IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM

Mrs.P.Jasmine Lois Ebenezer, Mrs. S.Cyciliya pearcine christy and S.Saranya
Department of Computer Applications, Sarah Tucker College, Thirunelveli-7.

ABSTRACT

The Corona virus disease 2019 (COVID-19) pandemic, which originated in Wuhan China, has had disastrous effects on the global community and has overburdened advanced healthcare systems throughout the world. Globally; over 4,063,525 confirmed cases and 282,244 deaths have been recorded as of 11th May 2020, according to the European Centre for Disease Prevention and Control agency. However, the current rapid and exponential rise in the number of patients has necessitated efficient and quick prediction of the possible outcome of an infected patient for appropriate treatment using AI techniques. This paper proposes a fine-tuned Random Forest model boosted by the Ada Boost algorithm. The model uses the COVID-19 patient's geographical, travel, health, and demographic data to predict the severity of the case and the possible outcome, recovery, or death. The model has an accuracy of 94% and a F1 Score of 0.86 on the dataset used.

1. INTRODUCTION

Severe acute respiratory syndrome (SARS) has recently emerged as a new human disease, resulting globally in 435 deaths from 6,234 probable cases (as of 3 May 2003). Here we provide proof from experimental infection of cynomolgus macaques (Macacafascicularis) that the newly discovered SARS-associated corona virus (SCV) is the aetiological agent of this disease. Our understanding of the aetiology of SARS will expedite the development of diagnostic tests, antiviral therapies and vaccines, and may allow a more concise case definition for this emerging disease.

Automated point-of-care molecular assays have greatly shortened the turnaround time of respiratory virus testing. One of the major bottlenecks now lies at the specimen collection step, especially in a busy clinical setting. Saliva is a convenient specimen type that can be provided easily by adult patients. This study assessed the diagnostic validity, specimen collection time and cost associated with the use of saliva.

An outbreak of pneumonia, caused by a novel corona virus (SARS-CoV-2), was identified in China in December 2019. This virus expanded worldwide, causing global concern. Although clinical, laboratory, and imaging features of COVID-19 are characterized in some observational studies, we undertook a systematic review and meta-analysis to assess the frequency of these features. We did a systematic review and meta-analysis using three databases to identify clinical, laboratory, and computerized tomography (CT) scanning features of rRT-PCR confirmed cases of COVID-19. Data for 3420 patients from 30 observational studies were included. Overall, the results showed that fever (84.2%, 95% CI 82.6-85.7), cough (62%, 95% CI 60-64), and fatigue (39.4%, 95% CI 37.2-41.6%) are the most prevalent symptoms in COVID-19 patients. Increased CRP level, decreased lymphocyte count, and increased D-dimer level were the most common laboratory findings. Among COVID-19 patients, 92% had a positive CT finding, most prevalently ground-glass opacification (GGO) (60%, 95% CI 58-62) and peripheral distribution opacification (64%, 95% CI 60-69). These results demonstrate the clinical, paraclinical, and imaging features of COVID-19.

2. LITERATURE REVIEW

In this Review, we summarize the current knowledge on the origin and evolution of these two pathogenic corona viruses and discuss their receptor usage; we also highlight the diversity and potential of spillover of bat-borne corona viruses, as evidenced by the recent spillover of swine acute diarrhea syndrome corona virus (SADS-CoV) to pigs.

[2]The severe acute respiratory syndrome (SARS) has recently been identified as a new clinical entity. SARS is thought to be caused by an unknown infectious agent.

[3]Severe acute respiratory syndrome (SARS) has recently emerged as a new human disease, resulting globally in 435 deaths from 6,234 probable cases (as of 3 May 2003). Here we provide proof from experimental infection ocynomolgus macaques (Macaca fascicularis) that the newly discovered SARS-associated corona virus (SCV) is the aetiological agent of this disease. Our understanding of the aetiology of SARS will expedite the development of diagnostic tests, antiviral therapies and vaccines, and may allow a more concise case definition for this emerging disease.

[4]This study assessed the diagnostic validity, specimen collection time and cost associated with the use of saliva.

[5]This review attempts to give a comprehensive view of the origin of the virus, the mode of its entry and infecting human beings, and further discusses the possibility of new drugs and vaccines against the virus.

[6] In December 2019, novel corona virus (2019-nCoV)–infected pneumonia (NCIP) occurred in Wuhan, China. The number of cases has increased rapidly but information on the clinical characteristics of affected patients is limited.

[7] Recently, we reported the discovery of three novel corona viruses, bulbul corona virus HKU11, thrush corona virus HKU12, and munia corona virus HKU13, which were identified as representatives of a novel genus, Delta corona virus, in the subfamily Corona virinae. In this territory-wide molecular epidemiology study involving 3,137 mammals and 3,298 birds, we discovered seven additional novel delta corona viruses in pigs and birds, which we named porcine corona virus HKU15, white-eye corona virus HKU16, sparrow corona virus HKU17, magpie robin corona virus HKU18, night heron corona virus HKU19, wigeon corona virus HKU20, and common moorhen corona virus HKU21. Complete genome sequencing and comparative genome analysis showed that the avian and mammalian delta corona viruses have similar genome characteristics and structures. They all have relatively small genomes (25.421 to 26.674 kb), the smallest among all corona viruses. They all have a single papain-like protease domain in the nsp3 gene; an accessory gene, NS6 open reading frame (ORF), located between the M and N genes; and a variable number of accessory genes (up to four) downstream of the N gene. Moreover, they all have the same putative transcription regulatory sequence of ACACCA. Molecular clock analysis showed that the most recent common ancestor of all corona viruses was estimated at approximately 8100 BC, and those of Alpha corona virus, Beta corona virus, Gamma corona virus, and Delta corona virus were at approximately 2400 BC, 3300 BC, 2800 BC, and 3000 BC, respectively. From our studies, it appears that bats and birds, the warm blooded flying vertebrates, are ideal hosts for the corona virus gene source, bats for Alpha corona virus and Beta corona virus and birds for Gamma corona virus and Delta corona virus, to fuel corona virus evolution and dissemination.

