The Prediction of COVID 19 Disease Using Feature Selection Techniques

The Prediction of COVID 19 Disease Using Feature Selection Techniques

Rasha H. Ali[1] and Wisal Hashim Abdulsalam[2]

[1]Computer Science Department/College of Education for Women/University of Baghdad, Iraq
[2]Computer Science Department/College of Education for Pure Science/Ibn-Al Haitham/University of Baghdad, Iraq

E-mail: wisal.h@ihcoedu.uobaghdad.edu.iq

Abstract. COVID 19 has spread rapidly around the world due to the lack of a suitable vaccine; therefore the early prediction of those infected with this virus is extremely important attempting to control it by quarantining the infected people and giving them possible medical attention to limit its spread. This work suggests a model for predicting the COVID 19 virus using feature selection techniques. The proposed model consists of three stages which include the preprocessing stage, the features selection stage, and the classification stage. This work uses a data set consists of 8571 records, with forty features for patients from different countries. Two feature selection techniques are used in order to select the best features that affect the prediction of the proposed model. These are the Recursive Feature Elimination (RFE) as wrapper feature selection and the Extra Tree Classifier (ETC) as embedded feature selection. Two classification methods are applied for classifying the features vectors which include the Naïve Bayesian method and Restricted Boltzmann Machine (RBM) method. The results were 56.181%, 97.906% respectively when classifying all features and 66.329%, 99.924% respectively when classifying the best ten features using features selection techniques.

Keywords: Feature selection, COVID 19, Recursive Feature Elimination, Extra Tree Classifier, Restricted Boltzmann Machine, Naïve Bayesian.

1. Introduction

COVID 19 is an infectious disease that spreads through the air. Besides, it can live on rooftops for about two days, and it was initially called the Wuhan virus because of its origin from Wuhan, China, where the first human case was recorded [3]. Then the disease spreads rapidly and within a month turned into a pandemic affecting people all over the world in a negatively way. This prompted the World Health Organization (WHO) to declare a general global health emergency by the end of January 2020. Symptoms may appear on the infected person or not, but in either cases, the infected person can spread the virus. The incubation period ranges from two days to two weeks. The symptoms are different, such as fever,

shortness of breath, cough, chest pain, loss of sense of smell and taste, vomiting, headache, nausea, laziness, and others [4, 5].
In general, there is no vaccine available for **COVID 19** and information about it remains limited. There are several clinical trials to understand how the virus could be redeveloped in an attempt to produce a vaccine [6].
To limit the virus spread, governments around the world have taken various measures such as closing borders and commercial complexes, suspending education in universities and schools and converting them to e-learning, and suspending gatherings, etc., which affected various aspects of life, especially the economic aspect. People have been made aware of the importance of practicing simple steps such as staying home, regularly washing and sterilizing hands, social distancing, and wearing masks. Figure (1) gives **COVID 19** cases distribution worldwide, in November 2020 [2].

