

Future forecasting prediction of Covid-19 using hybrid deep learning algorithm

Ganesh Yenurkar^{1,2} • Sandip Mal¹

Received: 16 September 2021 / Revised: 27 October 2022 / Accepted: 1 November 2022 / Published online: 18 November 2022 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Due the quick spread of coronavirus disease 2019 (COVID-19), identification of that disease, prediction of mortality rate and recovery rate are considered as one of the critical challenges in the whole world. The occurrence of COVID-19 dissemination beyond the world is analyzed in this research and an artificial-intelligence (AI) based deep learning algorithm is suggested to detect positive cases of COVID19 patients, mortality rate and recovery rate using real-world datasets. Initially, the unwanted data like prepositions, links, hashtags etc., are removed using some pre-processing techniques. After that, term frequency inverse-term frequency (TF-IDF) and Bag of Words (BoW) techniques are utilized to extract the features from pre-processed dataset. Then, Mayfly Optimization (MO) algorithm is performed to pick the relevant features from the set of features. Finally, two deep learning procedures, ResNet model and GoogleNet model, are hybridized to achieve the prediction process. Our system examines two different kinds of publicly available text datasets to identify COVID-19 disease as well as to predict mortality rate and recovery rate using those datasets. There are four different datasets are taken to analyse the performance, in which the proposed method achieves 97.56% accuracy which is 1.40% greater than Linear Regression (LR) and Multinomial Naive Bayesian (MNB), 3.39% higher than Random Forest (RF) and Stochastic gradient boosting (SGB) as well as 5.32% higher than Decision tree (DT) and Bagging techniques if first dataset. When compared to existing machine learning models, the simulation result indicates that a proposed hybrid deep learning method is valuable in corona virus identification and future mortality forecast study.

Ganesh Yenurkar ganesh.keshaorao2018@vitbhopal.ac.in

Sandip Mal sandip.mal@vitbhopal.ac.in

¹ School of Computing Science & Engineering, VIT Bhopal University, Bhopal, India

² Yeshwantrao Chavan College of Engineering, Wanadongri, Nagpur, India

Keywords Corona disease · Hybrid deep learning model · ResNet · GoogleNet · Feature extraction · Feature selection · As well as mayfly optimization (MO) algorithm

1 Introduction

COVID-19 is otherwise named as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is considered as an infectious disease which is the reason for the 2020 pandemic worldwide. This infection was initially identified in Wuhan, the capital Hubei province, China on late December 2019 and instigating primary epidemic of this period [36]. Due to the spread of infectious COVID-19, almost three million and above positive cases are identified in world-wide from the start of year 2020. Different kind of situations are reason for the spread of this pandemic disease. Some of them are mentioned as follows: (i) high transmission rate of that disease, (ii) unavailability of exact vaccine and proper medicine, and (iii) unknown nature of SARS-COVID-19 [26, 44]. Literally, this disease is identified in many countries like 210 countries and territories around the globe. Each and every country fight against the COVID-19 enlargement with a sequence of choices in the absence of a definitive vaccine to stop the progress of this virus. Whole world is suffered with billions in losses due to the paralyzed economy [1].

The reason for the speedy transmission of COVID-19 is the closer proximity of people and still, there is no proper vaccine is developed to control the diseases full [24, 33]. Viral pneumonia has been speculated due to the COVID19 disease in which it has almost a 5.8% mortality risk. When compared to Spanish flu pandemic on 1918, the expiry rate of corona is equal to 5% of death rate of that flu [17, 37]. Countries like Spain, India, USA, Italy, Germany, France, UK, China and Iran were recorded with the highest number of positive cases in which Saudi Arabia country is reported with a greater number of positive cases between all the Arab countries. In mainland China, information technology services like mobile applications have been developed to curb the possible risk of that pandemic from February 2020 [16, 18, 21]. When someone infected with the virus, the mobile app suggests the user to prepare the concerned health authorities and self-quarantine [25, 41]. Different kind of methods like next-generation sequencing, nucleic acid-based methods using polymerase chain reaction (PCR), computed tomography (CT) scan, chest X-ray (CXR) scan, and paper-based detection are available to diagnosis COVID-19. Generally, these techniques are utilized to monitor the changes in organs and check whether the patients undergoes their pathological tests. Still, there is an unavailability of exact and quick method for COVID-19 diagnosis [22].

AI techniques are considered as one of the best methods to resolve the main problems associated with the COVID pandemic. It has the ability to predict the risk of infection and monitor the affected population. The information like how many persons are infected with particular area and how much time a person pay out in an extremely affected region are used for the prediction process. Based on this info, AI created a spatial prediction model and envisage the infection transmission. Also, symptoms are used as another information which helps to identify the infection person through the AI based prediction model. In order to making proposed decisions, a lot of researchers apply traditional machine and deep learning techniques to assist the doctors. CT scan and X-ray methods are useful to predict the data into two classes named as normal and infected [4].

