International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I8P216
Volume 69 Issue 8, 126-137, August, 2021

Disease Forecasting and Severity Prediction Model for COVID-19 Using Correlated Feature Extraction and Feed-Forward Artificial Neural Networks

Jayaraj T [1], Dr. J. Abdul Samath [2]

[1]*Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, India.*
[2]*Assistant Professor, Chikkana Government Arts and Science College, Tiruppur, India*

[1]yoursjayan@gmail.com, [2]abdul_samath@yahoo.com

Abstract *- The digitization of the medical sector has led to an explosion of heterogeneous medical records. The contribution of big data in the medical field is used to effectively address certain unsolved issues. Effectively integrating and analyzing this big data can reveal many useful hidden medical information. The COVID-19 virus, which first appeared in China at the end of 2019, is suffocating the rest of the world. Traditional methods of preventing this unforeseen pandemic are the lengthy process that can take several years. Now, the second wave of COVID-19 is wreaking havoc throughout the world. The survival rate can be significantly increased by predicting the risk of COVID-19 infection based on the patient's early symptoms and health status. Various disease forecasting approaches based on machine learning, and deep learning has been developed to forecast the severity rate of COVID-19. However, these approaches are becoming useless as the virus mutates. In this proposed research, the prediction model is trained by the proposed Correlated Feature Extraction (CFE) method according to the virus behaviors. Furthermore, two prediction models were established in this study. The first model uses initial symptoms to forecast positive cases. Second, based on the health status of the positive cases, the risk rate is estimated. In this study, the disease is forecast by Feed-Forward Artificial Neural Networks (FFANN). Finally, a comparative study has been conducted with the recently developed COVID-19 disease prediction methods to demonstrate the training efficacy and accuracy of the proposed COVID-19 forecasting system.*

Keywords — *COVID-19 disease prediction, Big data, Artificial intelligence, Disease control, Expert system.*

I. INTRODUCTION

Due to medical digitalization, many petabytes of medical data are being produced every day [5][6]. Analyze this big medical data efficiently will provide a wealth of knowledge for medical advancements. Due to the advent of Graphical Processing Units (GPU) and the evaluation of deep learning algorithms, the opportunities for processing big data are very high during this period [19]. However, analyzing this data presents numerous obstacles due to the heterogeneous nature of this big medical data [7][8]. For example, medical data might be in a variety of data formats (.txt, .jpg, and others) and data types (numeric, string, and other data types) [9]. The primary objective of this research is to develop efficient ways for processing heterogeneous-natured medical big data, consequently increasing the overall accuracy of the prediction model. In addition, to reduce the computational overhead caused by the vast quantity of the data, a specialized high correlated feature extraction algorithm should be designed. The ultimate goal is to develop a very efficient disease prediction system based on deep learning with a proposed highly correlated feature extraction method.

The sudden epidemic of the coronavirus has paralyzed people all over the world. As a result, the entire planet faces a substantial loss of valuable human lives as well as significant economic losses [10]. The World Health Organization (WHO) and medical experts are battling to keep this medical emergency under pressure. This virus is highly infectious from one person to another due to its existence. Its global reach is increasing day by day. After the United States, India, Brazil, Russia, and Colombia are the next countries to be affected. Wearing masks, maintaining physical distance, and implementing lockdown are ineffective control mechanisms used by countries to restrict this malicious virus. Additionally, there are numerous obstacles to developing an effective vaccination due to the virus's mutations [11][12].

In India, the corona virus's Delta strain (B.1.617.2) caused the most havoc. B. 1.617.2's positive rate is increasing daily; as a result, the death rate is uncontrollably high. Consequently of this deadly virus, hospitals faced a shortage of beds and oxygen. The main reason for these enormous casualties is caused by lack of public awareness about the fast-spreading and deadly virus and the implementation of poor preventive strategies. Furthermore, the primary cause of this intolerable disaster is that many people do not seek proper medical advice and do not isolate themselves despite their symptoms. This research aimed to improve the computer-aided predictability of COVID-19 and

enhance traditional disease severity prediction methods.

As a consequence of the medical urgency created by COVID-19, numerous studies have presented a variety of techniques for controlling this disease. However, due to the COVID-19 virus's diverse nature and mutations, all of these methods become infective. The first wave of COVID-19 is most dangerous to immunocompromised and older people, whereas the second wave is more dangerous to middle-aged and healthy adults. Existing methods place little emphasis on feature selection, which has a significant impact on prediction efficiency. In this study, a dynamic correlated feature extraction method has been proposed to address this shortcoming. This will update the features depending on the behavior of the virus. This will definitely reduce the False Positive (FP) and False Negative (FN) rates of the proposed disease prediction model. This has been proven in experimental analysis.

In addition, two disease forecasting systems have been developed in this proposed method using Feed Forward Artificial Neural Networks (FFANN). The first methodology is to forecast COVID-19 positive cases based on initial symptoms. If the prediction outcome is positive, the severity rate is estimated based on the patient's medical condition. This enables patients to make an informed decision on whether or not to be admitted to the hospital. This has the potential to significantly lower mortality rates. The following sections summarize the significant contributions of the CFE-FFANN based COVID-19 disease forecasting system.

1. Using deep learning and big medical data to develop a highly efficient COVID-19 disease severity prediction method.
2. The Correlated Feature Extraction (CFE) method is introduced to prevent loss of prediction accuracy caused by the virus mutation.
3. Reducing unnecessary costs associated with COVID-19 medical investigations.
4. Reduce the mortality rate caused by patient's unawareness.

Section 2 examines the advantages and disadvantages of the COVID-19 disease predicting algorithms that were recently established. Section 3 details the foure critical modules of the proposed COVID-19 disease forecasting methodology. In section 4, experimental analysis and comparison study utilizing newly developed COVID-19 disease forecasting methods. Finally, the conclusion is discussed, as well as the feature research.

II. LITERATURE REVIEW

This section reviews recently developed COVID-19 disease prediction methods.

