A Hybrid Two-Phase Machine Learning Model for Early COVID-19 Diagnosis Prediction

Ahmed S. Salama
Computers and Information Systems Department
Sadat Academy for Management Sciences
Cairo, Egypt

ABSTRACT

Across the history, COVID-19 pandemic is considered one of the deadliest diseases that harvested more than one million souls and left thousands of patients with damaged fibrotic lungs that physicians called post COVID syndrome. The main aim of this study is to propose a hybrid two-phase machine learning model to early diagnose COVID-19 based on available laboratory tests results, clinical symptoms, CT results, and demographic data in case of the difficulty of applying or absence of PCR test. The proposed model employs unsupervised learning Scalable Expectation Maximization (SEM) soft clustering mining model in the first phase to identify the most relevant identifying clusters characteristics for the disease grades, and in phase two the proposed model applies two proposed supervised learning classification mining models which are Association Rules (AR) based on improved Apriori algorithm, and Multilayer Perceptron(MLP) Multiclass Artificial Neural Network (ANN) to predict the COVID-19 disease diagnosis. The implemented proposed ML hybrid COVID-19 prediction model has successfully classified COVID-19 patients into positive mild, positive severe patients and discriminated between COVID-19 and Influenza patients/normal cases (COVID-19 negative) with an overall accuracy of 97.3%, a sensitivity 96%, and specificity 98%. It outperforms other reviewed state-of-the art COVID-19 diagnosis prediction models.

General Terms

Artificial Intelligence, Machine Learning, COVID-19

Keywords

Covid-19; Hybrid Model; Machine Learning; Soft Clustering; Association Rules; MLP Multiclass Artificial Neural Networks

1. INTRODUCTION

In the city of Wuhan, China in December 2019, pneumonia of unidentified aetiologia (novel coronavirus) was detected with its first death recorded on January 10, 2020, has turned into a pandemic, and labeled by the World Health Organization (WHO) as COVID-19 (Coronavirus 2019) [1]. Latest research has shown the promising opportunities of AI and ML techniques and algorithms for different pandemic breakouts. They help health professionals in different transmissible diseases (SARS, EBOLA, HIV, COVID 19) [2,3,4] and outbreak of diseases that are not communicable (Cancer, Diabetic, Heart, and Stroke) [5,6]. Early and quick identification of diseases, whether they are contagious or non-contagious, is a vital activity that gives health experts enough time to save lives early and lowers medical cost and disease outbreak [7,8]. The paper is organized as follows: section 2 presents a literature review of the related work. Section 3 introduces the proposed two-phase hybrid machine learning based model parameters explanation, architecture, proposed AR, and MLP ANN algorithms for COVID-19 diagnosis and prediction. Section 4 presents the implementation issues and the experimental results of the proposed hybrid model. Finally, section 5 introduces conclusions and a look forward for the expected future work in this domain.

2. RELATED WORK

Several studies and research were done on using AI and machine learning techniques to diagnose COVID-19 disease. This section introduces a review for some of them.

Javor et al. [9] introduced a new deep learning-based machine learning classification (ML) with a simplified programming approach, with an open-source dataset consisting of 6,868 CT images from 418 patients. Receiver Operating Characteristics (ROC) analysis was used to calculate Diagnostic performance metrics Rule-in and rule out thresholds were determined and checked. At the rule-in operating point, sensitivity and specificity were 84.4 % and 93.3 %. At the rule-out threshold, sensitivity (100 %) and specificity (60 %). Goodman-Meza et al. [10] aimed at designing and testing the COVID-19 inpatient diagnostic machine learning algorithm. The proposed algorithm has been used as a screening method based on basic demographic and laboratory features. For the final diagnostic classification, seven machine learning models were tested and employed as a combination. The model has reached 0.93 sensitivity and 0.64 specificity. An et al. [11] introduced a diagnostic model for COVID-19 diagnosis based on machine learning techniques that include the least absolute shrinkage and selection operator (LASSO), linear support vector machine (SVM), SVM with radial basis function kernel, random forest (RF), and k-nearest neighbors were applied. In mortality prediction, LASSO and linear SVM has achieved high sensitivities (90.7% and 92.0% respectively) and specificities (91.4% and 91.8% respectively). Khanday et al. [12] proposed a four-class approach which classified textual clinical reports using traditional and hybrid machine learning algorithms. The features were given to classifiers for conventional and hybrid machine learning. By having 96.2 percent accuracy testing, logistic regression and multinomial Naïve Bays have shown better results than other ML algorithms. For greater accuracy recurrent neural network can be employed. Banerjee et al. [13] presented four machine-study models that have been tested on the basis of blood tests to initially screen suspect COVID-19 cases. The models used include random forest (RF), artificial neural network (ANN), linear regression (LR) and generalized regularized Lasso-elastic network (GLMNET). The models have reached 81–87% of accuracy; 43–65% of sensitivity and 81–91% of specificity. Bao et al. [14] introduced models for the early detection of COVID-19 based on routine blood tests were investigated by random forest (RF) and the support vector machine (SVM). Three classification activities have been carried out (moderate vs viral, severe vs. viral and severe vs. moderate). Up to 15 blood characteristics were chosen to train the models. The best performance of the SVM-based classifier was 84% accuracy, 88% sensitivity, 80% specificity and 92% precision. Barbosa et al. [15] built a cheap COVID-19 blood sample detection system, using several ML classifications, including SVM, random forest (RF) and Bayesian networks (BN). In order to further minimize the expense and length of blood tests, features of training and testing of their models were reduced to 24. With 95.159% of the overall accuracy, sensitivity of 96.8%, precision of 93.8%, and specificity of 93.6%, the result achieved high classification efficiency. Experiments showed that BN performed better in comparison to other models. Nan et al. [16] used five types of classification models to identify the most powerful early diagnostic model for the COVID-19 infection by using: logistic regression (LR), support vector machine (SVM), decision tree(DT), random forest (RF) and deep learning neural network(DNN). The best performance was achieved by the LR classification model among the five classifiers with 91% accuracy, 87% sensitivity and 95% specificity. Bayat et al. [17] established a predictive random forest (RF) model of COVID-19 by combining clinical signs with common lab tests. The model was trained with 40-54 parameters, which resulted in 88.3% accuracy, 83.4% sensitivity and 89.8% specificity. Zoabi Y. et al. [18] proposed a machine-learning approach in which predictions were made using a gradient-boosting machine model built with decision-tree base-learners. The model predicted with high precision COVID-19 test results using only eight binary characteristics: sex, age 60 years, documented contact with the infected person and five initial symptoms.

3. THE PROPOSED HYBRID TWO-PHASE MACHINE LEARNING BASED MODEL FOR COVID-19 DISEASE DIAGNOSIS

In this section, the proposed hybrid model input and output parameters, architecture, and proposed used mining models' algorithms are introduced.

3.1 The Proposed Hybrid Model Input and Output Parameters

Based on a review of practice, chest consultants' opinions surveys, and empirical literature, the proposed model empowers six important categories of input parameters (20 features) that are used to diagnose Covid-19: A complete blood count (CBC) parameters, immunology parameter, chemistry parameters, clinical symptoms parameters, CT results parameter and demographic data parameters. The CBC, immunology, and chemistry parameters belong to generic category called laboratory tests. The CT results are represented using a common CO-RADS: a categorical CT assessment scheme for patients with suspected COVID- 19 [19]. CO-RADS is a CT-based system that is used to assess the suspicion of pulmonary involvement in COVID-19. For a patient diagnosis with COVID-19, CO-RADS needs to be accompanied with other data, such as laboratory test results, clinical findings, and type and duration of symptoms. That is why this proposed ML hybrid model used all these types of categories for predicting accurate COVID-19 diagnosis. These categorized input parameters with their descriptions are shown in Table 1. In addition, main predicted output parameter for the proposed model which is Covid-19 Diagnosis with its description is shown in Table 2. Negative means that the investigated person either normal or infected with other disease that is not compatible with COVID-19 such as Influenza, Positive Mild means that the patient is COVID-19 patient with mild lab test results, clinical symptoms, and CT result (CO-RADS 4), and Positive Severe means that the COVID-19 patient has typical features for pulmonary involvement of COVID-19 (CO-RADS 4 or 5) and supportive severe lab test results, and clinical symptoms. In this case, the patient may need noninvasive (CPAP), or invasive ventilation. The proposed hybrid model parameters for predicting Covid-19 diagnosis accompanied with normal range values, domain values and types are presented in Table 3.