Therefore, in this research hybrid deep learning model is proposed to predict who may get infected with the deadly virus, how many percentages of recovery rate and how many percentages of mortality rate using datasets. Moreover, negative and positive cases of COVID-19 disease are also predicted with this method [2]. When compared to shallow

models, deep learning method predicts that hierarchical or a deep sequential model is more effective in classification or regression tasks. Due to the ability of processing variable length sequential data, the deep learning algorithms are frequently used in forecasting applications [42]. This research work concentrates on the detection of mortality rate, positive cases, as well as recovery rate in particular datasets using hybrid deep learning algorithm. Major contributions of the proposed methodology is given as below:

- To eliminate the unwanted data by performing different pre-processing steps. Usually, preprocessing involves transforming raw data into a well-formed data set to perform other processes.
- To extract the different kind of features by using different feature extraction techniques. The feature extraction methods applied in this research offers a plenty of flexibility customization on particular text data.
- To minimize the computational complexity of entire system by identifying the optimal features using optimization algorithm. Generally, optimization methods are utilized to find out the best among the possible combinations of feature subsets which helps to minimize the computational complexity.
- To design a hybrid deep learning technique that can effectively predict the COVID-19 disease negative and positive cases, confirmed cases, recovery rate and mortality rate. By using this hybrid deep learning algorithm, we can reduce the demerits of both learning algorithms and provide more accurate and less computationally expensive solutions. Overall, hybrid deep learning models performed better than standalone learning algorithms by means of different parameters and training time.

Other sections of the research paper is ordered as follows: An introduction about COVID-19 disease is delivered in section 1. Section 2 displays the review of recent research papers and problems faced by existing methods. Section 3 elaborates the proposed methodology with deep explanation. Simulation results and outcomes of proposed methodology is illustrated in section 4, and conclusion as well as future enhancement of our research methodology is explained in section 5.

2 Literature review

Corona virus pandemic is still considered as one of the global health disasters in which the world is still suffering with this disease. So, it requires a quick detection method to identify the diseases in human beings. In this article [31], Harsh Panwar et al. suggested a deep transfer learning technique to notice a disease using CXR and CT scan images. This deep learning process was applied to three different datasets to obtain better classification accuracy which in terms provides the correct prediction of COVID-19 in patients. R. G. Babukarthik et al. [6] developed a Genetic Deep Learning Convolutional Neural Network (GDCNN) method for corona virus prediction. This method was developed to recognize a pneumonia because of the COVID-19 using CXR images. Generally, high dimensional features from the images are obtained using the deep learning algorithms. The proposed algorithm in this paper achieves 98.84% classification accuracy.

By using CXR images, Neha Gianchandani et al. [12] developed an ensemble deep transfer learning method to diagnose the COVID-19 pandemic. Deep methods like Densenet201 and VGG16 model were utilized and they perform better classification with pre-trained models. Viral pneumonia, COVID-19, and bacterial pneumonia were diagnosed by this method. Moreover, this paper performs both multiclass as well as binary class classification on two different datasets. Prabira Kumar Sethy et al. [38] developed a deep feature plus support vector machine (SVM) technique to identify COVID-19 in infected people using CXR images. Rather than deep learning algorithm, this research developed SVM based classifier for training and validation process. Also, CNN model was used to fetch deep features from the images to perform classification process. Additionally, this classifier classifies the image into 3 classes like normal, pneumonia, and COVID-19. This method was very helpful to doctor to categorize between healthy people, corona virus patient and pneumonia patient.

For public medical system, it was great challenge as highly infectious people with COVID-19 continuously flood into hospitals. So, it requires to find the severity of the diseases in human being based on the priority of treatment. So, Fang et al. [11] suggested a deep learning method to sense corona disease in earlier time to reduce a count of death. Clinical data of outpatients and CT scans were collected to perform this detection process. Akib Mohi Ud Din Khanday et al. [19] developed a machine learning algorithm for corona disease prediction using medical text data. AI technique has been established to verdict the current deadly virus to control the spreading nature of that virus. Ensemble and classical machine learning methods are utilized to categorize the textual hospital reports into 4 classes. In order to achieve better classification accuracy, TF/IDF, BoW and report length methods are utilized to excerpt the relevant features from a dataset.

Machine learning based predictors were suggested by Dhruv Patel et al. [32] for covid-19 disease severity prediction. Machine learning algorithm was developed to forecast a need of intensive care and mechanical ventilation using blood panel profile data, clinical data, and socio-demographic data. Furqan Rustam et al. [34] developed a machine learning algorithm based forecasting method to prove their importance to enhance the decision making on the upcoming predictions. Different kind of algorithms like exponential smoothing (ES), least absolute shrinkage and selection operator (LASSO), SVM, and LR are developed in this study to predict a number of affected patients in upcoming days.

Adel Mehrpooya et al. [27] developed a Matrix Factorization (MF) as high dimensionality depletion method for system pharmacology. For the evolution and mechanisms of drug responses and drug resistance, it is necessary to minimize high dimensional space of features in in vitro, in vivo and clinical information. Three different kinds of feature extraction techniques are proposed in this research and those concepts are performed using mathematical conception of a basis. Eight different gene dataset are analysed by this MF and other three feature selection methods.

Kamal Berahmand et al. [7] detect a community in complex networks by enlarging core nodes over enlarged local comparability of nodes. There is a dramatical growth in complex networks due to the enhancement in data storage and computational power. Local technique is proposed in this research, due to the rapid extension of complex networks. Initially, community's central node is predicted based on comparability among graph's nodes, then, total weights of weighted graph's edges are created. For weighted graphs, node's membership function is created to expand the node. The local algorithm is considered as an efficient method due to the accurate detection of node similarity by weight of the edges. MF based feature selection method is suggested by Farshad Saberi-Movahed et al. [35] to decode a clinical biomarker space of corona virus. This research also suggested a machine learning algorithm to discover the set of clinical indicators from blood tests of corona virus patients. Feature selection and Prognosis classification are two different methods presented in this approach. RF algorithm is used for classification process and MF technique is utilized for the feature selection technique.