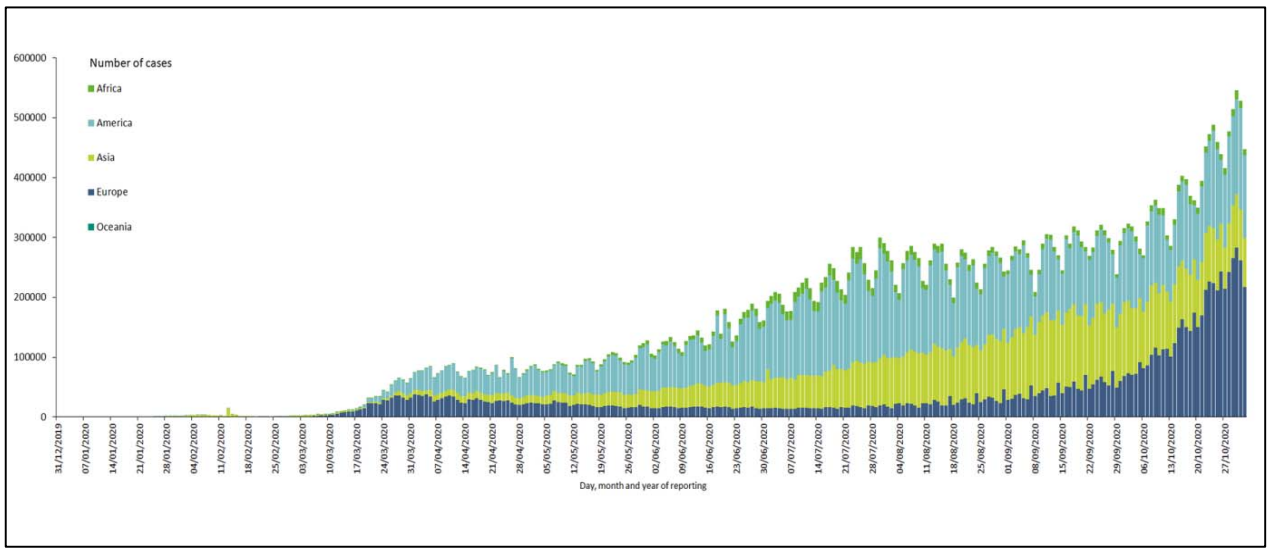

Fig. (1) **COVID 19** cases distribution worldwide, in November 2020 [2].

The goal of this work is to suggest a model for predicting the disease of **COVID 19** using feature selection techniques. The feature selection techniques are used for selecting the important features which affect the accuracy of the classification for the model.
The main contribution of this work is the prediction of the COVID 19 disease using two types of feature selection techniques, which are REF as wrapper feature selection method and ETC as embedded feature selection method and comparing between them by the results in order to select the most important features, then extracted the optimal number of features in the dataset contained different information, different types of data (text, number, and date), missing value, and large numbers. Two types of classification techniques are used including the Naïve Bayesian and RBM which are applied on different number of features for comparing the results.
This paper is ordered as follows: Section 2 presents the literature review. Section 3 describes the proposed work including the dataset used, basic preprocessing operations, feature selection, and classification. Section 4 presents the results and discussion while section 5 gives the conclusions.

2. Literature Review

This section presents the work related to **COVID 19** prediction based on feature selection techniques.

Warda M. Shaban et al. [5] suggested a strategy to diagnosis **COVID 19** using a hybrid feature selection that used filter method as a rapid feature selection and then used genetic algorithm as a wrapper method to select the most important features from those extracted from chest Computed Tomography (CT) images and then an Enhanced K-Nearest Neighbour (EKNN) classifier was used. This method achieved 96% of accuracy.

Liang Sun et al. [7] proposed an Adaptive Feature Selection guided Deep Forest (AFS-DF) for **COVID 19** classification based on chest CT images. AFS-DF was evaluated on a dataset with 1495 patients of **COVID 19** and 1027 patients of community and gained an accuracy of 91.79%.

Bejoy Abraham, and Madhu S. Nair [8] investigated the effectiveness of a combination of several pre-trained Convolutional Neural Networks (CNNs) for **COVID 19** detection from X-ray images. The method used a combination of features extracted from multi-CNN with correlation based feature selection technique in combination with subset size forward selection, and a linear forward selection based search technique to determine the optimal feature subset and Bayesnet classifier. They achieved an accuracy of 91.16% when tested on a dataset with 453 **COVID 19** images and 497 non-COVID images, and 97.44% on a dataset consisting of 71 **COVID 19** images and 7 non-COVID images.

Mohammad Pourhomayoun, and Mahdi Shakibi [9] applied different filter and wrapper methods for feature selection to select 42 features out of 112 features. Then they used several machine-learning algorithms to predict mortality in patients with **COVID 19**. They used a dataset of more than 117,000 laboratory-confirmed **COVID 19** patients from76 countries. The Neural Network algorithm achieved the best performance and accuracy of 93%.

3. The Proposed Work The proposed work consists of three stages. These are pre-processing stage, features selection stage, and finally the classification stage. In each stage, more than one step and techniques are used for achieving the goal of each stage. Figure (2) shows the structure of the proposed model.