[8]Emerging infectious diseases, such as severe acute respiratory syndrome (SARS) and Zika virus disease, present a major threat to public health1,2,3. Despite intense research efforts, how, when and where new diseases appear are still a source of considerable uncertainty. A severe respiratory disease was recently reported in Wuhan, Hubei province, China. As of 25 January 2020, at least 1,975 cases had been reported since the first patient was hospitalized on 12 December 2019. Epidemiological investigations have suggested that the outbreak was associated with a seafood market in Wuhan. Here we study a single patient who was a worker at the market and who was admitted to the Central Hospital of Wuhan on 26 December 2019 while experiencing a severe respiratory syndrome that included fever, dizziness and a cough. Metagenomic RNA sequencing4 of a sample of bronchoalveolar lavage fluid from the patient identified a new RNA virus strain from the family Corona viridae, which is designated here 'WH-Human 1' corona virus (and has also been referred to as '2019-nCoV'). Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that the virus was most closely related (89.1% nucleotide similarity) to a group of SARS-like corona viruses (genus Beta corona virus, subgenus Sarbeco virus) that had previously been found in bats in China5. This outbreak highlights the ongoing ability of viral spill-over from animals to cause severe disease in humans.

[9]An epidemic of severe acute respiratory syndrome (SARS) has been associated with an outbreak of atypical pneumonia originating in Guangdong Province, People's Republic of China. We aimed to identify the causative agent in the Guangdong outbreak and describe the emergence and spread of the disease within the province.

[10]Since the outbreak of severe acute respiratory syndrome (SARS) 18 years ago, a large number of SARS-related corona viruses (SARSr-CoVs) have been discovered in their natural reservoir host, bats1,2,3,4. Previous studies have shown that some bat SARSr-CoVs have the potential to infect humans5,6,7. Here we report the identification and characterization of a new corona virus (2019-nCoV), which caused an epidemic of acute respiratory syndrome in humans in Wuhan, China. The epidemic, which

started on 12 December 2019, had caused 2,794 laboratory-confirmed infections including 80 deaths by 26 January 2020. Full-length genome sequences were obtained from five patients at an early stage of the outbreak. The sequences are almost identical and share 79.6% sequence identity to SARS-CoV. Furthermore, we show that 2019-nCoV is 96% identical at the whole-genome level to a bat corona virus. Pair wise protein sequence analysis of seven conserved non-structural proteins domains show that this virus belongs to the species of SARSr-CoV. In addition, 2019-nCoV virus isolated from the bronchial veolar lavage fluid of a critically ill patient could be neutralized by sera from several patients. Notably, we confirmed that 2019-nCoV uses the same cell entry receptor—angiotensin converting enzyme II (ACE2)—as SARS-CoV.

3. METHODOLOGY

Multi Layer Perception

A multi-layer perceptron (**MLP**) has the same structure of a single layer perceptron with one or more hidden layers. The back propagation algorithm consists of two phases: the forward phase where the activations are propagated from the input to the output layer, and the backward phase, where the error between the observed actual and the requested nominal value in the output layer is propagated backwards in order to modify the weights and bias values.

Neural Network is made up of interconnected artificial neurons. They mimic human brain processing. The neurons interconnection link carries certain weight. The output of each neuron is determined by using an activation function such as sigmoid and step. In case neural networks (NN) are trained with training pattern of known classes, these are called supervised learning NN.

The supervised learning process of the neural network consists of a unique input signal and corresponding desired output signal .The network is trained until it reaches a stable state where the synaptic weights doesn't change and maps to their corresponding output. In recent years, there had been a great research in using neural network for classification of the mammography images.

In case of multi layer perception (MLP), neurons are connected in a network topology. They are placed in different layers and are connected through certain weights. We use 3 layer MLP containing input, hidden and output layers. The input layer consists of as many neurons as the number of features in a feature vector. Second layer, called *hidden layer*, contains *h* number of perceptions, where value of *h* is determined by experiment. Output layer contains only one neuron representing either benign or malignant value. We used sigmoid activation function for hidden and output layers. Batch learning method is used for updating weights between different layers.

The traditional feed-forward neural networks trained with the standard BP algorithm are called MLPs. In feed forward neural networks, the neurons of the first layer drive their output to the neurons of the second layer in a unidirectional manner (i.e., the neurons are not received from the reverse direction). The MLPs are supervised networks, thereby necessitating a desired response to be trained. Any input–output map can be approximated virtually by MLPs with one or two hidden layers. MLPs are employed in most neural network applications. A general structure of MLPNN comprising three layers is portrayed in Fig. 1.