Attributed graph embedding method was developed by Kamal Berahmand et al. [8] to construct an affinity matrix by performing a spectral clustering on protein-protein interaction networks. Both node features and topological structure are essential for protein complexes in protein-protein interaction. Spectral clustering method uses a eigen values of affinity matrix of the information to plot to low dimensional space. This clustering algorithm is considered as an effective technique in dimensionality reduction process. For attributed networks, this paper developed new version of spectral clustering known as text-associated DeepWalk-Spectral Clustering (TADW-SC) method due to structural cohesiveness and attribute homogeneity in protein complex networks. Also, text-associated DeepWalk (TADW) method is developed to compute embedding vectors of proteins. To enhance accuracy of affinity matrix, Cosine similarity is applied among the two low dimensional vectors to calculate the affinity matrix. To accurately control the COVID-19 spread, it is essential to develop real-time prediction models that can help different decisions both at international as well as national levels. So, it is necessary to develop real-time prediction models. Michał Wieczorek et al. [43] suggested a real-time neural network-based predictor for COVID-19 virus spread. ANN prediction model is developed to compute the COVID-19 spread and this model also act as online system as a real-time predictor. Numerical data and Geo-location from past 2 weeks are utilized to compute the future situation of the COVID-19 spread.

Madini O. Alassafi, et al. [28] developed a deep learning algorithm for COVID-19 time series prediction process. In order to perform the prediction, this research utilizes publicly available database provided by the European Centre for Disease Prevention and Control. The probable number of COVID-19 cases are predicted with the help of different deep learning algorithms like recurrent neural network (RNN) and long short-term memory (LSTM) networks. Sarwan Ali et al. [3] efficiently analyse the COVID-19 clinical data using machine learning algorithms. In big data analysis, usage of machine learning algorithm is considered as a natural approach which quickly extract the relevant information from the dataset to perform the data into fixed-length feature vector after that, feature selection technique is applied to perform the next process. For classification purpose, different kind of machine learning algorithms are applied on two clinical datasets. Comparative analysis of existing methods is listed in Table 1.

From the comparison analysis, it is shown that most of the existing methods have limited access on datasets. This is considered as one of the big motivations to our work. Also, most of the research works utilized machine learning concepts to predict the virus in patient. When compared to machine learning, deep learning analysis provides better prediction accuracy on large scale datasets. Moreover, the proposed hybrid deep learning algorithm tries to learn a relevant feature of data plus takes less time to complete the training and testing process.

3 Proposed methodology

Developing a hybrid deep-learning technique is a goal of this research to find out the COVID-19 positive cases, mortality rate as well as recovery rate on two different kind of datasets. This method significantly predict mortality rate, and recovery rate using the largest array of clinical variables. System model of proposed methodology is displayed in Fig. 1.

Initially, the dataset is pre-processed using different methods to remove the unwanted data like prepositions, links, hashtags etc. After that, different kind of features are extracted using

		g recumdues		
SI. no	Author name and year	Methods	Merits	Demerits
1 2	Harsh Panwar et al. [31], 2020 R. G. Babukarthik et al. [6], 2020	GDCNN	Proposed model is quicker than RT-PCR testing kit Faster and minimal cost	Limited dataset only used by this method Need to achieve higher classification accuracy with Jaros scole doneset
3	Neha Gianchandani et al. [12], 2020	ensemble deep transfer learning method	Achieves higher classification accuracy	Limited dataset only used
4	Prabira Kumar Sethy et al. [38], 2020	SVM	Can be used by the medical practitioner to classify the disease	If the patient is in a critical situation, he can't be able to attend a X-Ray scanning and this is considered as limitation of this research
5	Fang et al. [11], 2021	Domain adaptation approach	Improves emergency response capacity of medical framework	Limited samples are only available for malignant progression prediction
9	Akib Mohi Ud Din Khanday et al. [19], 2020	Ensemble and classical machine learning algorithms	Categorize the given dataset into 4 different classes	Does not considered deep learning approach
2	Dhruv Patel et al. [32], 2021	Different machine learning algorithms	Analyze multiple features to categorize the data	Limited number of datasets
8	Furqan Rustam et al. [34], 2020	Different machine learning algorithms	Forecasts a threatening factor of COVID-19 effi- ciently	Only use limited number of records
6	Michał Wieczorek et al. [43]	ANN	Easily accepts to new information and provide correct medictions	Need automatic adjustment to the newly incoming data
10	Madini O. Alassafi, et al. [28], 2022	LSTM, RNN	Predicts the probable number of COVID-19 cases	Limits the efficiency of the system
11	Sarwan Ali et al. [3], 2022	Different machine learning algorithms	Few-short learning to perform predictive analysis on COVID-19 data	Limited medical resources

 Table 1
 Comparative analysis of existing techniques



Fig. 1 System model of proposed methodology

TF-IDF and BoW model. Then, MO algorithm is utilized to select relevant features from large number of features. Finally, deep learning models like ResNet and GoogleNet are combined to achieve prediction process. This hybridization process effectively performs the detection/ prediction process and gives maximum accuracy on data prediction. The variables and their descriptions are defined in Table 2.

3.1 Pre-processing

Pre-processing is defined by transforming raw data into an understandable format also it is a data mining technique. The text pre-processing eliminates the links, prepositions, abbreviations, and hash tags. After data acquisition, it should be correctly pre-processed with different methods. So, pre-processing stage is vital to eliminate variability and unwanted data in the input data. The following difficulties are processed by the preprocessing step: (i) elimination of punctuations, (iii) elimination of stop-words such as, prepositions, articles, etc., (iv) removal of white spaces, unnecessary tabs and URLs, etc. and (v) alteration of upper-case letters to lower case [24]. The pre-processing is achieved by using the following procedures.