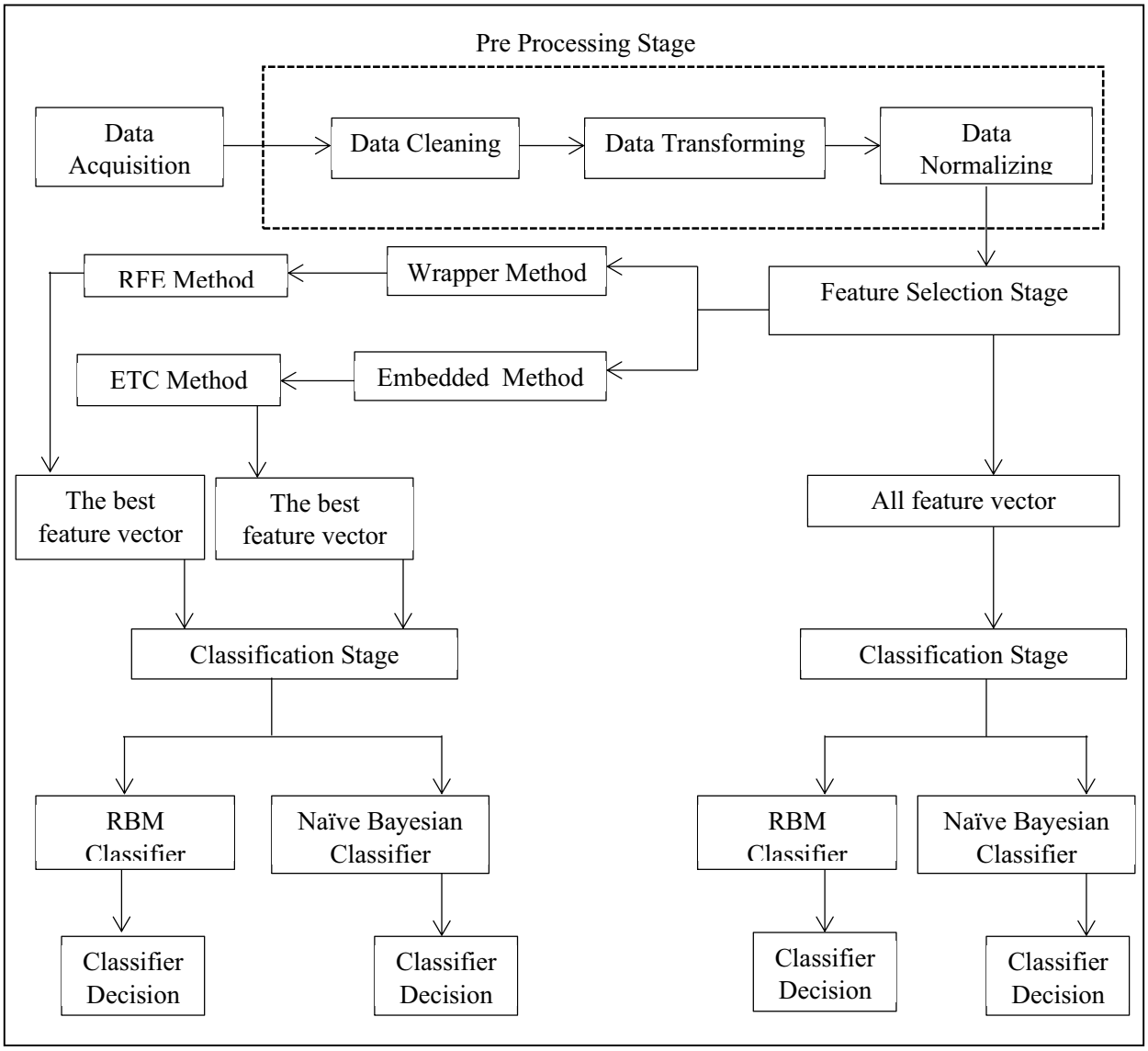

Fig. (2) The structure of the proposed model.

3.1 The Pre-processing Stage

This stage is also known as data preparation. It is a very important stage to build a good model and enhance the accuracy [10]. The dataset used in this work had been collecting by the WHO, which contained 8571 records from thirty-eight countries.

Figure (3) shows the number of records for each country in the dataset. The dataset included forty features with different types (time, text, and number), with type (CSV) file. The proposed model undergoes pre-processing in three steps, namely data cleaning, transforming and normalization

- **Data cleaning processing.** The data set may have insufficient data, missing data, too much data in rows or in columns, duplicate data and outlier values. In this work, the data has been cleaned through processing the missing value by fill each missing value with the nearest value.
- **Data transforming process** has been applied through transforming data into numeric forms.
- **Data normalizing** has been performed by applying min max scalar using the following equation