On the basis of data from Tongji Hospital, Li Yan et al. [1]. Developed a COVID-19 prediction system. This dataset contains 375 medical records of individuals who were admitted to the hospital as inpatients. The mortality rate is exceptionally high in this data set, with 174 people dying and 201 surviving. The XGBoost (XGB) model is used to train the classification system. Additionally, the disease prediction model is trained using the values for lactic lymphocytes, dehydrogenase, and high-sensitivity C-reactive protein. This research indicates a precession rate of 95% and an accuracy of 90%. However, this model was trained using a small amount of data, having a negative effect on its reliability. Sumayh S. Aljameel et al [2]. Developed the COVID-19 severity prediction system by combining the most widely used machine learning algorithms (random forest, extreme gradient boosting, and logistic regression (XGB)). In this technique, data from 287 COVID-19 patients at King Fahad University Hospital in Saudi Arabia were used to train machine learning models. Additionally, this approach is trained using the twenty features contained in the data set. According to this study, the random forest classifier has a maximum accuracy of 95%. Although the random classifier is more accurate, machine learning models are not trained to adjust for the virus's heterogeneity. Aziz Alotaibi et al. [3] have used the popular machine learning algorithms Artificial Neural Network, Support Vector Machine, and Random forest to predict COVID-19 at an early stage. These models are trained based on the clinical history and laboratory findings of the patients. This dataset was collected from the Peking University Clinical Research Institute. Out of the 52 clinical features in this data set, the top 20 features have been used to train machine learning models. In this method, the random forest gives more than 90% accuracy. Although this method gives excellent accuracy, the most considerable amount of lapse in total accuracy occurs when the behavior of the testing data set changes. Mortality analysis of COVID-19 infected patients using supervised and unsupervised learning methods by Manuel Sánchez-Montañés et al. [4]. Logistic Regression is utilized as the unsupervised learning model, and Decision Tree, Bayesian Network, is used as the supervised learning model. The information of 1696 COVID-19 confirmed cases had been taken for study. Logistic Regression according to this method yields a maximum of 86 percent accuracy. In this procedure, just a mortality analysis of the disease is done.

Limitations of existing methods

1. All COVID-19 prediction methods addressed in the literature review were trained on a small sample size of data. Due to the COVID-19 epidemic, a massive amount of heterogeneous clinical data is being generated. These present approaches make no recommendation for an algorithm or method for processing large amounts of data.
2. Due to the heterogeneous nature of the COVID-19 virus, the manner in which it attacks humans also varies. All the existing methods have the same features used to train the model. This affects prediction accuracy when virus mutation occurs.

III. PROPOSED METHODOLOGY

The proposed disease prediction methodology is divided into four key modules. With the help of the first module, two different types of medical data sets are included for noise reduction and type conversion using advanced pre-processing methods. The second module extracts the most important features of COVID-19 disease from the pre-processed data sets. COVID-19 positive cases are predicted by the third module. Finally, the patient's risk factor is predicted by the disease severity prediction module if positive results are obtained during the third phase. The entire architecture and process flow of the proposed COVID-19 disease prediction system are depicted in Figure 1.

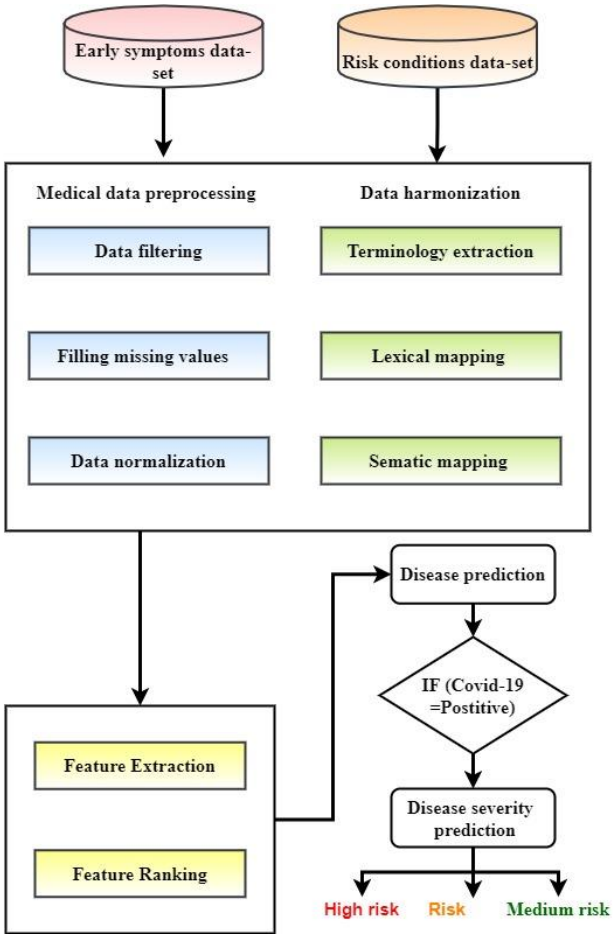

Fig1. The overall architecture and process flow of the proposed COVID-19 disease prediction system.

A. Dataset details

Two distinct data sets are analyzed in this study. The first data set comprises information on the COVID-19 test. The second dataset provides medical information about individuals impacted by COVID-19 who were admitted as inpatients. These data sets include information on COVID-19 patients who were registered at primary health clinics and government hospitals in South India between June 20, 2020, and April 1, 2021. The first dataset (early symptoms dataset) is used to train the disease prediction model, and the second dataset (risk conditions dataset) is used to train the severity prediction model.

Early symptoms dataset

The Early-Symptoms-Dataset set was gathered from a number of primary health centers in southern India. This data set comprises the initial symptoms and clinical characteristics of patients referred for COVID-19 testing. Early symptoms data set contains information for a total of 18960 patients, 1436 of whom are COVID-19 positive cases. Additionally, these data sets include clinical information on men and women of various age groups. In this study, Early-Symptoms-Dataset is being used to train the proposed CFE-FCANN- disease prediction model. The target class has two values, namely positive and negative. The data set for this study include the 12 most significant early symptoms of COVID-19. This is detailed in Table 1.

Risk conditions dataset

The risk conditions dataset is used to train the proposed CFE- FCANN severity prediction model. This data collection includes clinical information on 1463 patients (both genders) who were hospitalized as inpatients after testing positive for COVID-19. This information was gathered from Asaripallam Medical College in Kanyakumari, Rajaji Hospital in Madurai, and Tirunelveli Medical College Hospital in Palayamkottai. The risk conditions dataset contains 3238 records and 32 columns. Of these 3238 patients, 360 died. 763 people have been recovered from life-threatening situations. 2215 people have returned home after the treatment with mild symptoms. The target class has three values: death, severe and recovered. Table 3 contains the medical records of the patients in the risk conditions data set. The pre-processed data is converted to numeric values using the type conversion methods to reduce the number of computational resources required to train the proposed disease prediction models. It is given in Table 1 and Table 2.

B. Medical data pre-processing module

In this module, data filtering, filling missing values, and data normalization is used to reduce the high computational power required for data processing and improve the prediction model's accuracy.

Data filtering

To improve the prediction accuracy and reduce the computational power of this proposed study, unnecessary records such as patient name, address, phone number, email address, and others are first omitted or filtered.

Table 1. COVID-19 early symptoms data set details and the type conversion conditions.