Table 1. Categorized Covid-19 Model Diagnostic Input Parameters

1. CBC Parameters (Laboratory Tests)		
No	**Parameter Name**	**Description**
1	Haemoglobin (HB)	A protein found in the red blood cells that carries oxygen in your body
2	Red Blood Cell Count (RBC)	How many red blood cells (RBCs) a person has
3	Platelet Count (PC)	The average number of platelets in the blood.
4	Total leucocyte count (T.L.C.)	The number of leucocytes in the body
5	Lymphocytes%	Measures percentage of lymphocytes: white blood cells (immune cells)
2. Immunology (Laboratory Test)		
No	Parameter Name	Description
1	Ferritin in serum (F)	inflammatory marker: a marker of cellular damage
3. Chemistry (Laboratory Tests)		
No	Parameter Name	Description
1	level of lactate dehydrogenase (LDH) serum	High LDH blood or fluid means certain tissues in your body have been damaged by disease
2	C-reactive protein (CRP)	a blood test marker for inflammation in the body
3	D Dimer	a protein fragment (small piece) that has made when a blood clot dissolves in your body
4. CT Results		

No	Parameter Name	Description
1	CT COVID-19 grade	The level of suspicion of COVID-19 infection including the severity and stage of the disease graded from very low (negative) or CO-RADS 1 up to very high or CO-RADS 5

5. Clinical Symptoms

No	**Parameter Name**	**Description**
1	Body Temp (BT)	a measure of your body's ability to make and get rid of heat
2	Headache	a continuous pain in the head
3	Body aches	Body pain
4	Loss of Smell & Taste senses (anosmia and ageusia)	complete loss or absence of smell and taste
5	Diarrhea	loose, watery stools (bowel movements)
6	Dyspnea	Shortness of breath
7	Influenza Like Symptoms	A group of symptoms that are similar to those caused by the influenza (flu) virus
8	Other Symptoms	Recorded Covid-19 other rare symptoms include red eye, skin or CNS manifestations, stomachache

6. Demographic Data

No	Parameter Name	Description
1	gender	Person Sex Type (Male or Female)
2	Age	The length of time that a person has lived

Table 2. Covid-19 Predicted Output Parameter

No	Parameter Name	Description
1	Diagnosis	Decision whether there is Covid-19 from its signs and symptoms

Table 3. Categorized Covid-19 Diagnostic Model Parameters Normal Ranges, Domain, and Types

No	Parameter Name	Normal Range	Parameter Type
1	HB	11.5-16 g/dL	Continuous
2	RBC	3.8-5.4 *10^6 /uL	Continuous
3	Platelet Count (PC)	150-350 * 10^3/uL	Continuous
4	(T.L.C.)	4-11 * 10^3/uL	Continuous
5	Lymphocytes %	20-45%	Continuous
6	Ferritin in serum (F)	10-120 ng/mL	Continuous
7	(LDH) serum	0-247 U/L	Continuous
8	(CRP)	< 5 mg/L	Continuous
9	D Dimer	< 0.5 µg FEU/ml	Continuous
10	Body Temp (BT)	Normal:36.1°C to 37.2°C Fever: >37.5 to 38.3 °C High Fever: >40.0 or 41.0 °C	Continuous
No	Parameter Name	Domain	Parameter Type
11	Headache	No (0), Yes (1)	Discrete
12	Body aches	No (0), Yes (1)	Discrete
13	Loss of Smell & Taste senses	No (0), Yes (1)	Discrete
14	Diarrhea	No (0), Yes (1)	Discrete
15	Dyspnea	No (0), Yes (1)	Discrete
16	Influenza Like Symptoms	No (0), Yes (1)	Discrete
17	Other Symptoms	No (0), Yes (1)	Discrete
18	Gender	Male (1), Female(2)	Discrete
19	Age	1-115	Discrete

20	CT COVID-19 grade	0 (Negative) CO-RADS 1, 2, 3 1 (Positive Mild) CO-RADS 4 2 (Positive Severe) CO-RADS 5	Discrete
21	Diagnosis	0 Negative (COVID-19 Absence) 1 Positive Mild 2 Positive Severe	Discrete

3.2 The Proposed Hybrid Two-Phase Machine Learning Model for Predicting COVID-19 Diagnosis Architecture

The main components of the proposed hybrid model for predicting COVID-19 diagnosis architecture, and the two important implemented phases of the proposed unsupervised and supervised learning process to reach an accurate prediction for COVID-19 diagnosis and their outputs are presented in Fig. 1. Data cleaning, selection, and preparation operations involve removing of patients' medical data records that have missed values, redundancies, and noise or outliers. Next, a selection of the candidate parameters (features) for the mining process is performed. Finally, a data preparation process is performed that includes converting nominal attributes values to numerical values using integer encoding (Table 3). Last step is to convert the refined medical records dataset into a relational DB that is ready for mining. The Scalable Expectation Maximization (EM) algorithm as a soft clustering method [20], a Multilayer Perceptron (MLP) multiclass artificial neural network classifier and predictor mining model, and Association Rules (AR) based on improved Apriori algorithm classifier and predictor mining model are applied in this proposed model.

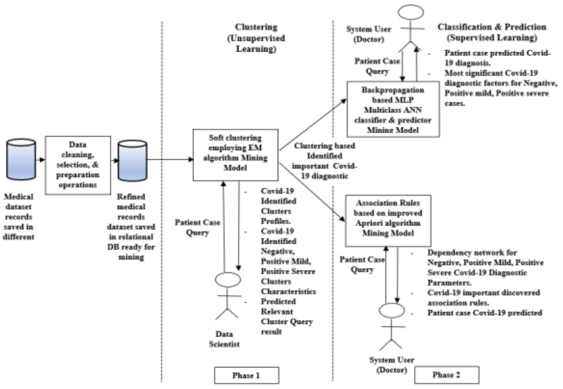

Clustering (Unsupervised Learning)
System User (Doctor)
Classification & Prediction (Supervised Learning)
Patient Case Query
- Patient case predicted Covid-19 diagnosis.
- Most significant Covid-19 diagnostic factors for Negative, Positive mild, Positive severe cases.
Backpropagation based MLP Multiclass ANN classifier & predictor Mining Model
Data cleaning, selection, & preparation operations
Soft clustering employing EM algorithm Mining Model
Clustering based Identified important Covid-19 diagnostic
Medical dataset records saved in different
Refined medical records dataset saved in relational DB ready for mining
Patient Case Query
- Covid-19 Identified Clusters Profiles.
- Covid-19 Identified Negative, Positive Mild, Positive Severe Clusters Characteristics
- Predicted Relevant Cluster Query result
Data Scientist
Association Rules based on improved Apriori algorithm Mining Model
Patient Case Query
- Dependency network for Negative, Positive Mild, Positive Severe Covid-19 Diagnostic Parameters.
- Covid-19 important discovered association rules.
- Patient case Covid-19 predicted
System User (Doctor)
Phase 1
Phase 2

Fig. 1: Architecture of the Proposed Hybrid Two-Phase Machine Learning Model for COVID-19 Diagnosis Prediction

3.3 Proposed Architecture and Algorithm for Back Propagation based MLP Multiclass ANN Mining Model

The proposed architecture of the Back Propagation based MLP multiclass ANN mining model for Predicting COVID-19 Diagnosis is shown in Fig. 2. The architecture of the proposed ANN consists of the input layer which contains a vector of twenty processing elements (PEs) which are the important identified COVID-19 diagnostic input parameters during clustering phase. Where x_1 .. x_{20} represent the input parameters for the ANN classified into 6 main categories: CBC parameters, Immunology parameter, Chemistry parameters, CT results, Clinical Symptoms, and Demographic data Parameter. The ANN also consists of one hidden layer with a vector of n PEs. h_1 .. h_n represent the n processing elements PEs in the hidden layer where the initial number of PEs in the hidden layer is ≈31 according to formula (4), and output layer with three PEs representing the three classes (states) of the COVID-19 predicted Diagnosis parameter, which are 0 Negative, 1 Positive Mild, and 2 Positive Severe. Listing of the input and output parameters of the proposed ANN model is explained previously in Table 1, and Table 2 section 4.1. Each W_{ij} represents the weight of the connection from the ith input PE x_i in the input layer to the jth hidden PE h_j in the hidden layer, and every w_{jk} represents weight of connection from the jth hidden PE h_j in the hidden layer to the kth output PE y_k in the output layer.