$$z = (x - \min) / (\max - \min) \quad \ldots(1)$$

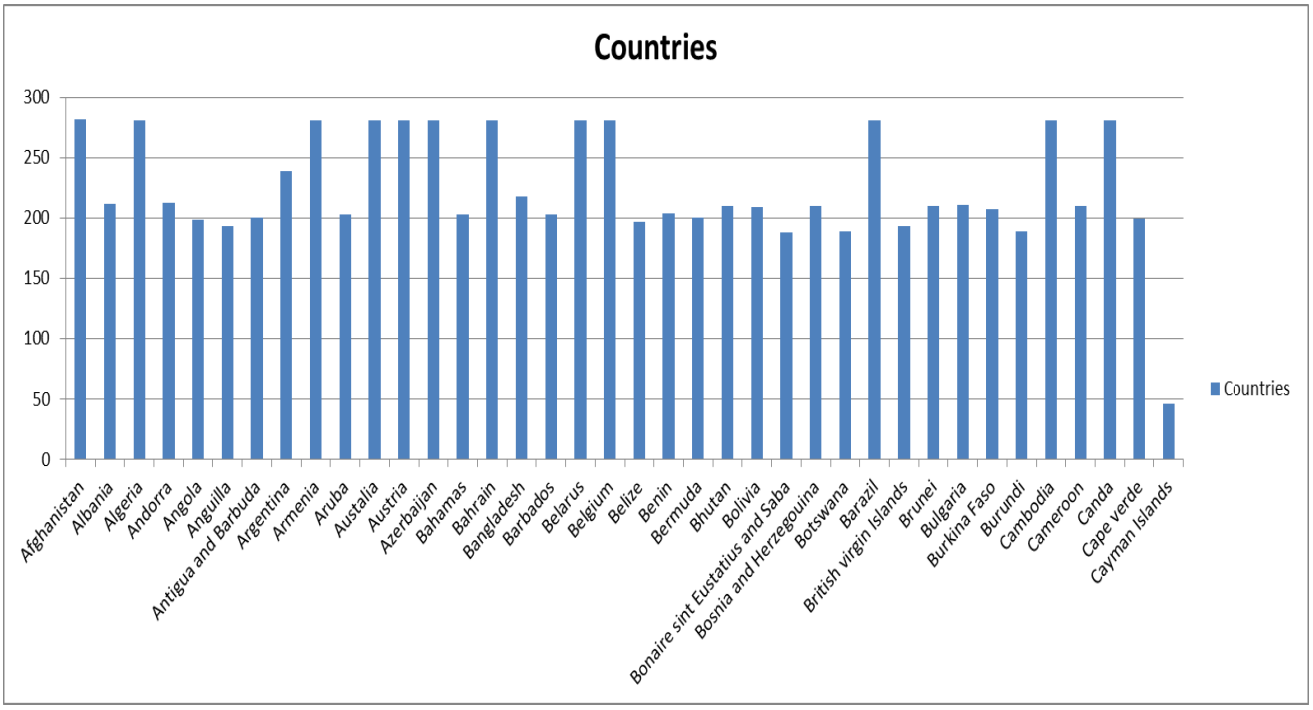

Fig. (3) The number of records and countries in the dataset.

3.2 The Feature Selection Stage

To improve learning performance, dimensionality reduction was used when there are a large number of input features, which divided into feature extraction and feature selection techniques. Feature extraction extracts a new set of features that is smaller than that resulting from feature selection, so the discriminative power is better. Because of that, it is more useful for signal processing, information retrieval, and image analysis, but a combination of features may have no physical meaning, so feature extraction is not a good approach in respect of transparency, interpretability, and readability. Feature

selection builds a subset of the original features. This is useful when interpretability knowledge extraction is crucial, as in medicine, and since this work is based on a medical problem, we will concentrate on feature selection. There are three types of feature selection: filters, wrappers, and embedded. Filters concentrate on the general characteristics of the data. So they are independent of any learning method, not computationally costly and have a good generalization capacity. Wrappers and embedded methods both are required for selecting features. For wrappers, an induction method evaluates candidate subsets of features. They are more computationally costly than filters but perform better. Embedded methods lie between filters and wrappers because the selection is part of the training process for the induction method. The search for the best subset of features is performed during the training of the classifier and, because of that; embedded methods are less computationally costly than wrappers. There is also a hybrid method, which combines different feature selection algorithms in a sequential manner [11].
In this work, the number of features in the dataset is forty. Table (1) presents the features of the dataset. The goal of this stage is selecting the important features which affect the prediction of the model. Two methods are used for selecting features: Recursive Feature Elimination (RFE) method and the Extra Tree Classifier (ETC) method. The RFE is one of the wrapper feature selection method. It works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance. It takes the model to be used and the number of required features as input and gives the ranking of all the variables, the value (one) being most important. It also gives its support, the value (True) being a relevant feature and (False) value being an irrelevant feature. Algorithm (1) shows the steps of the RFE.