S.No	Feature name	Description	Type conversion	Data type
1	Age	This denotes the outpatient's age. The variable D_AIt is used to denote this.	If(D_A>=70) { 1} If(D_A<=70 &&D_A >=50) { 2 } else {0}	Numeric
2	Gender	It relates to the outpatient's gender. The variable D_G is used to denote this.	If(D_G=M) {1} If(D_G=F) {2} else {0}	Numeric
3	Fever	Fever is a term that refers to a patient's elevated body temperature. Fahrenheit is the unit of measurement. The variable D_F is used to indicate this.	If(D_F >=102) {1} If(D_F<=102&&D_F >=98) {2} else {0}	Numeric
4	Dry cough	It refers to the state of a patient's dry cough during a clinical evaluation. The variable $D_{DC}$is used to indicate this.	If(D_{DC}=yes) {1} else {0}	Numeric
5	Tiredness	Tiredness is a term that refers to patients who exhibit signs of fatigue during clinical evaluations. The variable D_T is used to indicate this.	If(D_T=yes) {1}else{0}	Numeric
6	Aches and pains	It refers to the patient's aches and pains in the body. The variable D_{AP} is used to indicate this.	If(D_{AP}=yes) {1}else{0}	Numeric
7	Sore throat	It relates to the state of a patient's throats at the time of medical examination. The variable D_{ST} is used to indicate this.	If(D_{ST}=yes) {1}else{0)	Numeric
8	Diarrhea	This field specifies whether the patient experienced diarrhoea throughout the course of the clinical evaluation. The variable D_D is used to indicate this.	If(D_D = yes) {1}else{0}	Numeric
9	Conjunctivitis	This indicates whether the patient had conjunctivitis during clinical evaluation. The variable represents this D_C.	If(D_C.=yes) {1} else {0}	Numeric
10	Headache	This field shows whether or not the patient experienced a headache during the clinical examination. The variable D_H is used to indicate this.	If(D_H=yes) {1}else{0}	Numeric
11	Loss of taste or smell	This field shows whether the patient experienced a loss of taste or smell during the clinical assessment. The variable D_{TS} is used to indicate this.	If(D_{TS}=yes) {1}else{0}	Numeric
12	Rash on skin	This indicates whether the patient had a skin rash during the clinical evaluation. The variable represents this D_{RS}.	If(D_{RS}.=yes) {1}else{0}	Numeric

Table 2. COVID-19 patient's medical history and type conversion conditions of the risk conditions dataset.

S.No	Feature name	Description	Type conversion	Data type
1	Age	This denotes the inpatient's age. The variable S_A is used to denote this.	If(S_A>=70) { 1} If(S_A<=70&&S_A>=50) { 2 } else {0}	Numeric
2	Chronic medical illness	This field indicates if the patient was admitted with a chronic medical ailment. The variable S_{CML} is used to indicate this.	If(S_{CML}=yes) {1}else {0}	Numeric
3	Oxygen Level	It refers to the patient's oxygen saturation level. The variable S_{OL} is used to indicate this.	If(S_{OL} <=90&& S_{OL} >=85) {1} If(S_{OL} <=85&& S_{OL} >75) {2} Else {0}	Numeric
4	Blood glucose level	It refers to the patient's blood glucose level at the time of admission. The variable S_{BGL} is used to indicate this.	If(S_{BGL}=Normal) {0} Else{1}	Numeric
5	Blood pressure	It refers to the patient's blood pressure at the time of admission. The variable S_{BP} is used to indicate this.	If(S_{BP}=Normal) {0} Else {1}	Numeric
6	Respiratory rate	This refers to the respiratory rate of patients when admitted into the hospital. This is represented by the variable S_{RR}.	If(S_{RR}=Normal) {0} Else {1}	Numeric
7	Lipid profile status	It refers to the total cholesterol level of inpatients when they are admitted. The variable S_{LPS} is used to indicate this.	If(S_{LPS}=Normal) {0} Else {1}	Numeric
8	Renal problem	This indicates that the patient was admitted to the hospital with a renal issue. The variable S_{RP} Represents this.	If(S_{RP} =NO) {0} Else {1}	Numeric
9	Cardiac Problem	This signifies that the patient was admitted to the hospital with a Cardiac Problem. The variable S_C is used to indicate this.	If(S_C =NO) {0} Else {1}	Numeric
10	Shortness of breath	This shows that the patient is experiencing shortness of breath during hospitalization. The variable S_{SB} is used to indicate this.	If(S_{SB} =NO) {0} Else {1}	Numeric
11	Chest pain	It referred to the presence of chest pain when patients were admitted to the hospital. The variable S_{CP} is used to indicate this.	If(S_{CP} =NO) {0} Else {1}	Numeric
12	loss of speech	This signifies a loss of speech during the hospitalization process. The variable S_{LS} is used to indicate this.	If(S_{CP} =No) {0} Else {1}	Numeric
13	Fever	This shows that the patient had a fever at the time of his hospitalization. The variable S_F is used to indicate this.	If(S_F =No) {0} Else {1}	Numeric
14	Smoking history	This shows the patient's smoke history. The variable S_H is used to indicate this.	If(S_H =Yes) {1} Else {0}	Numeric

Filling missing values

Filling in missing values incorrectly will degrade the quality of extracted information and result in incorrect results. Hence, filling in missing values in the medical sector must be performed with extreme precision because incorrect findings may have significant consequences. In this proposed study, the Expectation-Maximization (EM) algorithm was used to fill in the missing values. An EM algorithm is an iterative method for determining the maximum probability or maximum a posteriori estimates of parameters in statistical models with unobserved latent variables.

Data normalization

Normalization is a term used in traditional database design to refer to grouping the fields and tables to reduce complexity and dependencies. Normalization is a term used in healthcare to describe the process of rationalizing data to a common vocabulary or definition, such as patient information, symptoms, lab results, and others.

C. Data harmonization

Data harmonization is the process of combining medical data from various healthcare sources in order to provide a comparable view of medical data. For data harmonization, three methods are used in this proposed study: terminology extraction, lexical matching, and semantic matching.

Terminology extraction

Medical terminologies such as disease names, signs, clinical trials, electronic health records, adverse case reports, emails, and others are often included in large volumes of biomedical data. However, these texts are primarily written in the language of the associated community. As a result, formalization and cataloging of these technical terms or definitions are needed. Additionally, these technical terms are critical for information retrieval. However, manually extracting these technical terms is a lengthy and laborious task. Using SQL queries, this proposed method extracts the most critical medical terminologies associated with COVID-19.

Lexical mapping

Lexical mapping is used to map the most significant relationship between disease and symptoms. In this proposed research, lexical matching is performed using SQL queries.

Sematic mapping

Semantic mapping is the method of associating medical keywords with their semantic meaning. Thus, different words can be related to the same definition in biomedical ontologies and have semantically similar writing, such as “cancer” and “neoplasm”. Semantic mapping significantly increases prediction accuracy in medical data analysis.