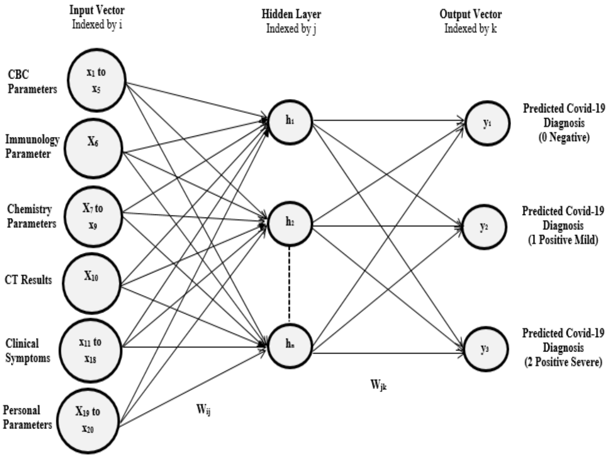


Fig. 2: The Back Propagation Based MLP ANN Based Mining Model for COVID-19 Diagnosis Prediction

Each PE (h_j) in the hidden layer or (y_k) in the output layer will do summation to combine and modify the inputs from the previous layer using the following equation:

$$m_j = \sum_{i=1}^{n} x_i \, w_{ij} + b_j \qquad (1)$$

where m_j is the net input to h_j in hidden or to y_k in output layer, x_i is the input to PE_j (or outputs of previous layer), i is the number of PE in previous layer, n is the number of inputs and b_j is the bias associated with PE_j. This weighted sum, then is passed to the activation function which is a hyperbolic tangent function (tanh) in the hidden layer PEs, whereas output PEs use a sigmoid function for activation. These used activation functions are as follows:

$$\text{(sigmoid function)} \quad f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

$$\text{(hyperbolic tangent function)} \; f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \qquad (3)$$

where x is the input and f is the output.

The formula used to determine the initial number of PEs in the hidden layer (h_n) is as follows:

$$hn = r * \sqrt{i * o} \qquad (4)$$

where r is the hidden node ratio (default value is 4.0), i is the total input PEs, and o is the total output PEs.

Algorithm 1: Proposed Backpropagation based MLP ANN Mining Model Algorithm for COVID-19 Diagnosis Prediction

```
1.  TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative
2.  D: Selected patients' dataset for identified COVID-19 diagnostic input parameters during clustering phase.
3.  X: The total number of cases COVID-19 diagnosis values to be predicted by the model.
4.  m: It takes value from 1 to 3 representing three diagnosis classes (Negative, Positive Mild, Positive Severe).
5.  Initialize all network weights and thresholds with random values between -1 to 1.
6.  Initialize macc = 0.0 // Initializes ANN mining model accuracy with 0.
7.  Calculate $hn$ //initial number of PEs in the hidden layer using formula (4)
8.  Determine percentage of the training and testing data of D.
9.  Repeat the following for the training dataset:
10. Calculate the weighted sum for each hidden PE ($h_j$) in the hidden layer using formula (1).
11.     Apply hidden layer activation function (3) to the calculated weighted sum.
12.     Calculate the weighted sum for each output PE ($y_k$) in the output layer using formula (1).
13.     Apply output layer activation function (2) to the calculated weighted sum.
14.     Calculate the difference between the predicted and actual value for the output PEs. (error)
15.     The gradients of the errors for the neurons in the output layer are calculated.
16.     The gradients for the weights between the hidden layer and the output layer are updated.
17.     The gradients of the errors for the neurons in the hidden layer are calculated.
18.     The gradients of the weights between the input layer and the hidden layer are updated.
19.     Update the weights of all the connections based on the gradients of the weights using this formula:
        $wij = wij + \eta \cdot \Delta wij$          (5)
        where $\eta$ is the learning rate.
20. Apply the learned ANN mining model on the testing dataset.
21. Calculate nmacc as follows: // nmacc is calculated new ANN mining model overall accuracy derived from 3-class confusion matrix
    $nmacc = \frac{\sum_{m=1}^{3} TP_m + \sum_{m=1}^{3} TN_m}{\sum_{m=1}^{3} TP_m + \sum_{m=1}^{3} TN_m + \sum_{m=1}^{3} FP_m + \sum_{m=1}^{3} FN_m}$
        (6)
22. If  nmacc > mcc Then
            Assign nmacc to macc
            Go to step 9
23. Endif
24. Stop // terminate training when the mining model accuracy no longer increases.
```

3.4 Proposed Improved Apriori Algorithm based Association Rules (AR) Mining Model

In this section, an improved Apriori algorithm for Association Rules mining model is presented. Improving the Apriori algorithm is done by defining the low minimum support, high minimum confidence, and minimum importance percentage to generate important COVID-19 diagnosis prediction association rules.

Algorithm 2: Proposed Association Rules mining model Algorithm based on Improved Apriori algorithm

```
1.  D: selected patients' dataset for identified COVID-19 diagnostic input parameters during clustering phase.
2.  Join step: C is generated by joining Lx-1 with itself.
3.  Prune step: Any (x-1 itemset) that is not frequent remove it.
4.  C: Candidate itemset of size x.
5.  L: Frequent itemset of size x.
6.  c : Candidate set in C.
7.  count: frequency support counter of candidate set c.
8.  cv: Confidence value.
9.  $\varepsilon$ : Support threshold // minimum support count
10. min_conf: Minimum confidence percentage or minimum probability percentage
11. min_imp: Minimum importance percentage or minimum lift percentage
12. ARS: Set of generated Association Rules.
13. A: Association Rules that their confidence values ≥ min_conf and importance value ≥ min_imp.
14. $\varepsilon$ = D * 25/100 // 25% as support threshold gave best results
15. L1 = {frequent items with count ≥ $\varepsilon$}
16. for (x=1; $L_x \neq \emptyset$; x++) do begin
17.       $C_{x+1}$ = candidate sets generated
                from $L_x$ // Applying Join step
18.       for each tuple t in D do
19.             Mt = {c ∈ $C_{x+1}$| c ⊆ t}
20.             for candidates c ∈ Mt
21.                 count[c] = count[c] +1
22.       $L_{x+1}$ = { c ∈ $C_{x+1}$| count[c] ≥ $\varepsilon$}
```

```
        // Applying Prune step
end // for
Generate ARS for $\cup_x L_x$
Calculate cv for each AR where AR ∈ ARS as follows:
    $cv(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$   (7)
Calculate iv for each AR where AR ∈ ARS as follows:
    $iv(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A) * Support(B)}$   (8)
A = {AR ∈ ARS| (cv[AR] ≥ min_conf) and (iv[AR] ≥ min_imp)}
Return A
```

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The used medical dataset was especially collected following a selective and thoroughly method to guarantee that it is highly representative for real negative, positive mild, and positive severe COVID-19 patients' cases. It represents real 291 COVID-19 investigated patients (127 COVID-19 Negative Patients includes 68 Influenza patients and 59 normal cases, 84 COVID-19 Positive Mild patients, and 80 COVID-19 Positive Severe patients). Based on a review of practice, chest consultants' opinions surveys, and empirical literature, the used dataset covers the most important COVID-19 diagnosis parameters. 70% of the prepared dataset was used in training and 30% of the dataset is used in testing the mining models. The proposed hybrid model was implemented using SQL Server Analysis Services and Visual Studio IDE. In the following subsections, the results of implementing the proposed hybrid model.