Algorithm (1) The steps of the RFE method.

Step1:- Train the logistic regression model using training set.

Step2:- Calculate the performance of the model.

Step3:- Calculate the variable ranking.

Step4:- For each subset size S_i do

1. keep the S_i most important variable
2. preprocessing the data
3. train the model using S_i predictors
4. calculate the performance of the model.
5. recalculate the ranking for each predictors.

End

Step5:- Calculate the performance for all S_i

Step6:- Determine the number of the predictors.

Step7:- END.

The (ETC) method is an embedded feature selection method. It depends on decision trees where each decision tree is created from the unusual training sample. Then, at each test node, each tree is provided with a random sample of N features from the feature set from which each decision tree must select the best feature to split the data based on the Gini Index using equation (2) as a mathematical criterion [12]. This random sample of features leads to the creation of multiple de-correlated decision trees.

$$I_G = 1 - \sum_{j=1}^{c} p_j^2 \quad \ldots(2)$$

Table (1) The features of the dataset [1].

No.	Feature's Name	No.	Feature's Name
1	iso_code	21	new_tests_smoothed_per_thousand
2	Continent	22	tests_per_case
3	Location	23	positive_rate
4	Date	24	tests_units
5	total_cases	25	stringency_index
6	new_cases_smoothed	26	population
7	total_deaths	27	population_density
8	new_deaths	28	median_age
9	new_deaths_smoothed	29	aged_65_older
10	total_cases_per_million	30	aged_70_older
11	new_cases_per_million	31	gdp_per_capita
12	new_cases_smoothed_per_million	32	extreme_poverty
13	total_deaths_per_million	33	cardiovasc_death_rate
14	new_deaths_per_million	34	diabetes_prevalence
15	new_deaths_smoothed_per_million	35	female_smokers
16	new_tests	36	male_smokers
17	total_tests	37	handwashing_facilities
18	total_tests_per_thousand	38	hospital_beds_per_thousand
19	new_tests_per_thousand	39	life_expectancy
20	new_tests_smoothed	40	human_development_index

3.3 The Classification Stage

Also, it is known as decision making. The goal of this stage is to classify the feature vector [13]. In this stage, the feature vector is classified using two methods: the Naïve Bayesian (NB) method and Restricted Boltzmann Machine (RBM) method. NB is a simple classifier based on the Bayes theorem which assumes that the features are independent. This classifier has worked well in many real-world situations. It is a very fast and has a good performance. The main advantages of this classifier are the conditional independence assumption, which helps to obtain a quick classification, and the probabilistic hypotheses. Parameters of probability distributions are obtained from the training material [14].

RBM is an algorithm useful for dimensionality reduction, feature learning, regression, classification, and collaborative filtering. RBM is a two-layered stochastic recurrent neural network capable of unsupervised learning and feature extraction [15]. It is an energy based model that can be viewed as a single layer undirected neural network which consists of two binary layers, the visible layer, and hidden layers. There are no effects between nodes in a single layer of RBM[16].

4 The Results and Discussion

This paper suggests a model for predicting the **COVID 19** disease using features selection techniques and two methods for classifying the features vector. The implementation of the proposed work was achieved using Python (V.3.5) as programming language, Jet Brains Pycharm (V.2018.2) as framework. As described earlier, the suggested method consists of three main stages. Firstly, the data pre-processing which was contained more than one process because of the nature of the collected dataset which has missing values, extra data, and different types of data. Secondly, the features selection was achieved by using two models for selecting features (the wrapper method and the embedded method). In this stage, the number of features selected was: ten features and twenty features. The RFE is one of the wrapper methods which was applied for selecting the 10 features and 20 features. Figure (4) presents the sorting of ten important features using RFE. The true value means the relevant feature while the false value means the irrelevant feature. Also, it can be viewed as ranking; the value (one) is the most important feature while the remainder values mean the rank of each feature depending on its importance.