D. Feature Extraction

In general, optimal feature selection methods determine the efficiency of deep learning or machine learning algorithms. The inclusion of incorrect features in medical data analysis considerably increases the probability of diagnosis mistakes. This is seriously damaging to someone who is utilizing the system. As a result, advanced feature extraction techniques are necessary to identify the most relevant features. The data sets used to forecast COVID-19 are multidimensional, having m number of rows and n number of columns. The dimension of the data set to change when the virus’s and disease’s behavior change. Features and class labels are constantly changing in the data set due to the heterogeneous nature of COVID-19 disease and virus mutation. This is specified in the following format.

$$F = \begin{bmatrix} \overrightarrow{f1} \\ \overrightarrow{f2} \\ \overrightarrow{f3} \\ . \\ . \\ \overrightarrow{fn} \end{bmatrix} , C = \begin{bmatrix} c1 \\ c2 \\ c3 \\ . \\ . \\ cm \end{bmatrix} \quad (1)$$

F has n the number of features vectors, and C has m number of class labels. n and m vary depending on the nature of the virus and the disease. The correlation between F and C is determined by formula 2.

$$Cor_{fc} = \frac{\sum(F_i - \bar{F})(C_i - \bar{C})}{\sqrt{\sum(F_i - \bar{F})^2 \sum(C_i - \bar{C})^2}} \quad (2)$$

Where $\bar{F}$ represents the observed mean value of the ith feature of the disease symptoms and $\bar{\bar{C}}$ Represents the average value of the class label. The formula 2 returns values between 1 and -1. These values vary depending on the correlation between the Patient’s symptoms and the class labels. If the correlation between the patient’s symptoms and the class label is very strong, it returns positive values between 1 and 0, and else it returns negative values. The correlation vales between the class labels and the patients’ symptoms are described in Table 3.

Table 3 correlation values between disease symptoms and classes labels.

Similarity Values	Similarity Range
1	Very high correlated feature
1-0.8	High correlated feature
0.8-0.6	Medium correlated feature
0.6-0.4	Low correlated feature
0	No correlation

In this proposed study, features of correlation vale between 1 and 0.6 are used to train the deep learning models to improve the accuracy of COVID-19 disease and severity prediction. The proposed research use formulas 3 and 4 to select the highly correlated features. To train the FFANN model, the top ten ranked symptoms are chosen from the COVID-19-Early-Symptoms and COVID-19-Risk-Conditions data sets. The ranking of features is depicted in Figures 2 and 3.

$$D_d = \begin{cases} 0.6 \le D_A \le 1; \\ 0.6 \le D_G \le 1; \\ 0.6 \le D_F \le 1; \\ 0.6 \le D_{DC} \le 1; \\ 0.6 \le D_T \le 1; \\ 0.6 \le D_{ST} \le 1; \\ 0.6 \le D_D \le 1; \\ 0.6 \le D_C. \le 1; \\ 0.6 \le D_H \le 1; \\ 0.6 \le D_{TS} \le 1; \\ 0.6 \le D_{RS} \le 1; \\ 0.6 \le D_{AP} \le 1; \end{cases} \quad (3)$$

$$S_d = \begin{cases} 0.6 \le S_A \le 1; \\ 0.6 \le S_{CML} \le 1; \\ 0.6 \le S_{OL} \le 1; \\ 0.6 \le S_{BGL} \le 1; \\ 0.6 \le S_{BP} \le 1; \\ 0.6 \le S_{RR} \le 1; \\ 0.6 \le S_{LPS} \le 1; \\ 0.6 \le S_{RP} \le 1; \\ 0.6 \le S_C \le 1; \\ 0.6 \le S_{SB} \le 1; \\ 0.6 \le S_{LS} \le 1; \\ 0.6 \le S_F \le 1; \\ 0.6 \le S_H \le 1; \\ 0.6 \le S_{cp} \le 1; \end{cases} \quad (4)$$

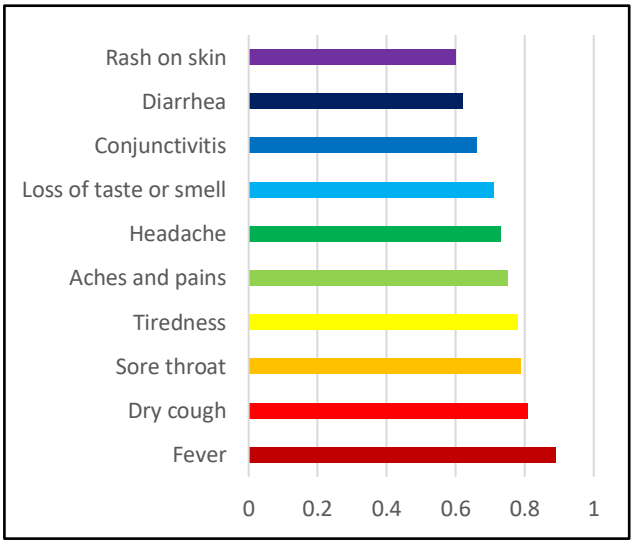

Fig 2. The correlation coefficient between the data set's class labels (COVID-19-Early-Symptoms data set) and its attributes are ranked.

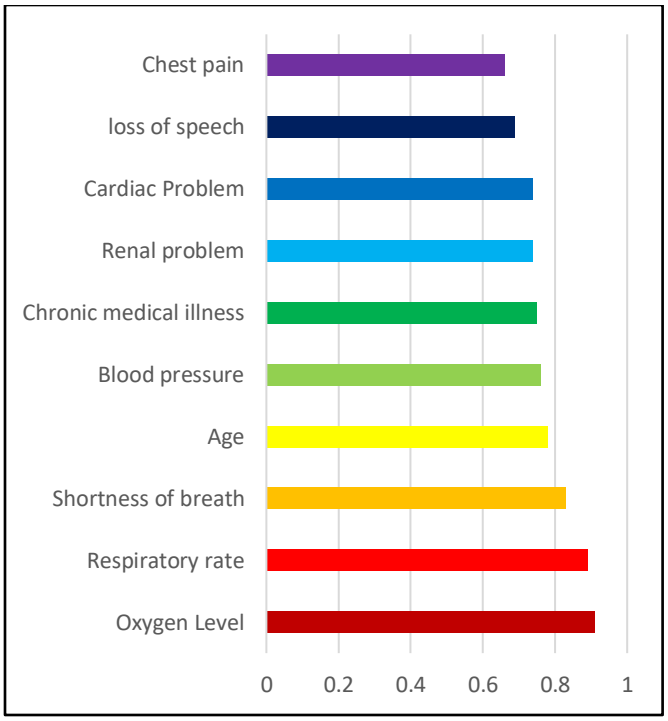

Fig 3. The correlation between the class labels of the dataset (COVID-19-Risk-Conditions dataset) and the attributes is ranked.

E. Feed Forward Artificial Neural Network (FFANN) for disease forecasting

Feed Forward Artificial Neural Network (FFANN) is a deep learning model inspired by biological neural networks [13][14]. This significantly improves the pattern regeneration task in general. Three key layers comprise the FFANN used to forecast COVID-19 disease: the input layer, the output layer, and the hidden layer. Figure 4 depicts the general architecture of FFANN.