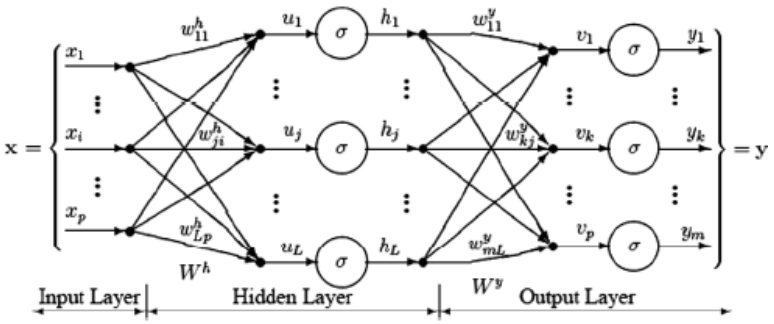


Figure 1

The structure of MLPNN

The only task of the neurons in the input layer is the distribution of the input signal xi to neurons in the hidden layer. Each neuron j in the hidden layer sums up its input signals xi after weighting them with the strengths of the respective connections wji from the input layer and computes its output yj as a function f of the sum [16], given by yj =f(_WjiXi). Here f can be a simple threshold function such as a sigmoid, or a hyperbolic tangent function [15]. The output of neurons in the output layer is computed in the

same manner. Following this calculation, a learning algorithm is used in adjusting the strengths of the connections so as to allow a network to achieve a desired overall behaviour.

DATASET

A collection database for machine learning oriented is available with the kaggle website which will be available from internet and it is open source. The data set are accommodated and preserved in the center for Machine Learning and Intelligent Systems in the University of California, Irvine. Each dataset contains separate webpage that presents the entire facts about its inclusion and any pertinent research that is examining it.

The datasets from the internet will take up the format of ASCII files, frequently the helpful CSV arrangement.

Benefits of the Repository

1. All datasets are obtained from the field, referring that they contain real-world potentials.
2. Datasets encompass broader variety of topics.
3. Particulars contained in the datasets will be abridged with the help of concepts such as types of attributes, amount of illustrations, quantity of features and year of establishment that might be arranged and examined.
4. Datasets are analyzed that refers which are recognized according to characteristics that are influencing anticipated to generate better Outcomes. This will deliver a helpful foundation for analysis.
5. Maximum number of datasets are minor in size stating that it can be viewed in a text editor that you can readily load them in a text editor or MS Excel and analyze them, and could be speedily simulated.

Dataset Collection

- https ://www.kaggl e.com/kimji hoo/coron avirusdata set/data.
- No. of records = 858
- No. of attributes = 36

Performance Metrics

The performance evaluation of proposed Neural Networks are simulated using PYTHON under windows environment. The implementation of this framework is performed on lung cancer dataset obtained from the UCI machine learning repository site [8]. The lung dataset given as input to the neural network and the data is divided into training data and test data. The training set for the neural network consists of 70% of the total dataset and the testing set is 30% of the total data. The proposed method is effectively compared with Back Propagation algorithm, Multi layer perceptron and stochastic gradient descent algorithm in terms of performance metrics obtained from confusion matrix shown in table 1.

Table 1 Confustion Matrix

		Predicted Class	
	Total Population	Prediction Positive	Prediction Negative
Actual Class	Condition Positive	True Positive (TP)	False Negative (FN)
	Condition Negative	False Positive (FP)	True Negative (TN)

Performance Metrics

- True Positive (TP) - The Extracted dataset containing cancer nodule is classified as cancerous.
- False Positive (FP) - The Extracted dataset without cancer nodule is classified as cancerous.
- True Negative (TN) - The Extracted dataset without cancer nodule is classified as non-cancerous.
- False Negative (FN) - The Extracted dataset containing cancer nodule is classified as non-cancer

Accuracy

This means as many times the different samples or images are tested with the same algorithm and the machine or system provides results how much accurate. The accuracy is the proportion of true results (both true positive and true negative) in the total data.

Accuracy(A) = (TP+TN) / (TP + TN + FP + FN)

Sensitivity

Sensitivity means that how accurately a cancer test identifies people as presence of lung cancer. Recall(R) or Sensitivity=TP(TP+FN)

Specificity

Specificity means that how accurately a cancer test identifies people who do not have lung cancer. Precision (P) or (1−Specificity) = TP (TP+FP)

Analysis

Here several performance metrics are used to check the segmentation.

Segmentation results of an image and ground truth of an image are compared to evaluate the performance.

1. Accuracy
2. Precision
3. Recall

Accuracy

An alternative metric to evaluate a semantic segmentation is to simply report the percent of pixels in the image which were correctly classified. The pixel accuracy is commonly reported for each class separately as well as globally

across all classes. When considering the per-class pixel accuracy we're essentially evaluating a binary mask; a true positive represents a pixel that is correctly predicted to belong to the given class (according to the target mask) whereas a true negative represents a pixel that is correctly identified as not belonging to the given class.

Accuracy=TP+TN/TP+TN+FP+FN

TP True Positive **TN** True Negative **FP** False Positive **FN** False Negative

Precision

Precision effectively describes the purity of positive detections relative to the ground truth.

Precision=TP/TP+FP

TP True Positive **TN** True Negative **FP** False Positive **FN** False Negative