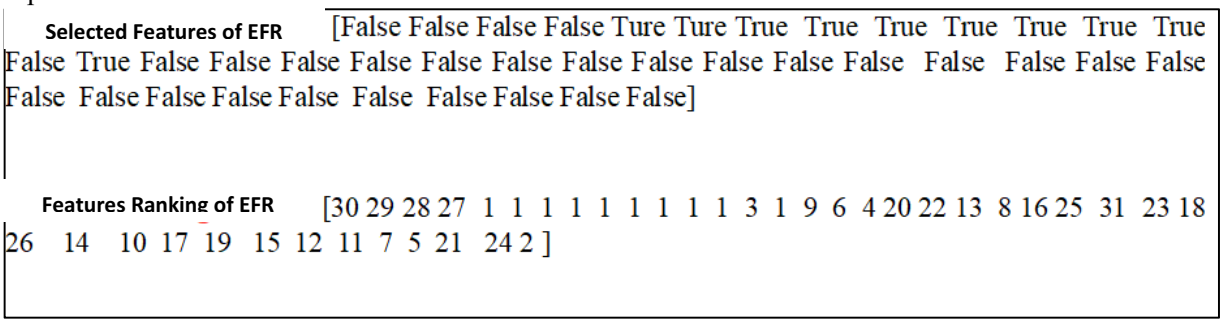

Fig. (4) The results of RFE method.

The second method applied for selecting features was ETC. Table (2) shows the values of features depend on their importance. While Figure (5) and (6) show the sorting of features depends on the importance of features for (10 most important features and 20 most important features respectively).

Table (2) The values of importance for features

Feature's Name	The value	Feature's Name	The value
iso_code	0.0000	new_tests_smoothed_per_thousand	0.017507
Continent	0.0000	tests_per_case	0.017407
Location	0.0000	positive_rate	0.015329
Date	0.0000	tests_units	0.0000
total_cases	0.075913	stringency_index	0.037782
New_deaths	0.040493	population	0.004229
new_cases_smoothed	0.075071	population_density	0.002263
total_deaths	0.047183	median_age	0.006317
new_deaths_smoothed	0.044856	aged_65_older	0.008087
total_cases_per_million	0.071144	aged_70_older	0.016779
new_cases_per_million	0.180063	gdp_per_capita	0.005155
new_cases_smoothed_per_million	0.076830	extreme_poverty	0.004050
total_deaths_per_million	0.047108	cardiovasc_death_rate	0.004266
new_deaths_per_million	0.036282	diabetes_prevalence	0.008391
new_deaths_smoothed_per_million	0.040993	female_smokers	0.002974
new_tests	0.016462	male_smokers	0.004460
total_tests	0.016644	handwashing_facilities	0.005238
total_tests_per_thousand	0.016404	hospital_beds_per_thousand	0.005043
new_tests_per_thousand	0.014537	life_expectancy	0.005591
new_tests_smoothed	0.019729	human_development_index	0.009421