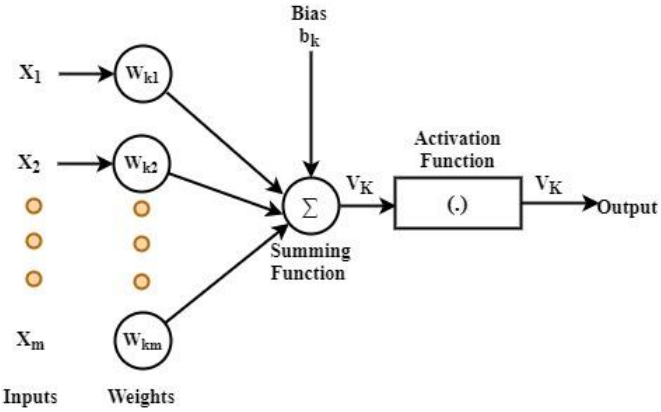

Fig 4. The overall structure of the feed-forward artificial neural network.

FFANN has a total of N input neurons and m output neurons. Using activation and bias, FFANN generates prediction results.

$$I_j = \sum_{i=1}^{n} x_i w_{ij} + b_j \qquad (5)$$

In formula 5, x_i denotes the FFANN's input neurons, b_j bias value and w_{ij} Denotes the FFANN's weight. This disease prediction model makes use of the sigmoid activation function. The activation function generates output utilizing the base value b_j Through hidden layers. The activation function is invoked by formula 6.

$$f(I) = (1 - e^{-2I}) / (1 + e^{-2I}) \qquad (6)$$

Equation 7 yields the FFANN model's prediction output, where $\hat{y}\imath$ Denotes the prediction outcome.

$$\hat{y}\imath = f(I_j) \qquad (7)$$

The proposed research developed a Corrected Feature Extraction-Based Feed Forward Artificial Neural Network (CFE-FCANN-DP) architecture for predicting COVID-19 positive cases and a Corrected Feature Extraction-Based Feed Forward Artificial Neural Network (CFE-FCANN-SP) architecture for predicting disease severity in COVID-19 positive cases. Figures 4 and 5 illustrate the architecture of these two prediction models.

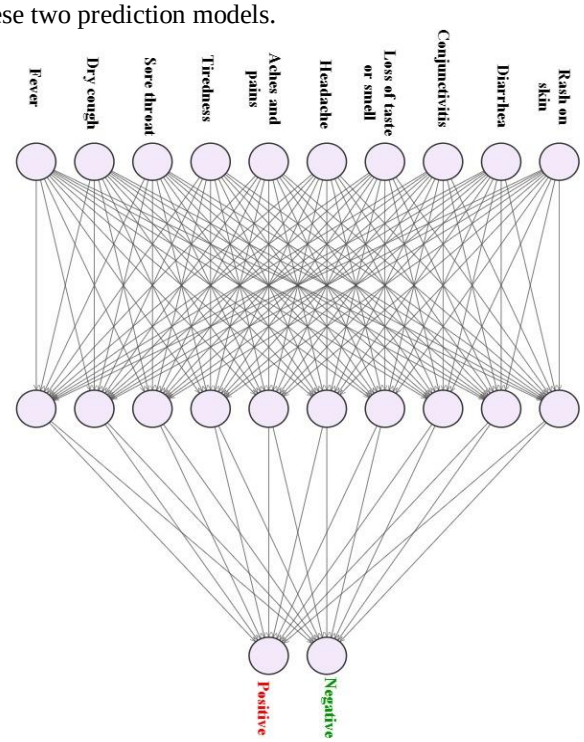

Fig 5. Feed Forward Artificial Neural Network (CFE-FCANN-DP) architecture for predicting COVID-19 positive cases.

The top ten features (fever, dry cough, sore throat, tiredness, aches and pains, headache, loss of taste or smell, conjunctivitis, diarrhea, and rash on skin) from the COVID-19-Early-Symptoms data set are used as input, as illustrated in figure 5. All of these values are numerical. The CFE-FCANN-DP model gives two outputs, returns 1 if COVID-19 is positive and returns 0 if COVID-19 is negative.

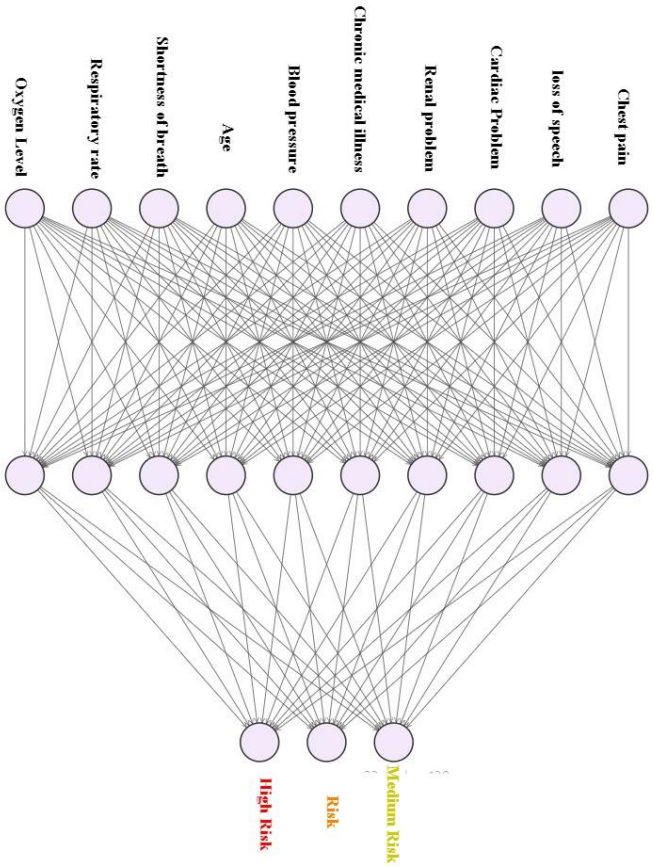

Fig 6. Feed Forward Artificial Neural Network (CFE-FCANN-SP) architecture for predicting COVID-19 patient's risk levels.

As illustrated in figure 6, the CFE-FCANN-SP model is fed the top ten features (oxygen Level, respiratory rate, shortness of breath, age, blood pressure, chronic medical illness renal problem, cardiac problem, loss of speech, and chest pain) extracted from the COVID-19-Risk-Conditions data set. Each of these values is a numeric value. The CFE-FCANN-SP model generates three outputs with a high-risk level. Returns 1 if the current risk level is exceptionally high, 0.5 if the risk level is medium, and 0 if the risk level is very low.

Backpropagation is used to train both of these FFANN models. Formula 2 illustrates this. The learning rate is stated to be 0.2.

$$w_{(t+1)} = w_t - \eta \frac{d(E)}{d(w_1)} \qquad (8)$$

Where $w_{(t+1)}$ denotes the weight updating procedure of

the proposed two FFANN models. The FFANN's learning rate is denoted by η. When the model is training, E signifies the total error rate.