3.1.1 Elimination of stop words

Stop words are some English words which doesn't give proper meaning to a sentence. Without losing the meaning of the sentence, these meaningless words are eliminated. The, he, have etc. are the examples for this word. Removing repeated words in sentence is called normalization process. Repeated words in social media text are not not matched with all dictionary words and are very difficult to deal with. The proposed technique make use of this pre-processing step to minimize all these occurrences. Some examples are, 'play' can be written as "played", "playing" and so on (reduced to play) [5].

Variables	Descriptions
$v_i(t)$	Current velocity
$p_i(t)$	Position
$f(x_{h_i})$	finest fitness value
$f(x_i)$	present fitness value
a_1 and a_2	Positive attraction constants
r_p	Cartesian distance among the historical finest position and individual position
$\dot{r_g}$	Cartesian distance among the individuals and its global finest location
\tilde{eta}	Fixed visibility coefficient
d	Arbitrary dance factor
r_1, r_2	Arbitrary number between -1 and 1
fl	Random walk coefficient
r _{mf}	Cartesian distance among male and female mayflies
L	Random value
S	Amount of features in the selected feature subset
D	Total amount of features in dataset
η	Classification error
Xi	Feature map
X^{i-1}	Input data
W ⁱ	Weight of filter
β^i	Bias term
ϕ	Rectifier or ReLU activation function
*	Convolution operator
x	Input data
AvgP	Average pooling function

 Table 2
 List of variables and their descriptions

3.1.2 Stemming

Process of dropping changed arguments to their base is defined as stemming, word stem, or root form. Stem need not be similar to the morphological root of the word, even if this stem is not in itself a valid root. Root is a stem of the word which created the word. Transformation of all the words in a text to their root or stem form is known as the procedure of stemming. For example, the stem for the words "playing", "played", and "player" is "Play" [30].

3.2 Feature extraction

Different kind of features are extricated from the pre-processed hospital data and are transformed into probabilistic values. Methods like TF-IDF and BoW are utilized to excerpt the features from the publicly available dataset. TF-IDF normally used to identify the feature from the dataset and unigram, bigram and BoW are used to extract the features.

3.2.1 TF-IDF features

TF-IDF procedure transform manuscript into an eloquent illustration of numbers. The features are extracted from various NLP applications by using this technique. It is recommended to convert the text into some vector or numeric illustration to use statistical technique on any kind of text. The numeric representation depicts the substantial characteristics of the text.

TF Higher value of a word means a greater importance in the given text. Naturally, the larger size documents must have more occurrences of the word with the size of the collection and it is termed as TF.

The TF of a word is demonstrated in the given below equation:

$$tf(w) = doc.count(w)/totalwordsincorpus$$
(1)

IDF Every single term has equal significant and specified a chance to contribute in vector illustration while computing term frequency. But particular words are so mutual in pamphlets and they donate very little in determining the meaning of it. For example, 'a', 'the', 'in', 'of' etc. are term frequency words and they overcome the weights of more meaningful words. To decrease this result, TF is reduced by a factor called IDF.

IDF(w) = log (total no of pamphlets/number of pamphlets containing word w) (2)

TF-IDF If one term occurs regularly in a particular document, then the vector representation assigns high values for a given term. The idf computed value will be zero, when the given term occurs in every single document. The dot product of TF-IDF is given in the below expression.

$$tf - idf(w) = idf(w) * tf(w)$$
(3)

More significant word in the phamplet get a higher TF-IDF score and vice versa. Finally, words in the tweets are represented as vector format [9].

3.2.2 BoW features

One of the commonly used text representations for various linear classifier is the BoW model. The BoW approach introduced in this research work improves the predictive power of low frequency. Further, the BoW model is enriched for out of vocabulary words with the relevant text consist in Word Vector model. This model is constructed based on a principle of encapsulating word co-occurrence data and trained on either domain specific or else general corpus. Also, consider that the contextually related words could be commonly took place with same set of surrounding words and the sequence of individual word tokens T = [t1, t2, ...] is defined by a text T. Every single text T is mapped to huge vector MT of dimension |V| by BoWs model, here V is a vocabulary that is defined as a set of all tokens combined with training data. To each instance of training set, this labelled training data of single class label is mapped. Frequency of each token from V that happens in T is included in BoW vector MT, here 0 is the vector value for each token in V that not transoire in T. Therefore, there is no element of vector for the token that does not present in training corpus's vocabulary. BoW model considers that the tokens frequencies and presence are significant for text classification and not the ordering of words [15]. Table 3 lists some of the sample features from the pre-processed words on dataset 1.

3.3 Feature selection

In data mining process, feature selection is considered as a significant pre-processing phase; which eliminates the repeated and irrelevant features from the group of features and supports to

lungs	chest	patient	multiple	peripheral	bilateral	lung	pneum	opacities
0.379	0	0	0	0	0.34539	0.379	0	0
0	0	0	0	0	0	0	0	0
0	0.612	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0.31916	0.2414	0.13287	0	0	0.16	0.119823

Table 3 Some sample features from the pre-processed words on dataset 1

accomplish higher prediction accuracy in deep learning models. In this research, MO [45] algorithm is adapted to obtain relevant features from the group of features also, this evolutionary algorithm is good for a suitable optimization task.

3.3.1 MO algorithm

MO algorithm is stimulated by a flight behaviour and mating procedure of mayflies. This algorithm merges main benefits of evolutionary and swarm intelligence algorithms. Nowadays, for an inspiring level of issues, MO algorithm has been enlarged and applied. The subsequent section offers a summary of MO algorithm.