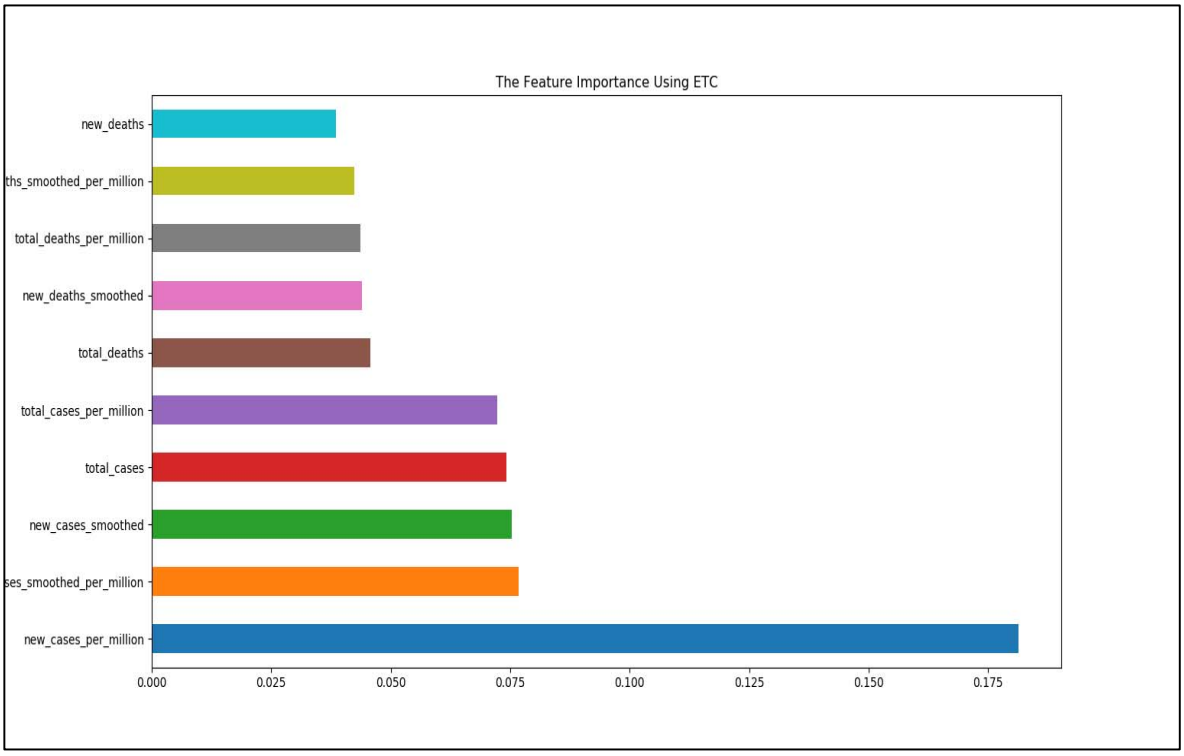

Fig. (5) The sorting of important ten features using ETC.

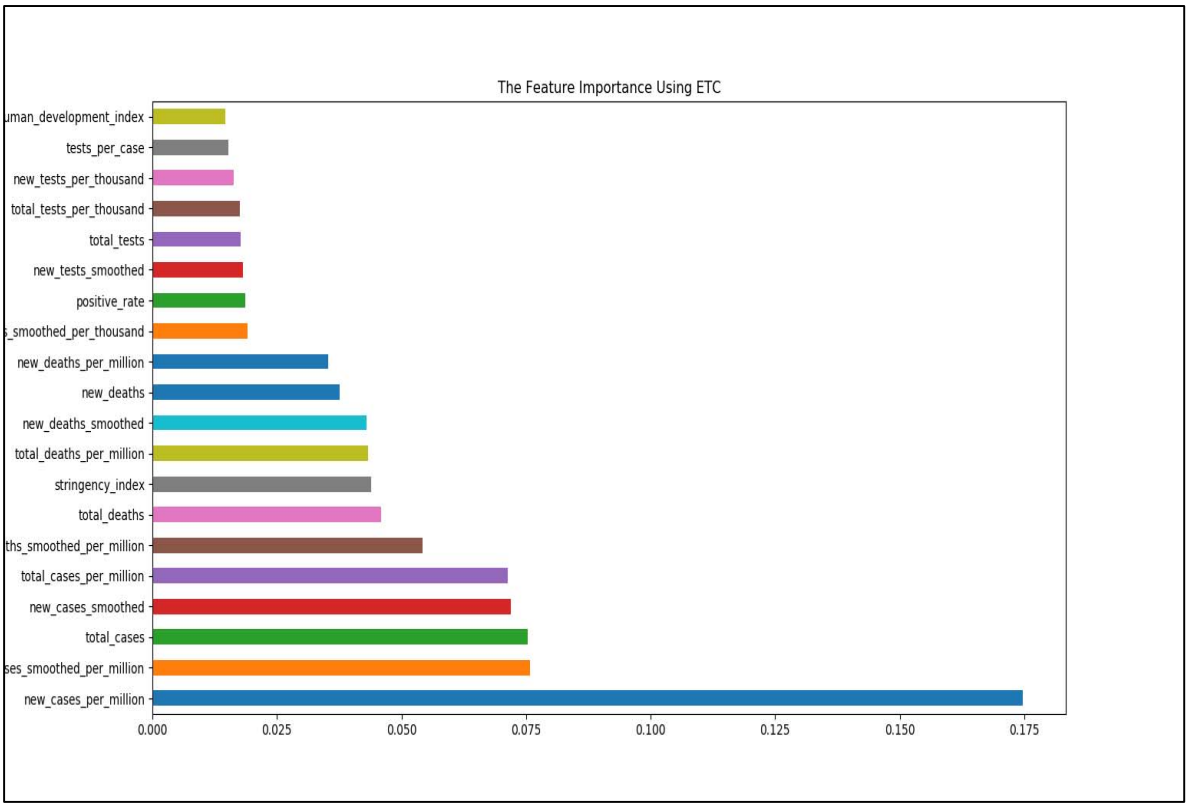

Fig. (6) The sorting of important twenty features using ETC.