. IV. EXPERIMENTAL ANALYSIS

A. System configuration and software details

The proposed deep learning-based approach for predicting COVID-19 disease and disease severity is implemented to prove the training efficiency and prediction accuracy. Software development tools such as Matlab 2016 and Matlab deep learning toolbox have been used to develop the proposed prediction models. To demonstrate the effectiveness of this proposed system, a Dell EMC DSS 8440 Server with NVIDIA RTX GPU, 24 GB GDDR6, 2 x Intel Xeon 6248, 20 C @ 2.5 GHz, and 2TB of storage is used. Furthermore, the software tools are implemented on the Windows 10 professional operating system.

B. Training efficiency analysis of proposed COVID-19 disease prediction model

The proposed COVID-19 disease prediction model is trained using highly significant COVID-19 disease behaviors derived using the proposed High correlated feature extraction and selection (CFE) method. The training is carried out using k-fold cross-validation. The COVID-19 early symptoms and risk conditions data sets are used to train the proposed disease prediction models. Features extraction is done through the proposed CFE method. The disease prediction model is trained using the ten leading features extracted from the COVID-19 early symptoms data set. The risk rate of COVID-19 disease is predicted using the top 10 features extracted from the COVID-19 risk factors data set.

Furthermore, the training efficiency of this disease forecasting approach is determined by the four most frequently used evaluation metrics, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) (RMSE) [17][18]. If the values of (MAE), (MSE), and (RMSE) are minimal, the proposed COVID-19 disease prediction model performs effectively. The mean absolute error (MAE), the mean square error (MSE), and the root mean square error (RMSE) are determined using formulas 9, 10, and 11.

MAE returns the n number of error statistics of the proposed COVID-19 prediction model. It determines the average values between the actual disease risk rate caused by COVID-19 and the predicted results for n number of samples. Where yi refers to the actual risk of the patient and $\hat{y}\imath$ Refers to the output (patient severity level) produced by the proposed method.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - \hat{y}\imath| \qquad (9)$$

MSE refers to the average square difference between the observed data and the predicted results of the proposed COVID-19 disease forecasting model [18-19].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}|yi - \hat{y}\imath|^2 \qquad (10)$$

RMSE refers to the standard deviation of the difference between the proposed COVID-19 disease forecasting model's predated result (patient risk rate predicted by the proposed model) and actual results (actual patient risk caused by COVID-19) [17].

$$RSME = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|yi - \hat{y}\imath|^2} \qquad (11)$$

To train the proposed CFE-FFANN models, the input data is divided into three configurations: 60%, 70%, and 80%. The data used for training and testing is randomly selected.

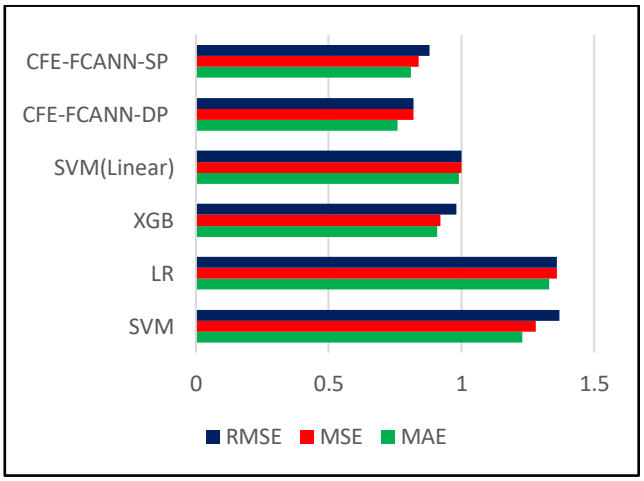

Fig 7. MAE, MSE, and RMSE values were obtained when training the proposed and existing disease prediction model with 60% of training data.

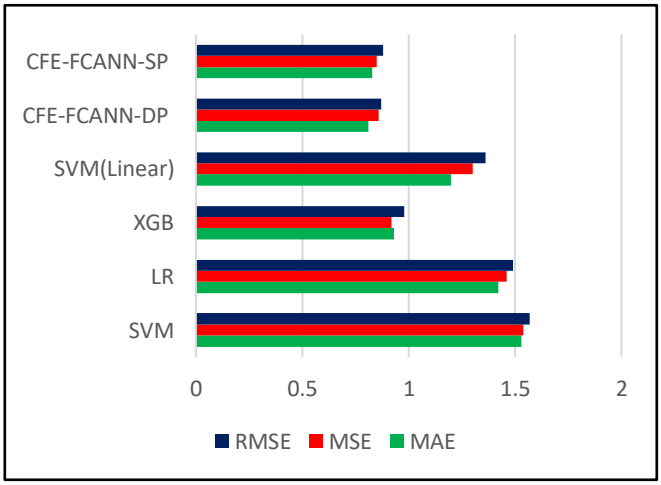

Fig 8. MAE, MSE, and RMSE values were obtained when training the proposed and existing disease prediction model with 70% of training data.

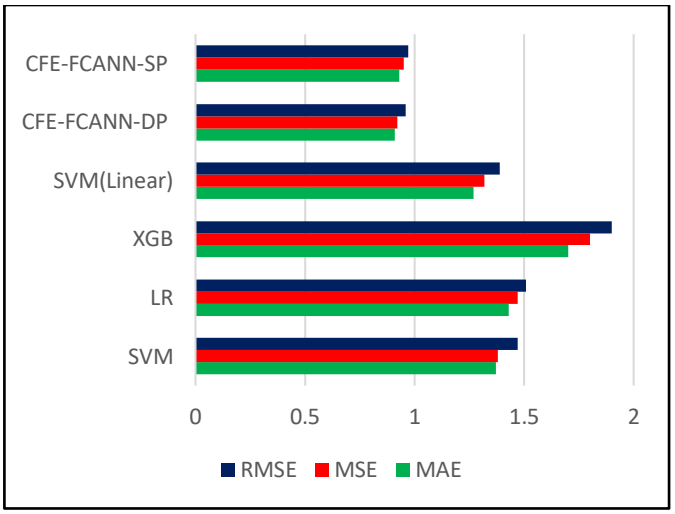

Fig 9. MAE, MSE, and RMSE values were obtained when training the proposed and existing disease prediction model with 80% of training data.

The MAE, MSE, and RMSE values obtained while training the proposed and existing disease prediction models with 60%, 70%, and 80% training data are displayed in figure 7, figure 8, and figure 9. Figure 7, Figure 8, and Figure 9 show that the proposed model's error rate is lower than the existing disease prediction model. The existing machine learning methods are not given much importance to Medical big data noise reduction and data integration. As the training data set grows in size, the error rate of existing disease prediction algorithms increases dramatically, which will affect the prediction accuracy. The proposed two disease predictions have an error rate of less than one in all three sizes of training data (60 %, 70 %, and 80 %).

The accuracy of proposed disease prediction methods and existing machine learning-based COVID-19 disease prediction models are evaluated using the most relevant accuracy measures, including Total Accuracy (TA), Precession (PRE), Recall (REC), and F1-Measure (F1-M). Each of these performance indicators varies according to the following accuracy variable: True Positive COVID-19 forecast, True Negative COVID-19 forecast, False Positive COVID-19 forecast, and False Negative COVID-19 forecast.