4.1 Proposed Soft Clustering mining model based on Scalable EM Results

Fig. 3 presents the generated 10 soft clusters diagram for the COVID-19 investigated patients according to the Diagnosis Parameter with labels identifying the highest probability levels for each grade of the diagnosis parameter: Negative, Positive Mild, and Positive Severe. The color density shows the distribution of the learning dataset population according to diagnosis value among the identified soft clusters (Positive Mild state 1 in this diagram). The connections show the relevant clusters. The shading of the line that connects one cluster to another represents the strength of the similarity of the clusters.

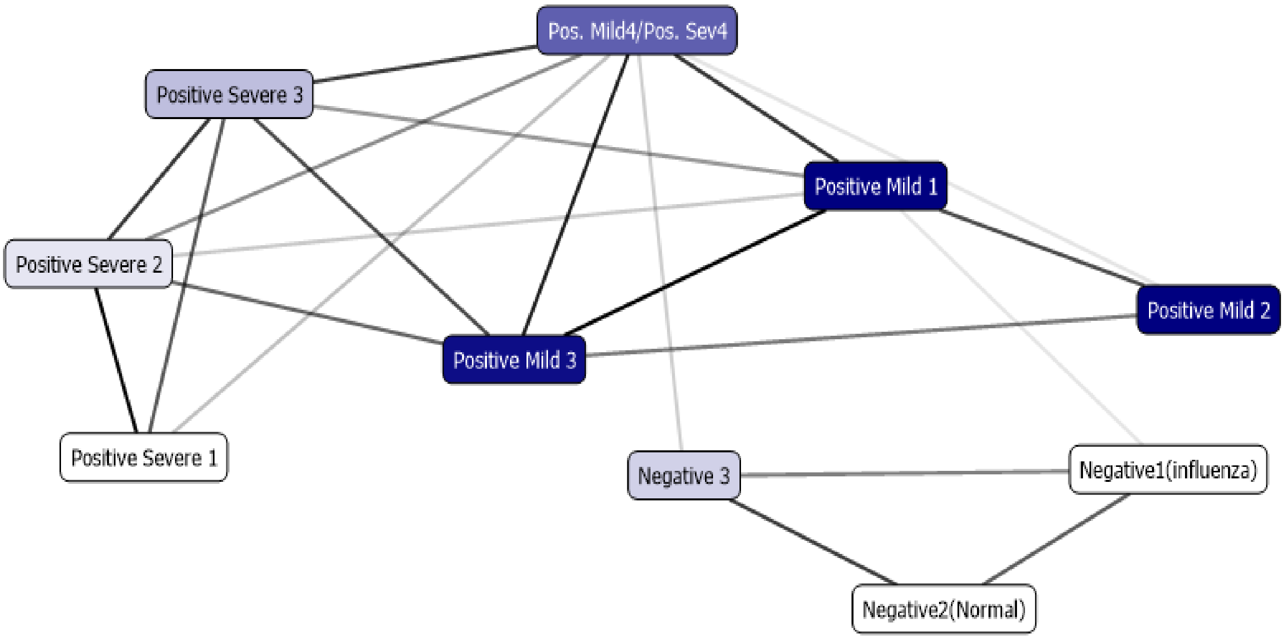


Fig. 3: Generated Soft Clusters Diagram from The Soft Clustering Mining Model (According to Positive Mild state 1 Distribution)

The implemented soft clustering mining model generated 10 clusters profiles for each one of all the COVID-19 diagnostic parameters and the diagnosis parameter with its states' distributions. Fig. 4 presents Part of the Generated COVID-19 Soft Clusters Profiles with States Distributions for each Variable. From Fig. 4, COVID-19 positive severe population distributions are found mainly in four clusters labeled Positive Severe 1, Positive Severe 2, and Positive Severe 3. COVID-19 positive mild population distributions are found also mainly in three clusters i.e., Positive Mild 1, Positive Mild 2, and Positive Mild 3, and so on. After reviewing the generated soft clusters, the importance of the 20 COVID-19 diagnostic parameters were confirmed in their strong relationships with the COVID-19 three grades diagnosis: Negative, Positive Mild, and Positive Severe. Fig. 5, Fig. 6, Fig. 7, and Fig. 8 present the main characteristics of the soft COVID-19 Negative1(Influenza) for patients' cases infected with Influenza and COVID-19 Negative 2(Normal) for normal cases, Positive Mild 1 and Positive Severe 1 with their associated probabilities. The attributes that the cluster contains are listed in the Variables column, and the state of the listed attribute is listed in the Values column.

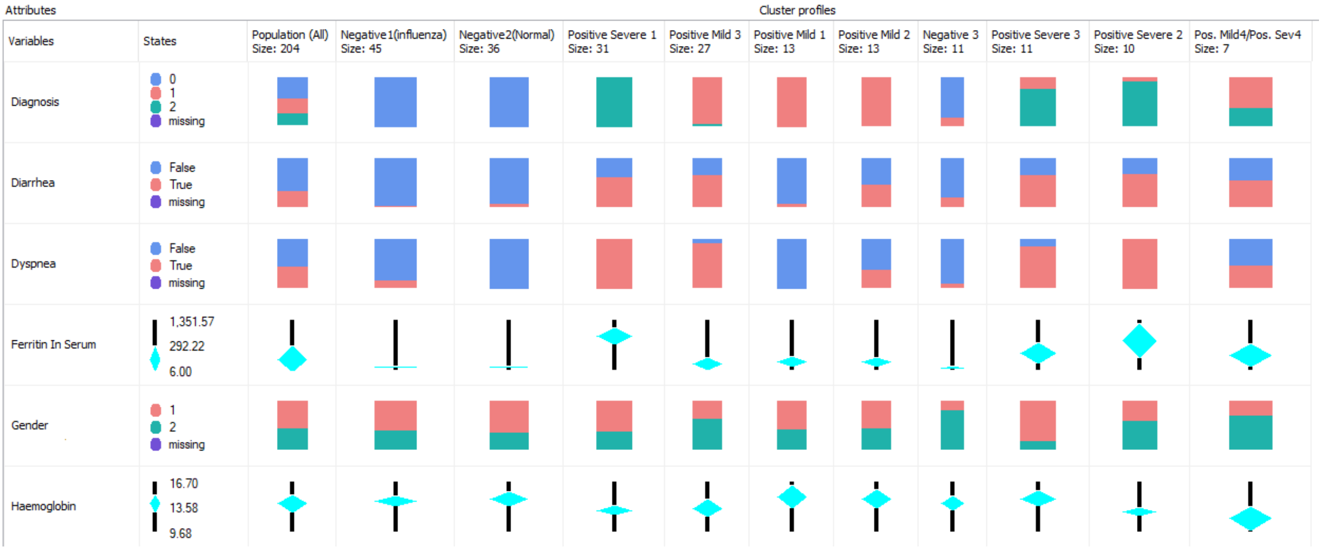
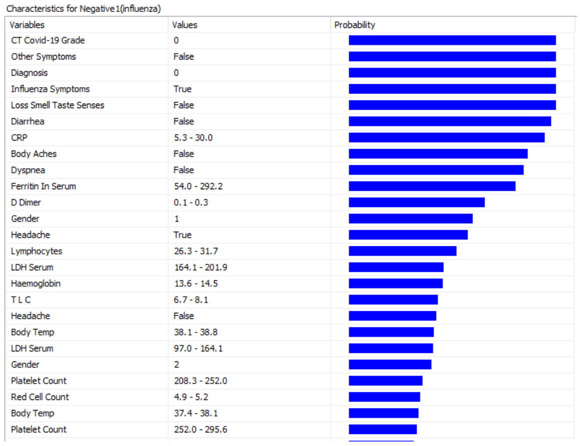
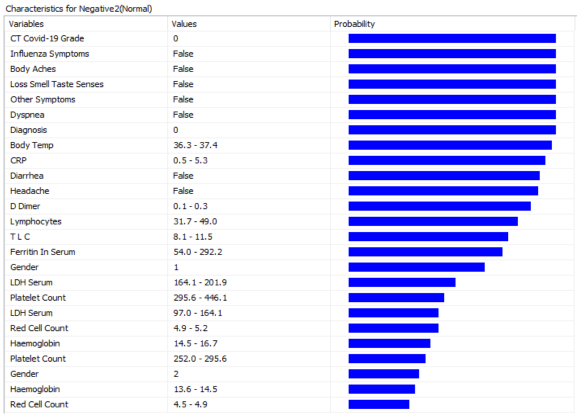