Finally, the classification stage aimed to classify the features vector. In this stage, two methods were applied including the NB method and RBM method. The dataset was split into 70% for training and 30% for testing to predicate the disease of **COVID 19**. The equation (3) was used for computing the accuracy of the proposed work.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad \ldots(3)$$

Table (3) presents the results of classification using two methods (NB and RBM) with three cases: firstly, for all features, secondly, for the best 20 features, and finally, for 10 best features which selected in features selection stage. The results show that there is an improvement in accuracy when using the techniques of selecting features. Also, the RBM classifier got a good accuracy comparing with (NB) classifiers.

Table (3) The accuracy of the methods.

The Method	The accuracy for all features	The accuracy for best 20 features	The accuracy for best 10 features
NB	56.181%	61.780%	66.329%
RBM	97.906%	98.949%	99.924%

As shown in figures (4) and (5), the results of the RFE method and ETC method in selecting the important features are approximate. As shown in table (3), the accuracy when using the best ten features is more than the accuracy of the best twenty features. This is because of the correlation between these features. So, it can eliminate the correlated features which may cause a decrease the accuracy and performance as all. By comparing the results of this work with the results of the related works, this work had more accuracy especially when using RBM method comparing with the related works in spite of increasing the size of the dataset except for the work in Reference [8], but this work got more accuracy comparing with that study. The features selection techniques had a more positive impaction on the classification of the feature vector.

5 The Conclusion

In this work, the prediction of **COVID 19** disease using the techniques of feature selection had been achieved. The dataset contained different information about the disease from different countries. The proposed system contains three stages. More than one steps has been applied for preparing the data for the next stage. In the features selection stage, two types of techniques (RFE and ETC) were used for selecting the important features from forty features. The results of the two methods for the important features were approximate. The selecting of the important features affected the accuracy of the proposed work. The accuracy was very good when used RBM technique more than the NB technique.

REFERENCES

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=CjwKCAiA_Kz-BRAJEiwAhJNY7zTX-h33i5Ty1mcFi-XXcMl7cjj7J9FFESmnKgvHxk7JcoynYR5oAhoCcVMQAvD_BwE#2020
[2] https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
[3] Cohen, J.: 'Wuhan seafood market may not be source of novel virus spreading globally', Science, 2020, 10
[4] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., and Gu, X.: 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', The lancet, 2020, 395, (10223), pp. 497-506
[5] Shaban, W.M., Rabie, A.H., Saleh, A.I., and Abo-Elsoud, M.: 'A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier', Knowledge-Based Systems, 2020, 205, pp. 106270
[6] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., and Wei, Y.: 'Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study', The Lancet, 2020, 395, (10223), pp. 507-513
[7] Sun, L., Mo, Z., Yan, F., Xia, L., Shan, F., Ding, Z., Song, B., Gao, W., Shao, W., and Shi, F.: 'Adaptive feature selection guided deep forest for covid-19 classification with chest ct', IEEE Journal of Biomedical and Health Informatics, 2020
[8] Abraham, B., and Nair, M.S.: 'Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier', Biocybernetics and biomedical engineering, 2020, 40, (4), pp. 1436-1445
[9] Pourhomayoun, M., and Shakibi, M.: 'Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making', medRxiv, 2020
[10] Abdulsalam, W.H., Alhamdani, R.S., and Abdullah, M.N.: 'Emotion Recognition System Based on Hybrid Techniques', International Journal of Machine Learning and Computing, 2019, 9, (4)
[11] Remeseiro, B., and Bolon-Canedo, V.: 'A review of feature selection methods in medical applications', Computers in biology and medicine, 2019, 112, pp. 103375
[12] Tangirala, S.: 'Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm'
[13] RASHA H. ALI, D.M.N.A.A.D.B.F.A.: 'SPEAKER IDENTIFICATION AND LOCALIZATION USING FUSION OF FEATURES AND SCORE LEVEL FUSION ', Journal of Theoretical and Applied Information Technology, 2018 96, (21), pp. 11
[14] Urbano Romeu, Á.: 'Emotion recognition based on the speech, using a Naive Bayes classifier', Universitat Politècnica de Catalunya, 2016
[15] Nasrin, S., Drobitch, J.L., Bandyopadhyay, S., and Trivedi, A.R.: 'Low power restricted Boltzmann machine using mixed-mode magneto-tunneling junctions', IEEE Electron Device Letters, 2019, 40, (2), pp. 345-348
[16] Hu, H., Gao, L., and Ma, Q.: 'Deep restricted boltzmann networks', arXiv preprint arXiv:1611.07917, 2016