True Positive COVID-19 forecast: If the proposed method accurately predicts the COVID-19 disease and severity based on the patients' symptoms and health history, it is called a True Positive COVID-19 forecast. It is described by the variable P_{CF} .

True Negative COVID-19 forecast: If the proposed method properly predicts COVID-19 negative cases based on the symptoms and health history of the patients, it is referred to as a True Negative COVID-19 forecast. The variable TN_{CF} is used to describe it.

False Positive COVID-19 forecast: If the proposed method incorrectly predicts COVID-19 negative cases based on patients' symptoms and health history, it is called a False Positive COVID-19 forecast. This is described by the variable FP_{CF}.

False Negative COVID-19 forecast: If the proposed method fails to predict positive cases based on patients' symptoms and health history, it is called a False Negative COVID-19 forecast. It is described by the variable FN_{CF}.

Equation 12 determines the total accuracy (TA) of the proposed COVID-19 prediction models.

$$TA = \frac{TP_{CF}+TN_{CF}}{TN_{CF}+TN_{CF}+FP_{CF}+FN_{CF}} \quad (12)$$

The suggested COVID-19 prediction models precession rate (PRE) is determined by equation 13.

$$PRE = \frac{TP_{CF}}{TP_{CF}+FP_{CF}} \quad (13)$$

The proposed COVID-19 prediction models recall rate (REC) is calculated by equation 14.

$$REC = \frac{TP_{CF}}{TP_{CF}+FN_{CF}} \quad (14)$$

The F1- Measure (F1) of the proposed COVID-19 prediction models is determined by equation 15.

$$F1 = \frac{2(PRE \times REC)}{\text{PRE+REC}} \quad (15)$$

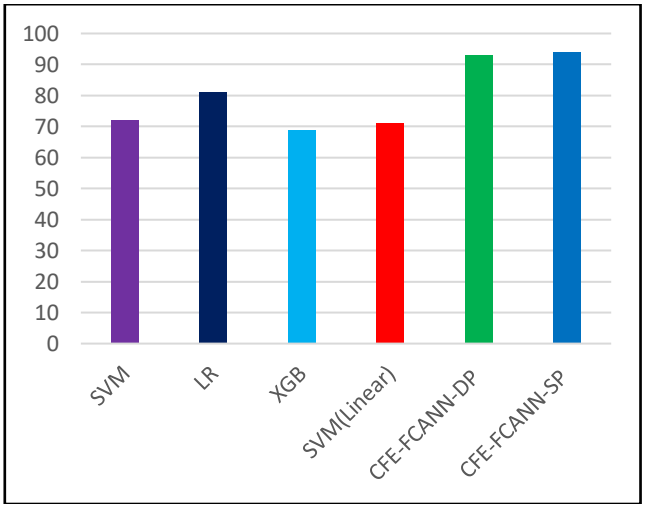

Fig 10. Total accuracy comparison of proposed COVID-19 disease prediction and existing methods.

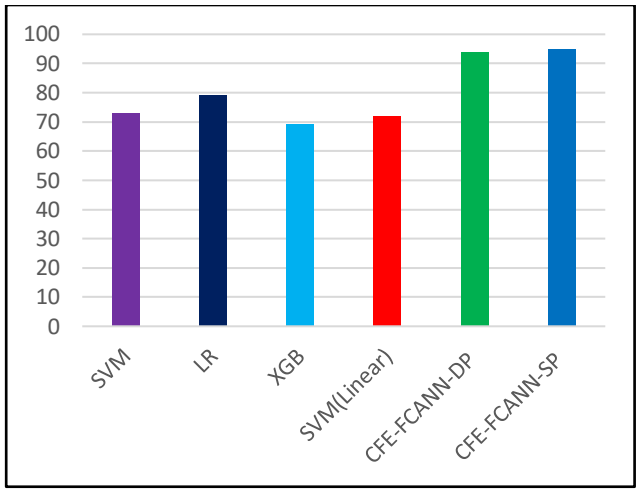

Fig 11. Precession rate comparison of proposed COVID-19 disease prediction and existing methods.

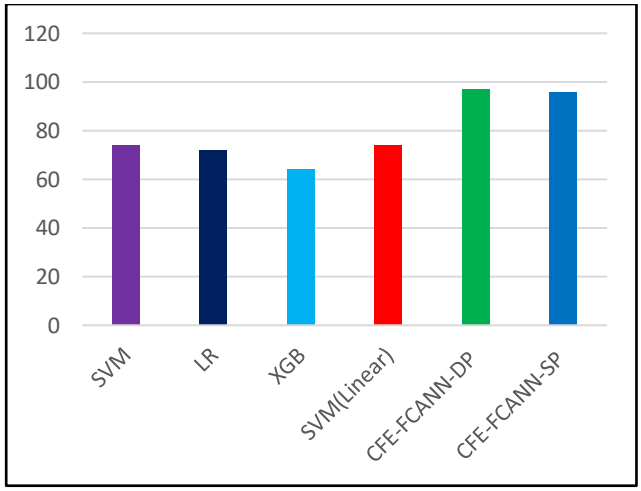

Fig 12. Recall rate comparison of proposed COVID-19 disease prediction and existing methods.

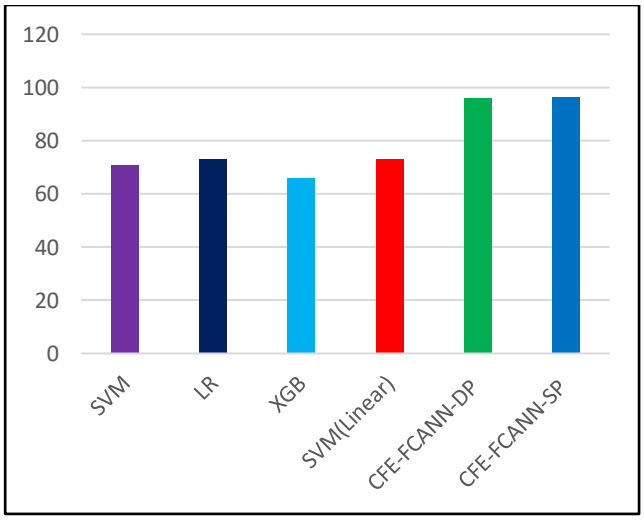

Fig 13. F1-measure comparison of proposed COVID-19 disease prediction and existing methods.

To prove the proposed COVID-19 disease forecasting method's prediction accuracy, test data from the first wave and test data from the second wave are randomly fed to FFAFN. Figure 10 compares the total accuracy of the proposed and existing COVID-19 prediction models. CFE-FCANN-DP and CFE-FCANN-SP, the proposed disease prediction models, achieve the highest accuracy rates of 93 percent and 94 percent, respectively. The main reason for this is that the Correlated Feature Extraction method is being used.