In MO approach, the mayflies are divided into female and male individuals. In optimization process, the male mayflies perform well due to their strong nature. Based on a current velocity $v_i(t)$ and position $p_i(t)$, the MO approach update the locations in the present iteration is given by.

$$p_i(t+1) = p_i(t) + v_i(t+1)$$
(4)

Using eq. (4), all the mayflies in swarm can be upgrade their locations. But, their velocity is upgraded in several manners.

Movements of male mayflies In the course of iterations, the male mayflies will perform on exploitation or exploration process. Based on the historical finest fitness values in trajectories $f(x_{h_i})$ and present fitness values $f(x_i)$, male mayflies are updated their velocity. Based on their present velocities, the historical finest trajectories, composed with distance among them and global finest position, male mayflies are updated their velocities if $f(x_{h_i})$ greater than $f(x_{h_i})$.

$$v_i(t+1) = g.v_i(t) + a_1 e^{-\beta r_p^2} [x_{h_i} - x_i(t)] + a_2 e^{-\beta r_g^2} [x_g - x_i(t)]$$
(5)

The parameter g is linearly reduced from the highest value to a lowest value. To balance the values, the variables a_1 , a_2 (positive attraction constants), and β (fixed visibility coefficient) are employed. Here, r_p mentions Cartesian distance between historical finest position and individual position, and r_g mentions the Cartesian distance among the individuals and its global finest location. For the distance array, Cartesian distance is considered as second norm.

$$||x_i - x_j|| = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$
 (6)

If $f(x_i)$ less than $f(x_{h_i})$, then the male mayflies update their velocities from present one with arbitrary dance factor *d*.

$$v_i(t+1) = g.v_i(t) + d.r_1$$
(7)

The arbitrary number is denoted as r_1 that lie between -1 and 1.

Movements of female mayflies Female mayflies update their velocities with a various manner. Hence, female mayflies are in hurry to discover a male mayfly to mate and replicate themselves. For this reason, male mayflies need to mate, and the female mayflies must update their velocities. In this algorithm, top finest male and female mayflies are considered as initial mate as well as second finest male and female mayflies are considered as the second mates. If $f(x_i)$ is greater than $f(y_i)$, the velocity is updated for the *i*th female mayfly is given by.

$$v_i(t+1) = g.v_i(t) + a_3 e^{-\beta r_{mf}^2} [x_i(t) - y_i(t)]$$
(8)

To balance the velocities, another parameter of a_3 is employed. The Cartesian distance between them is denoted as r_{nf} . If $f(y_i)$ is less than are equal to $f(x_i)$, then the female mayflies update their velocities from the present one with random walk coefficient fl.

$$v_i(t) = g.v_i(t) + fl.r_2$$
 (9)

The arbitrary number is denoted as r_2 that lie between -1 and 1.

Mating of mayflies All female and top half male mayflies are mated and produce the couple of children for themselves. From their parents, their offspring will randomly grow.

offspring1 =
$$L^*$$
male + $(1-L)^*$ female
offspring2 = L^* female + $(1-L)^*$ male (10)

In Gauss distribution, the random value is denoted as *L*. The offspring will be mutated, and the offspring will grew if the end iteration is satisfied. They will be sorted based on their fitness values and chosen as the female or male again for the next iteration.

Fitness function Number of features and classification accuracy are two different elements in our proposed optimization algorithm. We have decided to use classification error in order to achieve higher classification accuracy with low number of features because, these two elements are opposing in nature. This is because, high accuracy is obtained with low error and it is achieved by selecting minimized fitness function value. Equation (11) shows the fitness function in a given feature subset.

$$F(x) = \omega^* \eta + (1 - \omega)^* \frac{|S|}{|D|}$$
(11)

where, |S| and |D| denotes a amount of features in a selected feature set and total amount of features in a dataset, relative weight of features and classification error is denoted as $\omega \in [0, 1]$ and classification error is mentioned as η .

3.4 Classification

A new efficient hybrid algorithm is introduced in this paper to predict the COVID-19 disease and this hybrid method combines two various deep learning models like CNN. More than one fully connected (FC) layer follows the various convolution and subsampling layers. Generally, the FC layer is assumed as normal multilayer neural network and it provides an output is termed as class score. The convolutional layers combined all input data and employ various filters to convolve the information. Next, data is down sampled by a pooling layer. Max pooling and average polling are the two kinds of functions in the pooling layer. Then CNN changes input data on various stacked layers to accomplish a final class score. CNN frameworks are employed as building blocks for various semantic segmentation models. In the proposed work, two various deep learning algorithm such as ResNet [14] as well as Inception model or GoogleNet [45] model are combined to execute the prediction and mortality of the COVID-19 disease as well as recovery rate prediction in two various datasets. The next section has given the description of the Inception and ResNet model.

3.4.1 ResNet

A ResNet model has been developed by Microsoft researchers in 2016 that provides the highest classification accuracy of 96.4% and gained ImageNet Large Scale Visual Recognition Competition (ILSVRC). It involves an unique structure that provides residual blocks and also it contains 152 deep layers as exposed in Fig. 2.

Moreover, to discourse the issue of training a deep framework, this model employs the identity skip connections. Input of the layers are copied and forwarded to next layer by a residual function. Using an identity skip connection, a vanishing gradient problem is solved where a future layer acquires something various from the acquainted input.

3.4.2 GoogleNet model

GoogleNet [40] architecture was developed on 2014 by Google researchers which is also named as Inception model. It achieves 93.3% classification accuracy and won ILSVRC with top 5 position. The 1×1 convolution operation is performed by the convolution layers for dimensionality reduction. Number of parameters and operations are significantly reduced by dimensionality reduction and parallelism process by saving memory and computational cost. Figure 3 illustrates the Inception model architecture.