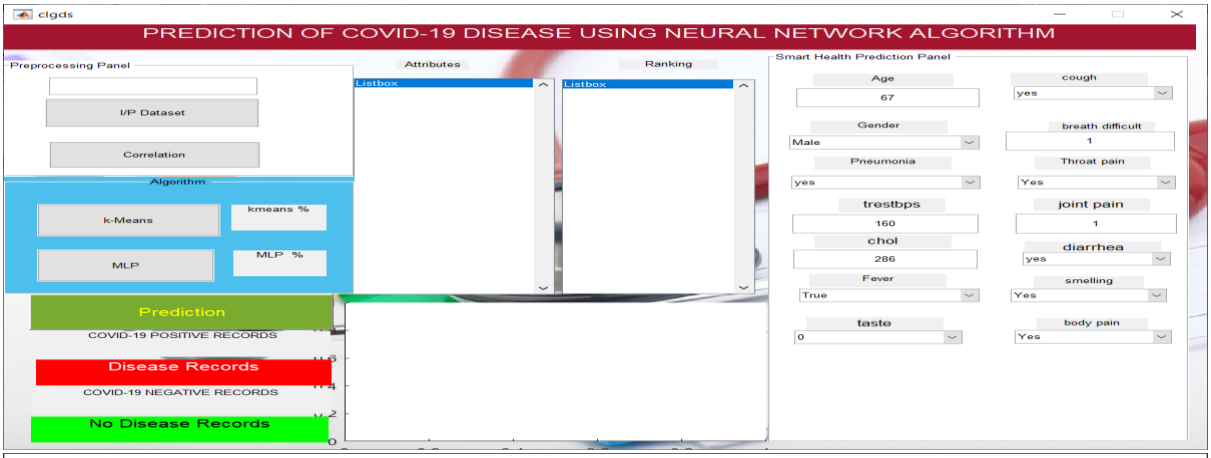
Recall

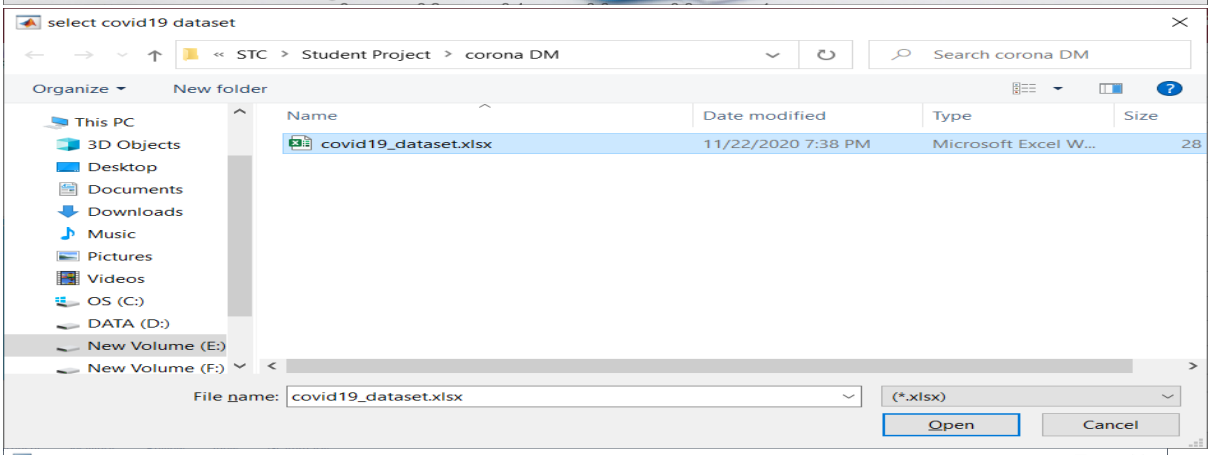
Recall effectively describes the completeness of positive predictions relative to the ground truth.