Fig. 4: Part of the Generated Soft Clusters Profiles with States Distributions for each Variable


Characteristics for Negative1(influenza)

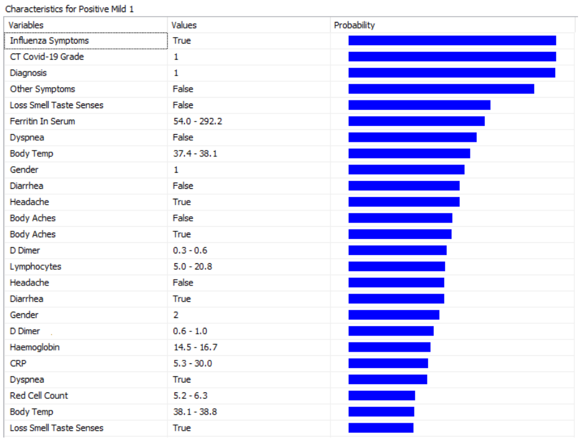
Variables	Values	Probability
CT Covid-19 Grade	0	
Other Symptoms	False	
Diagnosis	0	
Influenza Symptoms	True	
Loss Smell Taste Senses	False	
Diarrhea	False	
CRP	5.3 - 30.0	
Body Aches	False	
Dyspnea	False	
Ferritin In Serum	54.0 - 292.2	
D Dimer	0.1 - 0.3	
Gender	1	
Headache	True	
Lymphocytes	26.3 - 31.7	
LDH Serum	164.1 - 201.9	
Haemoglobin	13.6 - 14.5	
T L C	6.7 - 8.1	
Headache	False	
Body Temp	38.1 - 38.8	
LDH Serum	97.0 - 164.1	
Gender	2	
Platelet Count	208.3 - 252.0	
Red Cell Count	4.9 - 5.2	
Body Temp	37.4 - 38.1	
Platelet Count	252.0 - 295.6	

Fig. 5: Soft Cluster Negative 1 (Influenza) Characteristics


Characteristics for Negative2(Normal)

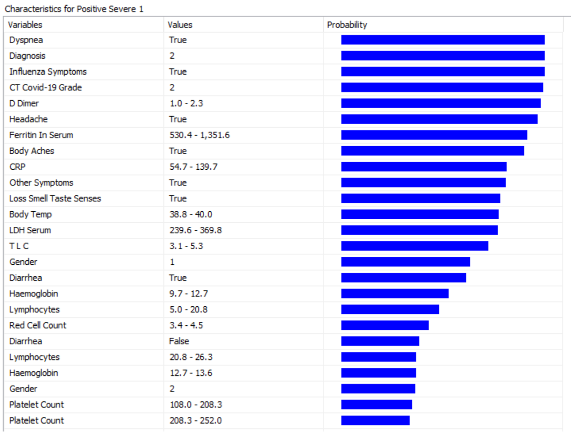
Variables	Values	Probability
CT Covid-19 Grade	0	
Influenza Symptoms	False	
Body Aches	False	
Loss Smell Taste Senses	False	
Other Symptoms	False	
Dyspnea	False	
Diagnosis	0	
Body Temp	36.3 - 37.4	
CRP	0.5 - 5.3	
Diarrhea	False	
Headache	False	
D Dimer	0.1 - 0.3	
Lymphocytes	31.7 - 49.0	
T L C	8.1 - 11.5	
Ferritin In Serum	54.0 - 292.2	
Gender	1	
LDH Serum	164.1 - 201.9	
Platelet Count	295.6 - 446.1	
LDH Serum	97.0 - 164.1	
Red Cell Count	4.9 - 5.2	
Haemoglobin	14.5 - 16.7	
Platelet Count	252.0 - 295.6	
Gender	2	
Haemoglobin	13.6 - 14.5	
Red Cell Count	4.5 - 4.9	

Fig. 6.: Soft Cluster Negative 2 (Normal) Characteristics


Characteristics for Positive Mild 1

Variables	Values	Probability
Influenza Symptoms	True	
CT Covid-19 Grade	1	
Diagnosis	1	
Other Symptoms	False	
Loss Smell Taste Senses	False	
Ferritin In Serum	54.0 - 292.2	
Dyspnea	False	
Body Temp	37.4 - 38.1	
Gender	1	
Diarrhea	False	
Headache	True	
Body Aches	False	
Body Aches	True	
D Dimer	0.3 - 0.6	
Lymphocytes	5.0 - 20.8	
Headache	False	
Diarrhea	True	
Gender	2	
D Dimer	0.6 - 1.0	
Haemoglobin	14.5 - 16.7	
CRP	5.3 - 30.0	
Dyspnea	True	
Red Cell Count	5.2 - 6.3	
Body Temp	38.1 - 38.8	
Loss Smell Taste Senses	True	

Fig. 7. Soft Cluster Positive Mild 1 Characteristics


Characteristics for Positive Severe 1

Variables	Values	Probability
Dyspnea	True	
Diagnosis	2	
Influenza Symptoms	True	
CT Covid-19 Grade	2	
D Dimer	1.0 - 2.3	
Headache	True	
Ferritin In Serum	530.4 - 1,351.6	
Body Aches	True	
CRP	54.7 - 139.7	
Other Symptoms	True	
Loss Smell Taste Senses	True	
Body Temp	38.8 - 40.0	
LDH Serum	239.6 - 369.8	
T L C	3.1 - 5.3	
Gender	1	
Diarrhea	True	
Haemoglobin	9.7 - 12.7	
Lymphocytes	5.0 - 20.8	
Red Cell Count	3.4 - 4.5	
Diarrhea	False	
Lymphocytes	20.8 - 26.3	
Haemoglobin	12.7 - 13.6	
Gender	2	
Platelet Count	108.0 - 208.3	
Platelet Count	208.3 - 252.0	

Fig. 8. Soft Cluster Positive Severe 1 Characteristics

The implemented soft clustering mining model successfully predicted accurately the relevant clusters for the different grades of COVID-19 diagnosis: Negative, Positive Mild, Positive Severe with their different states.

4.2 Proposed Backpropagation based MLP Multiclass ANN Mining Model for COVID-19 Diagnosis Prediction Results

After performing training and testing phases, the implemented proposed Backpropagation based MLP Multiclass ANN mining model for COVID-19 diagnosis prediction has generated very important results concerning the Negative, Positive Mild, and Positive Severe states (classes) for the diagnosis parameter. which are summarized in Table 4. Note that these results are related to the used relational dataset in training the model.