3.4.3 Hybrid learning algorithm for corona virus prediction

In this proposed work, merits of both models such as Inception and ResNet are integrated to increase the corona virus prediction accuracy and rate of recovery. The proposed hybrid model illustrates the ability to improvise the network layers by providing enhanced efficiency and superior performance. The difference between the hybrid models is that number of residual blocks are combined with identity mapping in residual network whereas the convolutional layers in DCN (Deep Convolution Network) incorporates the inception network. By using this hybrid deep learning algorithm, we can reduce the demerits of both learning algorithms and provide more accurate and less computationally expensive solutions. Overall, hybrid deep learning models performed better than standalone learning algorithms by means of different parameters and training time. The major goal of COVID-19 detection/prediction is to detect the recovery rate, death rate and forecast the number of positive specimens by tagging and allocating distinct data that have more frequency information. Figure 4 represents the structure of single fully connected cascaded residual block.





Previous layer

Fig. 3 Structure of GoogleNet model

1x1 Convolutions



In residual model, every single convolutional layer takes input from their prior convolutional layers. Overall structure of coupled ResNet and GoogleNet module is shown in Fig. 5.

Convolutional layers After applying convolution to the input picture, the data is transformed using a ReLU function. ReLU takes much less time to train than sigmoidal functions. Each kernel's operation is expressed in the following eq. (12).

$$X^{i} = \phi \left(W^{i*} X^{i-1} + \beta^{i} \right) \tag{12}$$

To obtain a feature map X^i , convolution operator convolves the filter W^i with input data X^{i-1} and adds bias term β^i then, apply rectifier operation ϕ also, * denotes the convolution operator. ReLU activation function ϕ initiates elementwise operation on input datax and it is depicted in eq. (13).



$$\phi(x) = \max(x, 0) \tag{13}$$

Figure 6 illustrates the architecture of hybrid ResNet-GoogleNet model that contains only 2 residual blocks.

Expression (14) defines the function of upper residual model in which the 2 residual models are ultimate connected in the network.

$$X^{1} = \phi(W^{1*}X^{0} + \beta^{1})$$

$$X^{2} = \phi(W^{2*}(X^{0} + X^{1})\beta^{2})$$

$$X^{3} = \phi(W^{3*}(X^{0} + X^{1} + X^{2})\beta^{3})$$

$$X^{4} = \operatorname{AvgP}(X^{3})$$
(14)

Lower residual model functions are defined in eq. 15.

$$X'^{1} = \phi \left(W'^{1*} X^{0} + \beta^{1} \right)$$

$$X'^{2} = \phi \left(W'^{2*} \left(X'^{0} + X'^{1} \right) \beta^{2} \right)$$

$$X'^{3} = \phi \left(W'^{3*} \left(X'^{0} + X'^{1} + X'^{2} \right) \beta^{3} \right)$$

$$X'^{4} = \operatorname{AvgP} \left(X'^{3} \right)$$
(15)

First 3 lines in the equation mentions the convolution function in which 3rd convolution layer X^3 and X^3 give outputs to average pooling layer and finally concern dropout procedure.

Pooling layers With stride and filter size of 2, one pooling layer performs average pooling operation and the average pooling function is displayed in expression (16).

$$X^{i} = A v g P(X^{i-1}) \tag{16}$$

🖄 Springer



Fig. 6 Whole architecture of hybridized deep ResNet with GoogleNet model

where, X^{i-1} and AvgP denotes input data from previous convolutional layer and average pooling function. Interdependent learning between the neurons are reduced with the help of dropout method in neural networks.

Softmax classifier In output layer, probabilities of each and every single input element belongs to label is obtained using Softmax activation function and it signifies the absolute distribution over class labels.

4 Simulation results and analysis

Simulation analysis of corona virus prediction using hybrid deep learning method is explained in this section and Python tool is used to implement the proposed methodology. Pre-processing is performed by 'NLTK' tool and performance of hybrid technique is calculated by means of different metrics such as recall, precision, F-1 measure, detection accuracy and so on. Different algorithms like Multinomial Naive Bayesian (MNB), LR, Decision tree (DT), SVM, RF, Bagging, Adaboost, as well as Stochastic gradient boosting (SGB) [19] are used to estimate the performance of proposed methodology in first dataset. Also, four standard forecasting models like LASSO, SVM, LR, and ES [34] has been used to compare the performance of proposed methodology to forecast an aggressive features of corona virus in second dataset. Given below section describes the simulation results and their outcomes.

4.1 Description about datasets

Three different kinds of datasets were utilized by this research to effectively predict the disease in human beings. Scientists and clinical provide open access to COVID-19 information as W.H.O announced that pandemic as Health Emergency. Data was gathered from the open-source data GitHub (https://github.com/Akibkhanday/Meta-data-of-Coronavirus). This dataset contains the details of 212 patients who have indications of Corona as well as other viruses. Also, it contains patient id, age, sex, offset, intubated, finding, needed supplemental O2, extubated, went icu, temperature, modality, date, location, neutrophil count, filename, DOI, URL leukocyte count, pO2 saturation, lymphocyte count, view, folder, survival, and License. This is first dataset [19]. Second dataset is obtained from Center for Systems Science and Engineering, Johns Hopkins University [34] which is accessible in GitHub repository. In GitHub storage, csse covid 19 time series file contains dataset files. The file comprises the number of confirmed cases, deaths, daily time series summary tables, and recoveries. The third dataset is taken from Indian government website. World weather page was utilized to collect the data like average temperature, average rainfall and total infected cases for the month of May, June, July and August. Simulated dataset was created for seven different cities in India and other features such as area, population and population density are obtained from the Indian government website. Table 4 shows the sample data format of second dataset which is utilized in this research [10]. Fourth dataset contains daily death cases, confirmed deaths from Saudi Arabia, Malaysia and Morocco and this dataset is taken from European Centre for Disease Prevention and Control [28].