Figure 11 depicts the precession rate of the proposed disease forecasting methods with the existing disease forecasting methods. Correlated Feature Extraction and pre-processing methods significantly reduce the FP rate of the proposed CFE-FCANN-DP and CFE-FCANN-SP. As a result, the proposed methods achieve the highest precession rate of 94% and 95%, respectively.

The recall rate for the proposed and existing disease forecasting systems is depicted in Figure 12. Correlated Feature Extraction and pre-processing methods greatly minimize the FN rate of the recommended CFE-FCANN-DP and CFE-FCANN-SP, so that CFE-FCANN-DP and CFE-CANN-SP achieves the highest recall of 97% and 95%. The F1 measure is generally greater when precession and recall are improved. The proposed approaches are capable of achieving 96% and 96.3% F1-measure, which is illustrated in figure 13.

C. Discussion

Numerous strategies have been developed to identify and classify COVID-19 patients according to their clinical symptoms. Recently published COVID-19 prediction algorithms focus primarily on a patient's early symptoms. Due to virus mutation, the disease's symptoms change, making all previously developed COVID-19 prediction methods infective. For example, the first wave of COVID-19 severely affected the elderly and co-morbidities, whereas the second wave had a significantly positive impact on the healthier population. In this research, the medical details of the patients infected by the first and second waves of COVID-19 were used to train the FFANN. Prediction errors in the medical area can have serious consequences for patients. Correlated Feature Extraction (CFE) is employed to correct prediction errors caused by virus mutations. The CFE method significantly reduces both the FP and FN rates. As a result, the proposed COVID-19 disease prediction model's overall accuracy, precession, and recall rate have considerably enhanced. Furthermore, employing this CFE-FFANN model significantly reduces the risk of subsequent infections caused by a virus-like COVID-19.

Massive volumes of disease-related data are frequently employed to increase the accuracy of disease prediction models. Simultaneously, the noise in the data has a

significant impact on the training efficiency. The most efficient pre-processing methods are not used in recently developed approaches to process large amounts of data, which significantly impacts training efficiency. This proposed method efficiently implements data filtering and data harmonization in order to reduce noise generated by data size and errors caused by medical big data heterogeneity. The MAE, MSE, and RMSE values available during the proposed method's training are less than one. The fundamental reason for this is because the training data was efficiently handled utilizing modern pre-processing approaches. The experimental results demonstrate that the suitable data processing methods and proposed CFE approach significantly improves the proposed method's training efficiency.

V. CONCLUSION

The new coronavirus (COVID-19) is wreaking havoc on the global economy and threatening the health of entire populations. Treatment of COVID-19 at an early stage can improve survival rates. Two FFANN-based disease prediction models have been developed in this proposed research. Furthermore, a correlated feature extraction method has been introduced to reduce prediction errors caused by virus mutation. COVID-19 patients can be effectively identified using this proposed method, which will benefit doctors, patients, and the health sector. The use of advanced data processing methods and proposed feature extraction techniques can significantly enhance training efficiency and prediction accuracy. This has been demonstrated experimentally and by comparative analysis.

References

[1] L. Yan, H.-T. Zhang, Y. Xiao, et al., Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan, medRxiv, (2020).

[2] Sumayh S. Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri, Eman S. Alsulmi, Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients, Scientific Programming, (2021) Article ID 5587188, 10 pages, 2021. https://doi.org/10.1155/2021/5587188.

[3] Alotaibi, A.; Shiblee, M.; Alshahrani, A. Prediction of Severity of COVID-19-Infected Patients Using Machine Learning Techniques. Computers 10(2021) 31. https://doi.org/10.3390/computers10030031.

[4] Sánchez-Montañés, M.; Rodríguez-Belenguer, P.; Serrano-López, A.J.; Soria-Olivas, E.; Alakhdar-Mohmara, Y. Machine Learning for Mortality Analysis in Patients with COVID-19. Int. J. Environ. Res. Public Health 17(2020) 8386.

[5] Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis, and future prospects. J Big Data 6(54) (2019).

[6] Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 15(3) (2018) 20170030. Published 2018 May 10. doi:10.1515/jib-2017-0030.

[7] Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 15(3) (2018) 20170030. Published 2018 May 10. doi:10.1515/jib-2017-0030

[8] A. Krithara et al., iASiS: Towards Heterogeneous Big Data Analysis for Personalized Medicine, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), (2019) 106-111, doi: 10.1109/CBMS.2019.00032.

[9] S. Kumar and M. Singh., Big data analytics for the healthcare industry: impact, applications, and tools, in Big Data Mining and Analytics, 2(1) (2019) 48-57, March 2019, doi: 10.26599/BDMA.2018.9020031.

[10] Singhal, T. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 87 (2020) 281–286. https://doi.org/10.1007/s12098-020-03263-6.

[11] V. Chamola, V. Hassija, V. Gupta and M. Guizani., A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact., in IEEE Access, 90225-90265, 8(2020) doi: 10.1109/ACCESS.2020.2992341.

[12] T. Xin., The Model of COVID-19 Pandemic, International Conference on Computing and Data Science (CDS), (2020) 429-432, doi: 10.1109/CDS49703.2020.00090.

[13] Lowe, D., Tipping, M. Feed-forward neural networks and topographic mappings for exploratory data analysis. Neural Comput & Applic 4 (1996) 83–95 (1996). https://doi.org/10.1007/BF01413744.

[14] Mičušík, D., Stopjaková, V. & Beňušková, L. Application of Feed-forward Artificial Neural Networks to the Identification of Defective Analog Integrated Circuits. Neural Comput Applic 11 (2002) 71–79. https://doi.org/10.1007/s005210200018.

[15] Hecht-Nielsen., Theory of the backpropagation neural network., International Joint Conference on Neural Networks, 1(1989) 593-605 doi: 10.1109/IJCNN.1989.118638.

[16] M. Roopa and S. S. K. Raja., Artificial neural network using backpropagation algorithm in distributed MANETs., International Conference on Information Communication and Embedded Systems (ICICES), (2016) 1-4.

[17] N. B. M. Khairudin, N. B. Mustapha, T. N. B. M. Aris, and M. B. Zolkepli., Comparison of Machine Learning Models For Rainfall Forecasting, International Conference on Computer Science and Its Application in Agriculture (ICOSICA), (2020) 1-5, doi: 10.1109/ICOSICA49951.2020.9243275.

[18] S. I. Popoola et al., Determination of Neural Network Parameters for Path Loss Prediction in Very High-Frequency Wireless Channel, in IEEE Access, 150462-150483, 7(2019) doi: 10.1109/ACCESS.2019.2947009.

[19] S.Sunitha, Dr.S.S. Sujatha., Combined Feature Learning And CNN For Polyp Detection In Wireless Capsule Endoscopy Images., International Journal of Engineering Trends and Technology 69(6)(2021) 206-215.