Table 4. Backpropagation based MLP ANN Mining Model for COVID-19 Diagnosis Prediction Important Results

Diagnosis State Parameter Value(s)	Covid-19 Negative (Normal)	Covid-19 Negative (Influenza)	Covid-19 Positive Mild	Covid-19 Positive Severe
(HB)	13.5-16.7	13.5-16.7	13.5 – 16.7	9.95 – 13.5
(RBC)	4.8 - 6.3	4.8 - 6.3	4.8 - 6.3	3.4 – 4.5
(PC)	201 - 444	201-444	108 - 290	108 - 290
(T.L.C.)	7.9- 11.0	5.1 – 7.9	5.1 – 7.9	3.2 – 5.1
Lymphocytes%	31.3 – 43	20.5 - 31.3	3.9 – 20.5	3.9 – 20.5
Ferritin in serum (F)	6 – 81	81 - 331	331 - 580	580 – 1439.7
(LDH) serum	97 – 166	166 – 206	206 - 380	206 - 380
(CRP)	0.5 - 7	7 - 35	35 - 64	64 – 163
D Dimer	0.1 – 0.29	0.29 – 0.6	0.689 – 1.09	1.09 – 2.5
CT COVID-19 grade	0	0	1	2
Body Temp (BT)	36.3 – 37.4	37.4 – 38.1	38.1 – 38.9	38.9 – 40
Headache	F (0)	F (0) / T(1)	F (0) / T (1)	T (1)
Body aches	F (0)	F(0)/ T(1)	F (0) / T (1)	T (1)
Loss of Smell & Taste senses	F (0)	F(0)	T (1)	T (1)
Diarrhea	F (0)	F(0)	F (0) / T (1)	F (0) / T (1)
Dyspnea	F (0)	F (0)	F (0) / T (1)	T (1)
Influenza Like Symptoms	F (0)	T (1)	T (1)	T (1)
Other Symptoms	F (0)	F (0)	F (0) / T (1)	T (1)
Gender	1 , 2	1 , 2	1 , 2	1 , 2
Age	9 - 74	9 – 74	4 - 80	23 – 80

4.3 Proposed Improved Apriori Algorithm based Association Rules (AR) mining model for COVID-19 Diagnosis Prediction Results

After performing training and testing phases, the implemented Proposed Improved Apriori Algorithm based Association Rules (AR) mining model for COVID-19 Diagnosis Prediction has generated the three dependency networks in Fig.9, Fig. 10, and Fig. 11 for the Negative, Positive Mild, and Positive severe states (classes) of the Diagnosis parameter.

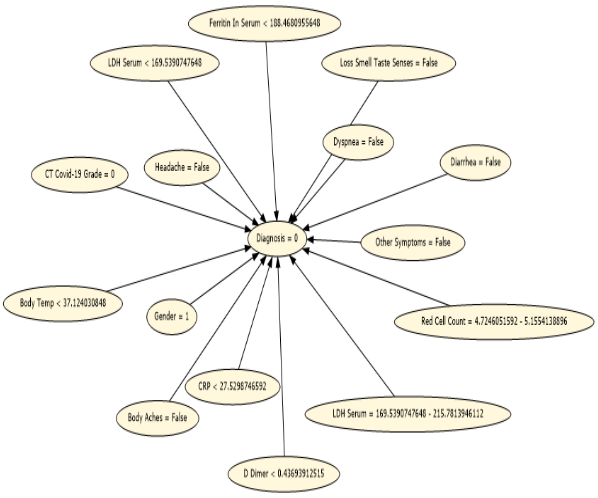


Fig. 9. COVID-19 Negative Dependency Network

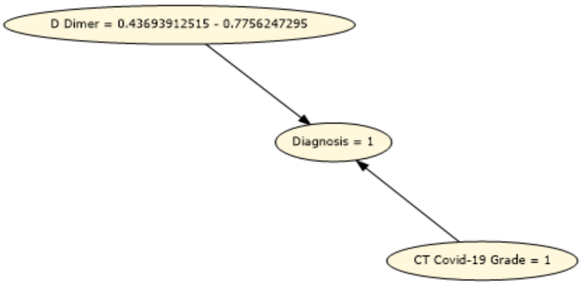


Fig. 10. COVID-19 Positive Mild Dependency Network

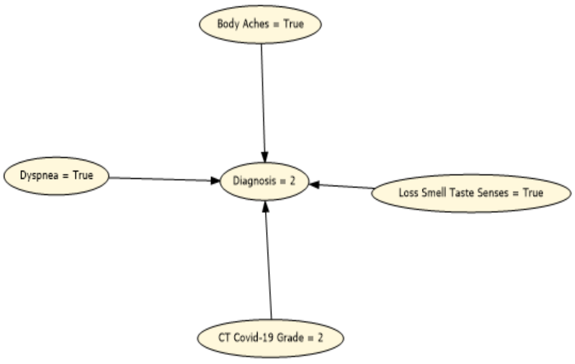
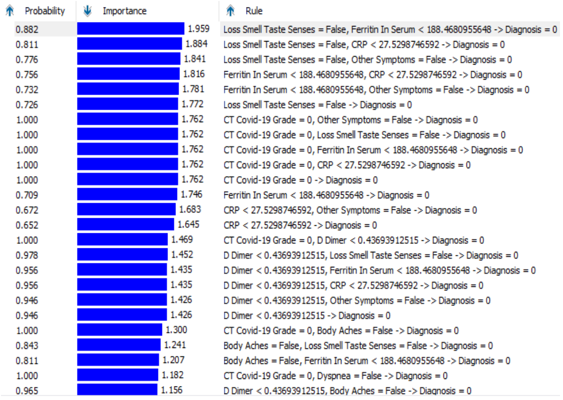


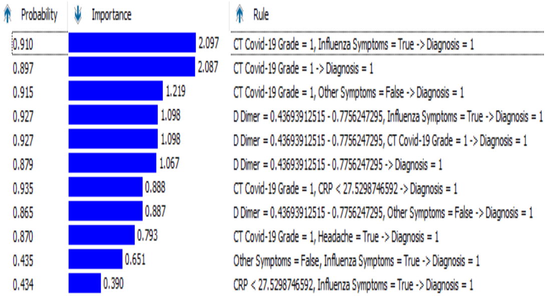
Fig. 11. COVID-19 Positive Severe Dependency Network

In addition, Fig. 12, Fig. 13, and Fig. 14 present the top generated association rules sorted in descending order according to importance related to diagnosing COVID-19 Negative (0), Positive Mild (1), and Positive Severe (3) states consecutively.



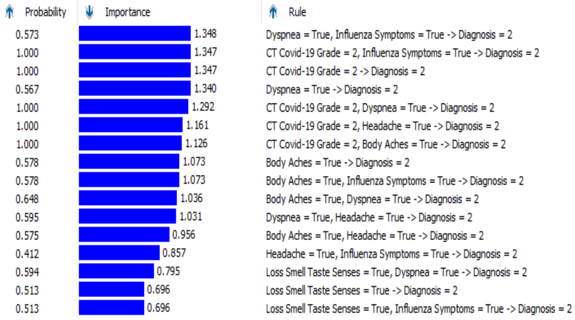
Probability	Importance	Rule
0.882	1.959	Loss Smell Taste Senses = False, Ferritin In Serum < 188.4680955648 -> Diagnosis = 0
0.811	1.884	Loss Smell Taste Senses = False, CRP < 27.5298746592 -> Diagnosis = 0
0.776	1.841	Loss Smell Taste Senses = False, Other Symptoms = False -> Diagnosis = 0
0.756	1.816	Ferritin In Serum < 188.4680955648, CRP < 27.5298746592 -> Diagnosis = 0
0.732	1.781	Ferritin In Serum < 188.4680955648, Other Symptoms = False -> Diagnosis = 0
0.726	1.772	Loss Smell Taste Senses = False -> Diagnosis = 0
1.000	1.762	CT Covid-19 Grade = 0, Other Symptoms = False -> Diagnosis = 0
1.000	1.762	CT Covid-19 Grade = 0, Loss Smell Taste Senses = False -> Diagnosis = 0
1.000	1.762	CT Covid-19 Grade = 0, Ferritin In Serum < 188.4680955648 -> Diagnosis = 0
1.000	1.762	CT Covid-19 Grade = 0, CRP < 27.5298746592 -> Diagnosis = 0
1.000	1.762	CT Covid-19 Grade = 0 -> Diagnosis = 0
0.709	1.746	Ferritin In Serum < 188.4680955648 -> Diagnosis = 0
0.672	1.683	CRP < 27.5298746592, Other Symptoms = False -> Diagnosis = 0
0.652	1.645	CRP < 27.5298746592 -> Diagnosis = 0
1.000	1.469	CT Covid-19 Grade = 0, D Dimer < 0.43693912515 -> Diagnosis = 0
0.978	1.452	D Dimer < 0.43693912515, Loss Smell Taste Senses = False -> Diagnosis = 0
0.956	1.435	D Dimer < 0.43693912515, Ferritin In Serum < 188.4680955648 -> Diagnosis = 0
0.956	1.435	D Dimer < 0.43693912515, CRP < 27.5298746592 -> Diagnosis = 0
0.946	1.426	D Dimer < 0.43693912515, Other Symptoms = False -> Diagnosis = 0
0.946	1.426	D Dimer < 0.43693912515 -> Diagnosis = 0
1.000	1.300	CT Covid-19 Grade = 0, Body Aches = False -> Diagnosis = 0
0.843	1.241	Body Aches = False, Loss Smell Taste Senses = False -> Diagnosis = 0
0.811	1.207	Body Aches = False, Ferritin In Serum < 188.4680955648 -> Diagnosis = 0
1.000	1.182	CT Covid-19 Grade = 0, Dyspnea = False -> Diagnosis = 0
0.965	1.156	D Dimer < 0.43693912515, Body Aches = False -> Diagnosis = 0