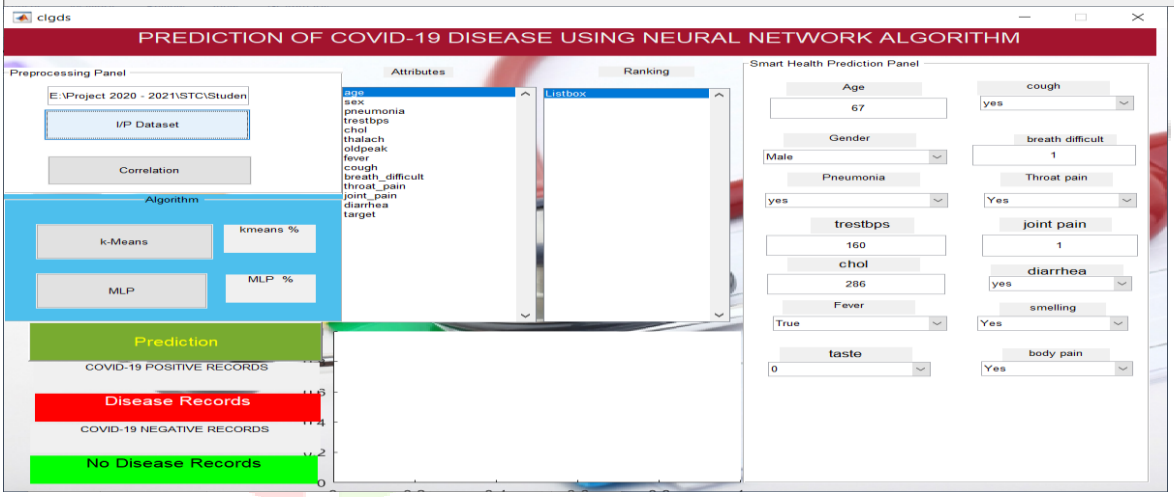
Recall=TP/TP+FN

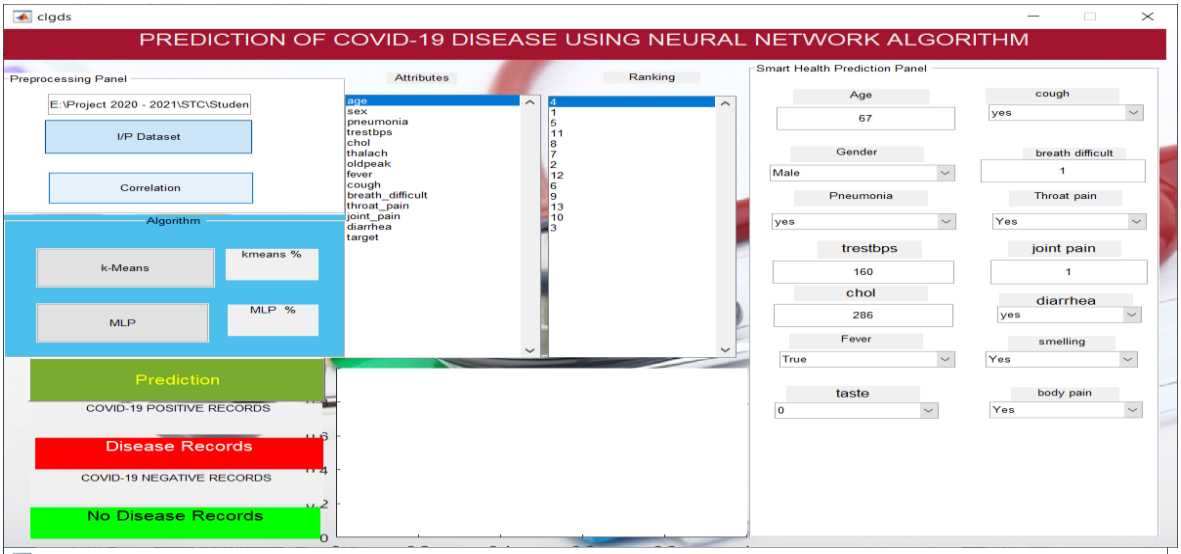
TP True Positive **TN** True Negative **FP** False Positive **FN** False Negative

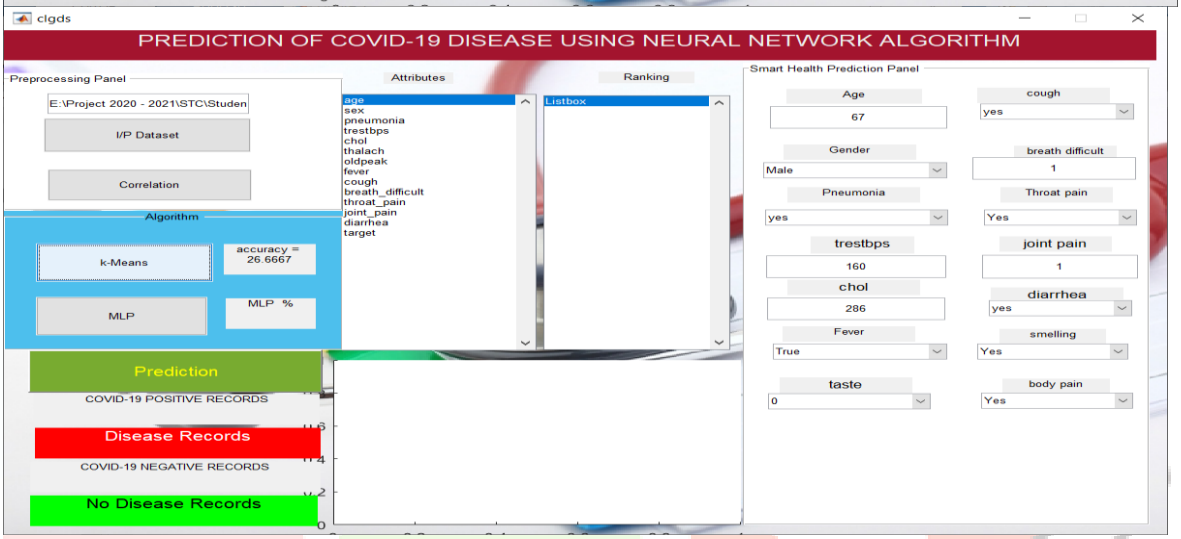
4. RESULTS


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
I/P Dataset
Correlation
Algorithm
k-Means
kmeans %
MLP
MLP %
Prediction
COVID-19 POSITIVE RECORDS
Disease Records
COVID-19 NEGATIVE RECORDS
No Disease Records
Attributes
Ranking
Listbox
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes


select covid19 dataset
STC > Student Project > corona DM
Search corona DM
Organize
New folder
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
DATA (D:)
New Volume (E:)
New Volume (F:)
Name
Date modified
Type
Size
covid19_dataset.xlsx
11/22/2020 7:38 PM
Microsoft Excel W...
28
File name:
covid19_dataset.xlsx
(*.xlsx)
Open
Cancel


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
E:\Project 2020 - 2021\STC\Studen
I/P Dataset
Correlation
Algorithm
k-Means
kmeans %
MLP
MLP %
Prediction
COVID-19 POSITIVE RECORDS
Disease Records
COVID-19 NEGATIVE RECORDS
No Disease Records
Attributes
Ranking
age
sex
pneumonia
trestbps
chol
thalach
oldpeak
fever
cough
breath_difficult
throat_pain
joint_pain
diarrhea
target
Listbox
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
E:\Project 2020 - 2021\STC\Studen
I/P Dataset
Correlation
Algorithm
k-Means
kmeans %
MLP
MLP %
Prediction
COVID-19 POSITIVE RECORDS
Disease Records
COVID-19 NEGATIVE RECORDS
No Disease Records
Attributes
Ranking
age
sex
pneumonia
trestbps
chol
thalach
oldpeak
fever
cough
breath_difficult
throat_pain
joint_pain
diarrhea
target
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
E:\Project 2020 - 2021\STC\Studen
I/P Dataset
Correlation
Algorithm
k-Means
accuracy = 26.6667
MLP
MLP %
Prediction
COVID-19 POSITIVE RECORDS
Disease Records
COVID-19 NEGATIVE RECORDS
No Disease Records
Attributes
Ranking
Listbox
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes

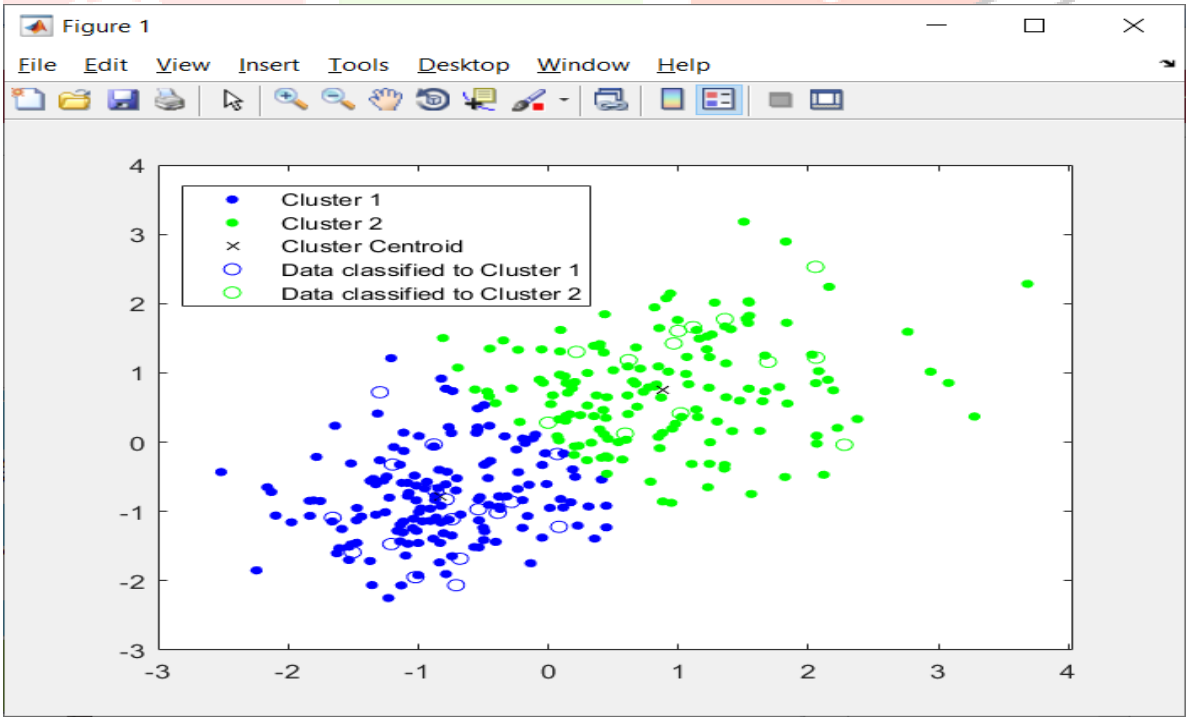
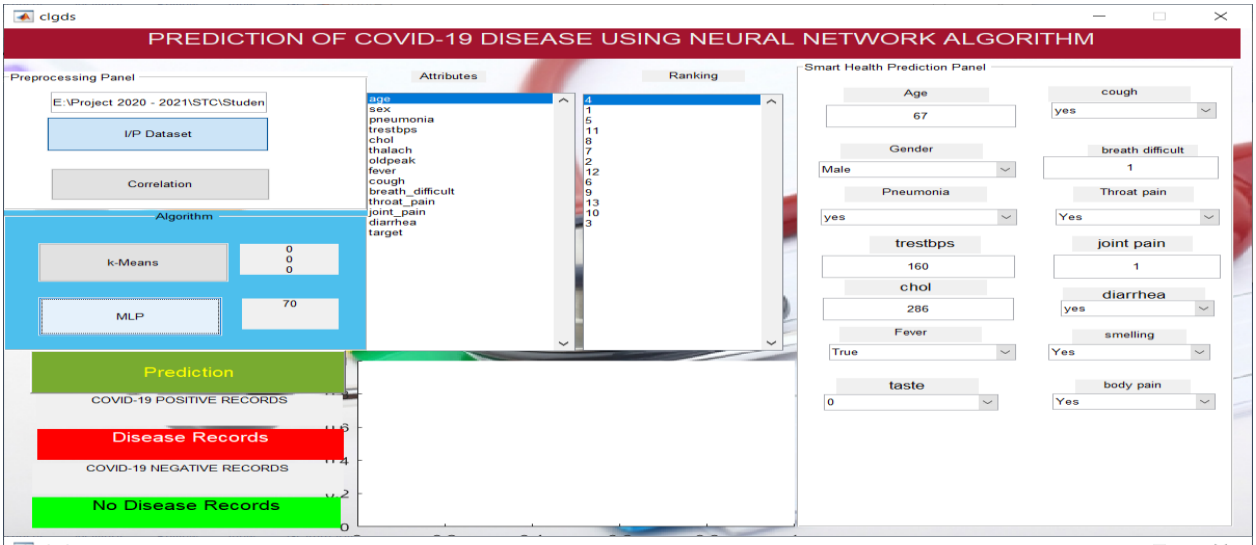
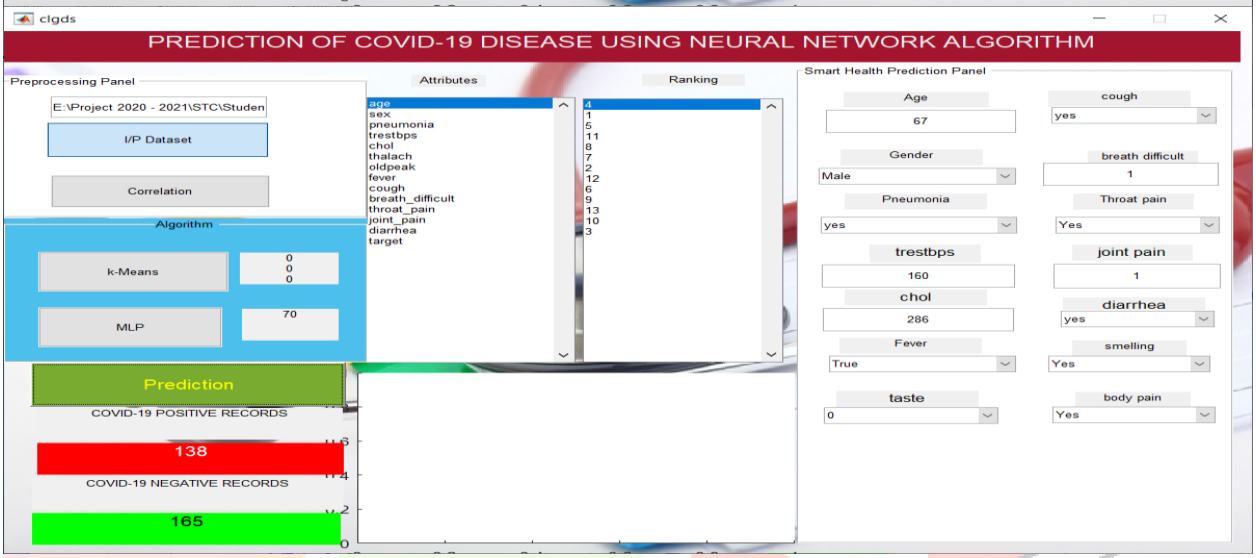
