian Zhang, Qingli Li, Simon X. Yang,“Face Mask Assistant: Detection of Face Mask Service Stage Based on Mobile Phone”,vol. 21, Issue: 9,2021.

[16] Malikhah, Riyanarto Sarno, Sozo Inoue, M. Syauqi Hanfi Ardani, Doni Putra Purbawa, Shoffi Izza Sabilla, Kelly Rossa Sungkono, Chastine Fatichah, Dwi Sunaryono, Arief Bakhtiar, Libriansyah, Cita R. S. Prakoeswa, Damayanti Tinduh, Yetti Hernaningsih,”Detection of Infectious Respiratory Disease Through Sweat From Axillary Using an E-Nose With Stacked Deep Neural Network”,vo. 10.,2022.

[17] Om Prakash Singh, Ramaswamy Palaniappan, Mb Malarvili, “Automatic quantitative analysis of Human Respired Carbon Dioxide Waveform for asthma and Non - Asthma classification Using Support vector Machine”,vol. 6., 2018.

[18] Geumkyung Nah, Eun-A Choi, Jiwon Kim, Woojin Kim, Kwangha Yoo, Young-Youl Kim, Dankyu Yoon “Gene expression analysis of known COPD loci revealed its varied levels by disease severity”, 2021.

[19] Eun-A Choi, Ji-Won Kim, Guemkyung Nah, Woojin Kim, Kwangha Yoo, Young-Youl Kim, Dankyu Yoon, “Prediction of COPD severity based on clinical data using Machine Learning”,2021.

[20] Geumkyung Nah, Jiwon Kim, Eun-A Choi, Jeomkyu Lee, Woojin Kim, Kwangha Yoo, Dankyu Yoon, “Genome-wide association study identified genetic variants associated with severity of COPD in Korean participants”, 2020.

[21] Ali Hussain, Ikromjanov Kobiljon Komil Ugli, Beom Su Kim, Minji Kim, Harin Ryu, Satyabrata Aich, Hee-Cheol Kim, “Detection of Different stages of COPD Patients Using Machine Learning Techniques”,2021.

[22] Mohamed Chetoui, Moulay A. Akhloufi,”Automated Detection of COVID-19 Cases using Recent Deep Convolutional Neural Networks and CT images”,2021.

[23] Julia Zofia Tomaszewska, Christos Chousidis, Eugenio Donati, “Sound-Based Cough Detection System using Convolutional Neural Network”,2022.

[24] Ganesh Babu C, Gowri Shankar M, Gomathi P, Priyanka G S, Vidhya B,”Performance Analysis of Least Square Linear Regression with Various Classifiers for Cardiovascular Respiratory Detection from Capnography”,2022.