Fig. 12. COVID-19 Negative Association Rules



Probability	Importance	Rule
0.910	2.097	CT Covid-19 Grade = 1, Influenza Symptoms = True -> Diagnosis = 1
0.897	2.087	CT Covid-19 Grade = 1 -> Diagnosis = 1
0.915	1.219	CT Covid-19 Grade = 1, Other Symptoms = False -> Diagnosis = 1
0.927	1.098	D Dimer = 0.43693912515 - 0.7756247295, Influenza Symptoms = True -> Diagnosis = 1
0.927	1.098	D Dimer = 0.43693912515 - 0.7756247295, CT Covid-19 Grade = 1 -> Diagnosis = 1
0.879	1.067	D Dimer = 0.43693912515 - 0.7756247295 -> Diagnosis = 1
0.935	0.888	CT Covid-19 Grade = 1, CRP < 27.5298746592 -> Diagnosis = 1
0.865	0.887	D Dimer = 0.43693912515 - 0.7756247295, Other Symptoms = False -> Diagnosis = 1
0.870	0.793	CT Covid-19 Grade = 1, Headache = True -> Diagnosis = 1
0.435	0.651	Other Symptoms = False, Influenza Symptoms = True -> Diagnosis = 1
0.434	0.390	CRP < 27.5298746592, Influenza Symptoms = True -> Diagnosis = 1

Fig. 13. COVID-19 Positive Mild Association Rules



Probability	Importance	Rule
0.573	1.348	Dyspnea = True, Influenza Symptoms = True -> Diagnosis = 2
1.000	1.347	CT Covid-19 Grade = 2, Influenza Symptoms = True -> Diagnosis = 2
1.000	1.347	CT Covid-19 Grade = 2 -> Diagnosis = 2
0.567	1.340	Dyspnea = True -> Diagnosis = 2
1.000	1.292	CT Covid-19 Grade = 2, Dyspnea = True -> Diagnosis = 2
1.000	1.161	CT Covid-19 Grade = 2, Headache = True -> Diagnosis = 2
1.000	1.126	CT Covid-19 Grade = 2, Body Aches = True -> Diagnosis = 2
0.578	1.073	Body Aches = True -> Diagnosis = 2
0.578	1.073	Body Aches = True, Influenza Symptoms = True -> Diagnosis = 2
0.648	1.036	Body Aches = True, Dyspnea = True -> Diagnosis = 2
0.595	1.031	Dyspnea = True, Headache = True -> Diagnosis = 2
0.575	0.956	Body Aches = True, Headache = True -> Diagnosis = 2
0.412	0.857	Headache = True, Influenza Symptoms = True -> Diagnosis = 2
0.594	0.795	Loss Smell Taste Senses = True, Dyspnea = True -> Diagnosis = 2
0.513	0.696	Loss Smell Taste Senses = True -> Diagnosis = 2
0.513	0.696	Loss Smell Taste Senses = True, Influenza Symptoms = True -> Diagnosis = 2

Fig. 14. COVID-19 Positive Severe Association Rules

4.4 Proposed Hybrid Machine Learning Model for Predicting COVID-19 Disease Diagnosis Evaluation

The implemented hybrid ML model was evaluated using different evaluation measures for the used predicting mining models. These used evaluation measures are derived from the generated confusion matrix related to each mining model which are presented in Table 5 and Table 6. It must be noted that the built confusion matrix in our proposed ML model is for 3-class classification (multi-class classification problem) with 3 classes: Negative (0), Positive Mild(1), and Positive Severe(3) and generated for the testing dataset (87 data samples) which represents 30% of the total dataset. Unlike binary classification, there is no direct positive or negative classes here. As a result, in this case TP, TN, FP and FN for each individual class are calculated first, then each mining model accuracy, precision, recall(sensitivity), specificity, micro F1 score, and Matthews correlation coefficient (MCC) are calculated as follows: *Accuracy* formula is (6) listed in section 3.3.

$$Precision = \sum_{m=1}^{3} TP_m \;/(\sum_{m=1}^{3} TP_m \;+ \sum_{m=1}^{3} FP_m \;) \qquad (9)$$

$$Recall(Sensitivity) = \sum_{m=1}^{3} TP_m \;/\; (\sum_{m=1}^{3} TP_m \;+ \sum_{m=1}^{3} FN_m \;) \qquad (10)$$

$$Specificity = \sum_{m=1}^{3} TN_m \;/(\sum_{m=1}^{3} TN_m \;+ \sum_{m=1}^{3} FP_m \;) \qquad (11)$$

$$Micro\ F1 - score = \frac{2*Total\ Recall*Total\ Precision}{Total\ Recall+Total\ Precision} \qquad (12)$$

$$\text{Matthews } correlation\ coefficient\ (MCC) = \frac{\sum_{m=1}^{3} TP_m * \sum_{m=1}^{3} TN_m \; - \sum_{m=1}^{3} FP_m * \sum_{m=1}^{3} FN_m}{\sqrt{(\sum_{m=1}^{3} TP_m + \sum_{m=1}^{3} FP_m)(\sum_{m=1}^{3} TP_m + \sum_{m=1}^{3} FN_m)(\sum_{m=1}^{3} TN_m + \sum_{m=1}^{3} FP_m)(\sum_{m=1}^{3} TN_m + \sum_{m=1}^{3} FN_m)}} \qquad (13)$$

Where TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative, and m: It takes value from 1 to 3 representing three diagnosis classes (Negative(0), Positive Mild(1), Positive Severe(3)).

Table 5. Confusion Matrix for Proposed Backpropagation based MLP ANN Mining Model

Predicted	0 (Actual)	1 (Actual)	2 (Actual)	Sum
0	37	1	0	38
1	1	26	0	27
2	1	1	20	22
Sum	39	28	20	

Table 6. Confusion Matrix for Proposed Improved Apriori Algorithm based Association Rules (AR) Mining Model

Predicted	0 (Actual)	1 (Actual)	2 (Actual)	Sum
0	37	2	0	39
1	0	21	1	22
2	0	0	26	26
Sum	37	23	27	

Table. 7 presents the calculated evaluation measures for the proposed two COVID-19 prediction mining models, and the evaluation measures average for the proposed hybrid machine learning model as a whole. From Table. 7 we can conclude that the proposed hybrid machine learning model has successfully not only classified COVID-19 patients into positive mild and positive severe patients, but also discriminated successfully between COVID-19 patients and Influenza patients/normal cases with an average accuracy of 97.3%, an average precision 96%, an average recall (sensitivity) 96%, an average specificity 98%, and average micro F1 score 96% on an independent testing dataset of 87 patients. Unlike the other evaluation metrics, MCC takes all the cells of the Confusion Matrix into consideration in its formula. The range of values of MCC lie between -1 to +1. A model with a score of +1 is a perfect model and -1 is a poor model. The average MCC of the hybrid model is 0.94 which means that the proposed model achieved a high degree of perfection.

Table. 8 and Fig. 15 show a comparison between the proposed model and the state-of-the art models regarding to the model accuracy, sensitivity, and specificity. The results show that the proposed model outperforms the other reviewed COVID-19 diagnosis prediction models.