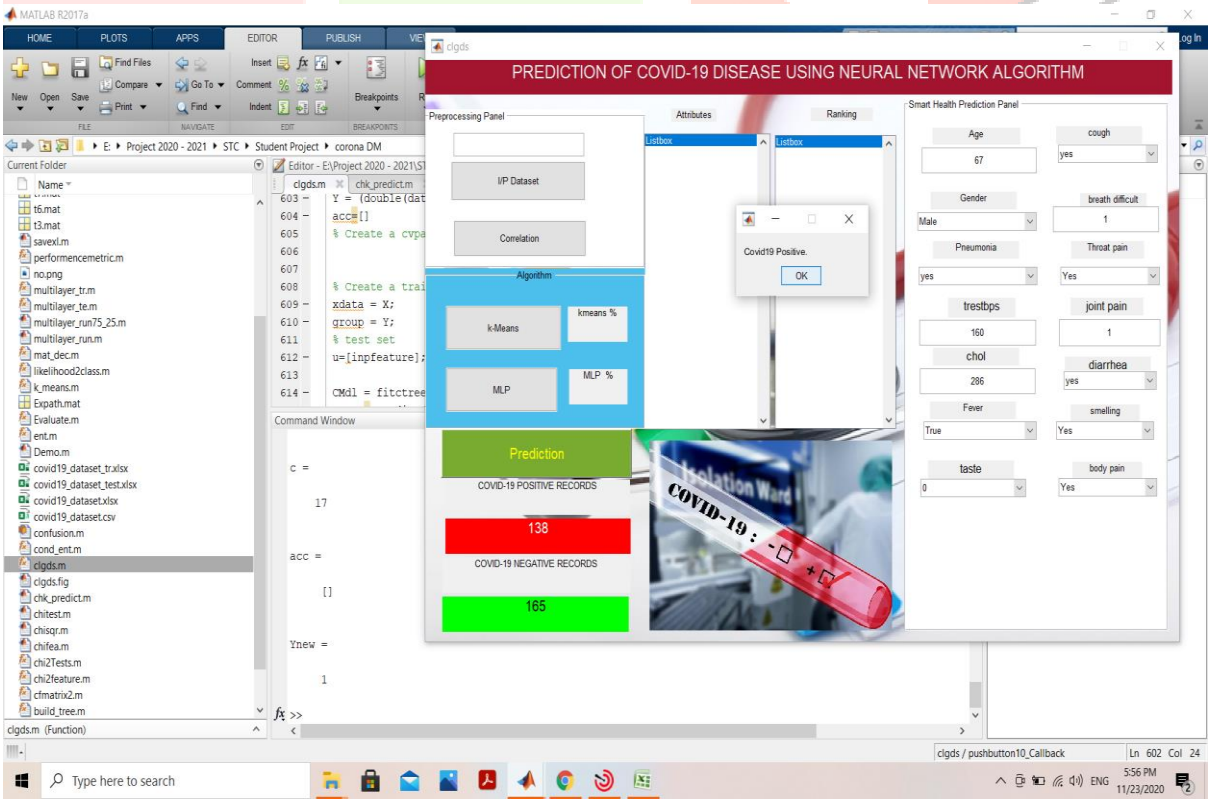
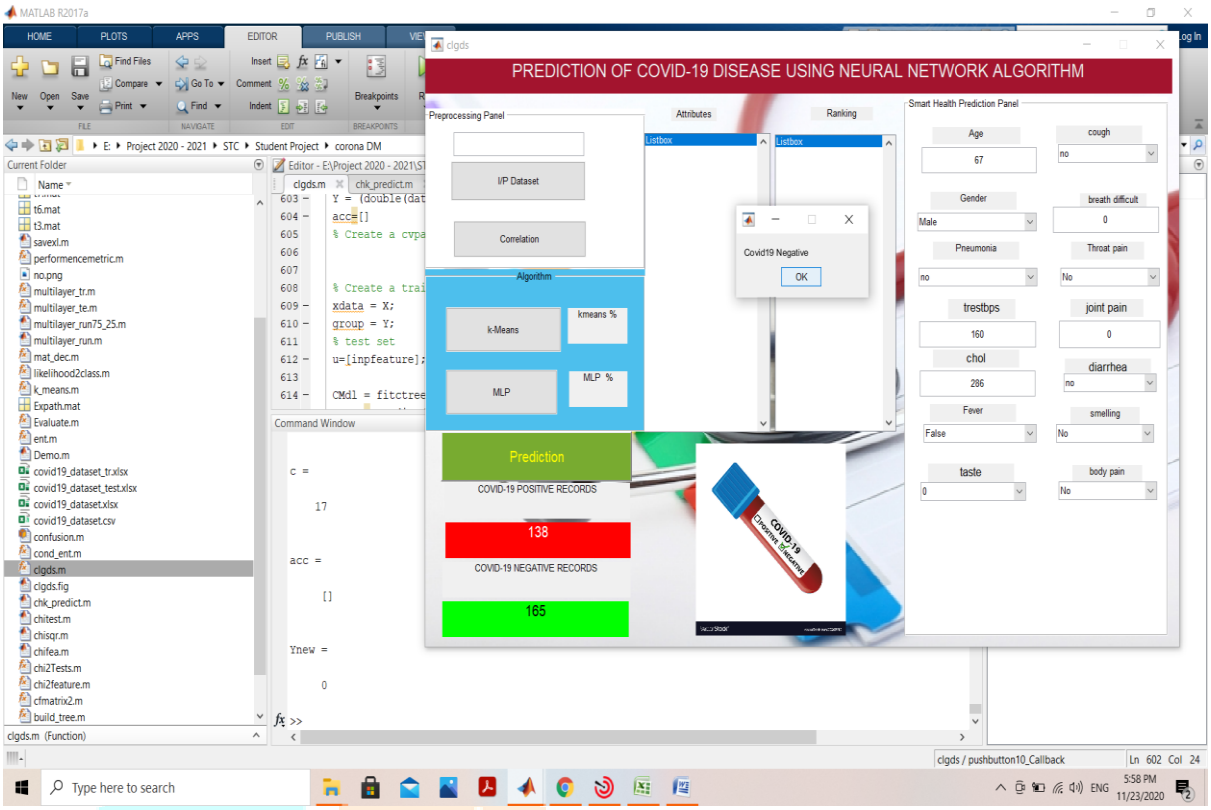