Second dataset is again considered for performance analysis with different deep learning methods [14]. Configuration parameters of the proposed hybrid model is displayed in Table 5. In this hybrid deep learning method, 3 convolutional layer are utilized with stride size 1 as well as ReLu-Inception activation function is applied to all the convolutional layers in proposed hybrid model. Drop out method is applied with 0.25 probability value. In final point, Softmax layer is joined in the deep learning algorithm to get final output. In this architecture, batch size is set as 17 and initial learning rate is set as 0.001.

Class	State	Region	Lat	Long	1/22/20	 2/27/20
Death	Victoria	Australia	-12.4	130.84	0	 0
2000	Nan	Canada	0.000	0.00	0	 1
	Nan	Algeria	28.03	1.65	0	 19
Recovery	Colombia	Canada	49.28	-123.1	0	 4
	Victoria	Australia	-37.8	144.96	0	 70
	Nan	Algeria	28.03	1.65	0	 65
New confirmed	Nan	Afghan	33.0	65.00	0	 74
	Victoria	Australia	-37.8	144.96	0	 411
	Nan	Algeria	28.03	1.65	0	 264

 Table 4
 Sample data from time-series dataset 2

Layer name	Size of Data	Size of kernel	Filter amount	Stride	Activation
Input	28*38	_	_	_	_
Convolution	24*24	5*5	9	1	Relu-Inception
Convolution	12*12	3*3	9	1	Relu-Inception
Convolution	10*10	2*2	9	1	Relu-Inception
Max pooling	5*5	3*3	2	2	Relu-Inception
Drop out (0.25)	5*5	_	_	_	-
Output	_	_	-	_	Softmax

Table 5 Configuration parameters and their values for proposed hybrid deep learning model

4.2 Performance metrics

For the simulation analysis, confusion matrix is utilized to derive the performance parameter. False Negative (FN), True Positive (TP), False Positive (FP), and True Negative (TN) parameters of confusion matrix were used to compute the metrics presented in this paper [32]. Ratio of recognized instances that are significant is defined as precision, ratio of important instances that are retrieved is mentioned as recall. Correctness of the classification is defined as accuracy and harmonic average of both recall as well as precision is denoted as F1-score. Accuracy, precision, recall, and F1-score are computed using the eqs. (17–20).

$$Precision = \frac{TP}{TP + FP}$$
(17)

$$Recall = \frac{TP}{TP + FN}$$
(18)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
(19)

$$\frac{F1_Score = 2^*Precision^*Recall}{Precision + Recall}$$
(20)

Percentage of variation independent variable is denoted as R^2 which is a linear model and it is represented in eq. (21).

$$\frac{R^2 = \text{variance explained by model}}{\text{Totalvariance}}$$
(21)

Improved form of R^2 is the Adjusted R-squared ($R^2_{adjusted}$) that also like R^2 displays how well the data points suitable to the curve and this is defined in eq. (22).

$$R_{adjusted}^2 = 1 - (1 - R^2) \frac{n - 1}{n - (k + 1)}$$
(22)

where, number of independent variables in regression expression is mentioned as k and n is sample size. Average magnitude of errors in model prediction is defined as mean absolute error (MAE).

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left(\left| \widehat{y}_i \neg y_i \right| \right)$$
(23)

The amount of error in statistical model is measured by the Mean square error (MSE). Generally, MSE defines the average square difference among the actual output value as well as predicted output value. The given below expression is used to calculate the MSE.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i \right)^2$$
(24)

Standard deviation of the prediction error is defined as the Root Mean Square Error (RMSE) and it is defined in eq. (25).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2}$$
(25)

Measure of prediction accuracy of a forecasting method in statistics is defined as Mean Absolute Percentage Error (MAPE). It is calculated based on the given below expression.

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\left(\left| \widehat{y}_i - y_i \right| \right)}{y_i} \right) *100\%$$
(26)

Here, \hat{y}_i mentions the actual output and y_i represents the predicted output.

4.3 Description about existing methodologies

This section briefly explains the process of different existing methods which are utilized for the comparison purpose. SVM is a supervised machine learning method to categorize text into various class and for that process, it takes n number of features from specific text with given label [20]. An ensemble machine learning algorithm is helpful to enhance the performance of other regression and classification of machine learning techniques [13]. Overfitting problem is avoided with the help of Bagging algorithm. This ensemble learning algorithm works with weighted instances of the dataset [39]. For classification purpose, ensemble machine learning algorithms are works like a DT. RF method is trained with the help of bootstrap aggregating algorithm. Averaging the predictions of all individual regression trees are used to make the overall prediction [23]. Boosting method permits tree to be generated from sample training dataset and it helps to minimize correlations among the trees in gradient boosting algorithm. From the full training dataset, a subsample of training data is randomly selected at every single iteration. In order to fit base learner, randomly chosen subsample is used rather than whole sample [29].

4.4 Performance analysis on dataset 1

In this research, dataset is separated into 70:30 ratio in which training utilize 70% data and testing process utilize remaining 30% data. The proposed model has been compared with machine learning algorithms like LR, MNB, SVM, DT, Bagging, Adaboost, RF, and SGB [19] to show the efficiency of the proposed model. Table 6 lists a comparison of proposed methodology with conventional methodologies by means of precision, recall, accuracy, and F1 measure.