Table 7. The Proposed Hybrid Machine Learning Evaluation Measures

Measure Mining Model	Accuracy	Precision	Recall/ Sensitivity	Specificity	Micro F1 Score	Matthews correlation coefficient (MCC)
Backpropagation based MLP ANN Mining Model for COVID-19 Diagnosis Prediction	96.9 %	95.4%	95.4%	97.7%	95.4%	0.93
Association Rules mining model based on Improved Apriori algorithm for COVID-19 Diagnosis Prediction	97.7 %	96.6%	96.6%	98.3%	96.6%	0.95
The proposed hybrid machine learning model evaluation measures Average	97.3%	96%	96%	98%	96%	0.94

Table 8. COVID-19 Diagnosis Models Evaluation Measures Comparison

	Accuracy%	sensitivity%	specificity%
COVID-19 Diagnosis proposed Model	97.3	96	98
Banerjee et al. Model	87.00	65.00	91.00
Bao et al. Model	84.00	88.00	80.00
Barbosa et al. Model	95.20	96.80	93.60
Nan et al. Model	91.00	87.00	95.00

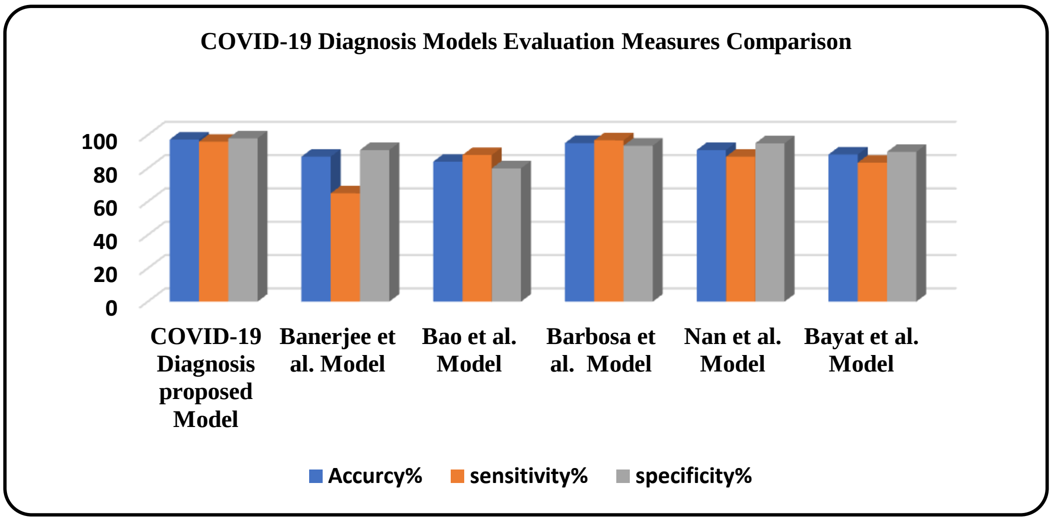


Fig. 15. COVID-19 Diagnosis Models Evaluation Measures Comparison Graph

5. CONCLUSION AND FUTURE WORK

In this study, a hybrid two-phase machine learning model for COVID-19 diagnosis prediction was proposed. The aim of this proposed model is to early diagnose COVID-19 based on available laboratory tests results, clinical symptoms, CT results, and demographic data in case of the difficulty of applying or absence of PCR test. The dataset consisted of 291 data samples (127 COVID-19 Negative Patients include 68 Influenza patients and 59 normal cases, 84 COVID-19 confirmed Positive Mild patients, and 80 COVID-19 confirmed Positive Severe patients). The soft clustering mining model was able to successfully identify the most relevant identifying clusters characteristics for the COVID-19 disease grades and confirmed the selected 20 features strong relationships with the COVID-19 three grades diagnosis: Negative, Positive Mild, and Positive Severe. The implementation of the two proposed ML mining models for COVID-19 diagnosis achieved an overall accuracy 97.3%, sensitivity 96%, and specificity 98%. These achieved results outperformed the other reviewed state-of-the art COVID-19 diagnosis prediction models, taking into consideration the proposed unique features combination used efficiently for COVID-19 diagnosis in his study. Future work will focus on integrating this proposed hybrid ML model with an intelligent COVID-19 treatment protocols advisor system according to the disease grade, patient history, clinical signs and symptoms, and physical examination.

6. REFERENCES

[1] Sohrabi, C., AlsafiZ OfiNeill, N., Khan, M. and Kerwan A. et al. 2020. World health organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery. vol. 76, 71-76.

[2] Barbat, MM., Wesche, C., Werhli, AV., and Mata, MM. 2019. An adaptive machine learning approach to improve automatic iceberg detection from SAR images . ISPRS Journal of Photogrammetry and Remote Sensing. vol. 156, 247–259.

[3] Shang, R., Qi, L., Jiao, L., Stolkin, R., and Li, Y. 2014. Change detection in SAR images by artificial immune multi-objective clustering. Engineering Applications of Artificial Intelligence. vol. 31, 53–67.

[4] Choi, S., Kang, M-G., Min, H., Chang, Y-S., and Yoon, S. 2017. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. Methods. vol. 129, 50–59.

[5] Vaka, AR., and Soni, B. 2020. Breast cancer detection by leveraging Machine Learning. ICT Express. vol. 6. no. 4, 320-324.

[6] Saxena, S., and Gyanchandani, M. 2019. Machine learning methods for computer-aided breast cancer diagnosis using histopathology: a narrative review. Journal of Medical Imaging and Radiation Science. vol. 51. no. 1, 182-193.

[7] Vaishya, R., Javaid, M., Khan, IH., and Haleem, A. 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. vol.14. no. 4, 337–339.

[8] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, “Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases”, Radiology, vol. 296, no. 2, pp. 32-40, 2020.

[9] Javor, D., Kaplan, A., Puchner, S.B., Krestan, C., and Baltzer, P. 2020. Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. European Journal of Radiology. vol. 133. no. 109402, 2020.

[10] Goodman-Meza, D., Adamson, PC., and Ebinger, J. et al. 2020. A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. PLOS ONE Journal. vol. 15. no. 9, e0239474.

[11] An, C., Lim, H., and Kim, DW. et al. 2020. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study . Scientific Reports Journal. Vol. 10. no. 18716.

[12] Khanday, A.M.U.D., Rabani, S.T., and Khan, Q.R., et al. 2020. Machine learning based approaches for detecting COVID-19 using clinical text data. Int. Journal of Information Technology. vol. 12. 731–739.

[13] Banerjee, A., Ray, S., Vorselaars, B., and Kitson, J. et al. 2020. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. Int.

immunopharmacology. vol. 86. no. 106705.

[14] Bao, FS., He, Y., Liu, J., Chen, Y., Li, Q., and Zhang, CR. et al. 2020. Triaging moderate covid-19 and other viral pneumonias from routine blood tests. arXiv e-prints. arXiv:2005.06546.

[15] Barbosa, V., Gomes, JC., Santana, MA., and Almeida Albuquerque, JE. et al. Heg. ia: an intelligent system to support diagnosis of covid-19 based on blood tests. medRxiv, 2020.

[16] Nan, SN., Ya, Y., Ling, TL., and Nv, GH., et al. A prediction model based on machine learning for diagnosing the early covid-19 patients. medRxiv, 2020.

[17] Bayat, V., Phelps, S., and Ryono, R., et al. A Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Prediction Model from Standard Laboratory Tests, Clinical Infectious Diseases. ciaa1175. 2020.

[18] Zoabi, Y., Deri-Rozov, S., and Shomron, N. 2021. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medical. vol. 4. no. 3.

[19] Prokop, M., van Everdingen, W., and Van Rees, VT. et al. 2020. CO- RADS – a categorical CT assessment scheme for patients with suspected COVID- 19: definition and evaluation. Radiology. vol. 296. no. 2, 97-104.

[20] Bradley P., Fayyad, U., and Reina, C. 1998. Scaling EM (Expectation-Maximization) Clustering to Large Databases. MSR-TR-98-3.1998.