Figure 1
File Edit View Insert Tools Desktop Window Help
Cluster 1
Cluster 2
Cluster Centroid
Data classified to Cluster 1
Data classified to Cluster 2
4
3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3 4


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
E:\Project 2020 - 2021\STC\Studen
I/P Dataset
Correlation
Algorithm
k-Means
MLP
Attributes
Ranking
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes
Prediction
COVID-19 POSITIVE RECORDS
Disease Records
COVID-19 NEGATIVE RECORDS
No Disease Records


clgds
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Preprocessing Panel
E:\Project 2020 - 2021\STC\Studen
I/P Dataset
Correlation
Algorithm
k-Means
MLP
Attributes
Ranking
Smart Health Prediction Panel
Age
67
Gender
Male
Pneumonia
yes
trestbps
160
chol
286
Fever
True
taste
0
cough
yes
breath difficult
1
Throat pain
Yes
joint pain
1
diarrhea
yes
smelling
Yes
body pain
Yes
Prediction
COVID-19 POSITIVE RECORDS
138
COVID-19 NEGATIVE RECORDS
165


MATLAB R2017a
PREDICTION OF COVID-19 DISEASE USING NEURAL NETWORK ALGORITHM
Covid19 Positive.
OK
COVID-19 :
Prediction
COVID-19 POSITIVE RECORDS
138
COVID-19 NEGATIVE RECORDS
165



5. CONCLUSION

In the present study, data mining models were developed for the prediction of COVID-19 infected patients' recovery using epidemiological dataset of COVID-19 patients of South Korea. MLP is applied directly on the dataset using MATLAB programming language. The model developed with MLP was found to be the most efficient with the highest percentage of accuracy of 99.85%, followed by RF with 99.60% accuracy, then SVM with 98.85% accuracy, then K-NN with 98.06% accuracy, then NB with 97.52% accuracy and LR with 97.49% accuracy. The developed models would be very helpful in healthcare for the combat against COVID-19.

6. REFERENCE

[1]Cui J, Li F, Shi ZL. Origin and evolution of pathogeni coronaviruses. Nat Rev Microbiol 2019;17(3):181–92.

[2]Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 2003;348(20):1967–76.

[3]Fouchier RA, Kuiken T, Schutten M, van Amerongen G, van Doornum GJ, van den Hoogen BG, et al. Aetiology: Koch's postulates fulfilled for SARS virus. Nature 2003;423(6937):240 .

[4]To KKW, Yip CCY, Lai CYW, Wong CKH, Ho DTY, Pang PKP, et al. Saliva as a diagnostic specimen for testing respiratory virus by a point-of-care molecular assay: a diagnostic validity study. Clin Microbiol Infect 2019;25(3):372–8 .

[5]Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020;395(10223):470–3.

[6]Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020.

[7]Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, et al. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol 2012;86(7):3995–4008.

[8]Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579(7798):265–9.

[9]Zhong NS, Zheng BJ, Li YM, Poon, Xie ZH, Chan KH, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet 2003;362(9393):1353–8.

[10]Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579 (7798):270–3.