Existing methods	Precision	Recall	F1 score	Accuracy (%)
LR	0.94	0.96	0.95	96.2
MNB	0.94	0.96	0.95	96.2
SVM	0.82	0.91	0.86	90.6
DT	0.92	0.92	0.92	92.5
Bagging	0.92	0.92	0.92	92.5
Adaboost	0.85	0.91	0.88	90.6
RF	0.93	0.94	0.93	94.3
SGB	0.93	0.94	0.93	94.3
Proposed model	0.95	0.97	0.97	97.56

 Table 6
 Comparison analysis of hybrid methodology with different machine learning algorithms on dataset 1

Figure 7 displays the performance analysis of proposed technique with existing techniques by means of (a) accuracy, (b) precision, (c) recall and (d) F1 measure. Our proposed hybrid deep learning model outperforms all existing machine learning methods by means of different parameters and the performance values of proposed methodology are highlighted in Table 6. When compared to the existing machine learning algorithms, the deep learning algorithm have many advantages as it is presented with cascade of layers. So, each of its successive layers utilize the output data from the previous layers as input and when this is applied to solve variety of problems, the features and number of features presented in the algorithm provide specific advantages to those problems. Moreover, accuracy level of proposed hybrid algorithm is increased



Fig. 7 Performance analysis of proposed methodology with existing methodologies

while the number of error decreases also, classical machine algorithm takes more time to process the data which is considered as a time-consuming process.

4.5 Performance analysis on dataset 2

COVID-19 patient states dataset [34] offered by Johns Hopkins is evaluated in this section. This dataset is separated into section in which 85% of the dataset is utilized for training procedure and 15% of the dataset is utilized for testing process. Significant measures including R^2 score, $R^2_{adjusted}$, MSE, MAE and RMSE are used to analyse the performance of proposed methodology.

4.5.1 Future forecasting of death rate

The study performs predictions on death rate based on different parameters like R^2 score, $R^2_{adjusted}$, MSE, MAE and RMSE. The comparison is done against the different kind of machine learning algorithms like LR, LSSO, SVM and ES [34]. Table 7 displays the results on dataset 2.

When compared to the existing machine learning algorithms, our proposed method achieves lesser error value. In all machine learning algorithms, SVM performs worst in this situation. Moreover, machine learning algorithms require structured data for prediction as well as detection process. While, the proposed hybrid algorithm is capable of becoming smarter itself as it can work with unstructured data. This leads to achieve higher productivity in a short period of time with less error rate.

4.5.2 Future forecasting of newly infected confirm cases

Corona virus disease is enlarging day by day and it automatically increases the newly confirmed cases and the result analysis on this model is displayed in Table 8.

When compared to machine learning algorithms, the proposed hybrid model outperforms all existing methods in terms of different error calculation parameters. Machine learning algorithms like LASSO and ES lead the table by means of performance after proposed model, LR also achieved good, whereas SVM achieves very poorly by means of all the evaluation metrics Moreover, machine learning algorithms require structured data for prediction as well as detection process. While, the proposed hybrid algorithm is capable of becoming smarter itself as it can work with unstructured data. This leads to achieve higher productivity in a short period of time with less error rate.

Model	R^2 score	R ² adjusted	MSE	MAE	RMSE
LR	0.96	0.95	840,240.11	723.11	916.64
LSSO	0.85	0.81	3,244,066.79	1430.29	1801.12
SVM	0.53	0.39	16,016,210.98	3129.74	4002.02
ES	0.98	0.97	66,228.72	406.08	813.77
Proposed model	0.997	0.987	58,990.9	359.87	786.98

 Table 7
 Death rate future forecasting

Model	<i>R</i> ²	$R^2_{adjusted}$	MSE	MAE	RMSE
LR	0.83	0.79	1,472,986,504.96	30,279.55	38,390.51
LSSO	0.98	0.97	234,489,560.99	1163.97	15,322.11
SVM	0.59	0.47	5,760,890,969.3	60,177.90	75,911.28
ES	0.98	0.97	283,201,302.2	8867.43	16,828.58
Proposed model	0.995	0.98	213,425,689.2	995.89	12,879.99

4.5.3 Future forecasting of recovery rate

The proposed hybrid algorithm performs better than other existing machine learning models in recovery rate future forecasting and result analysis are displayed in Table 9.

When compared to all other existing machine learning models, the proposed hybrid model performs better prediction rate in terms of newly infected cases. Moreover, machine learning algorithms require structured data for prediction as well as detection process. While, the proposed hybrid algorithm is capable of becoming smarter itself as it can work with unstructured data. This leads to achieve higher productivity in a short period of time with less error rate.

4.6 Performance analysis on third dataset

Predicting the status of infection in India is considered as a pivotal role as India is one of the most affected countries that have higher infected cases. For taking preventive measures, prediction tools support the public health administration to identify the number of cases in the near future. For analysis, second dataset is splits into two in which 80% of the data is used for training and the remaining 20% is used as test data. Performance of proposed model is compared with existing algorithms like Long Short-Term Memory (LSTM), Stacked LSTM (SLSTM), Auto-Regressive Integrated Moving Average (ARIMA), and Prophet approaches [10] by considering different cases like confirmed, death and recovered global cases.

Table 10 displays the comparison analysis of future forecasting data which contains confirmed, death and recovered cases. Different kinds of error calculation methods like RMSE, MAE and MAPE are analysed for confirmed, deaths and recovered cases. When compared to existing methods like LSTM, SLSTM, ARIMA and PROPHET, the proposed method takes less error rate for forecasting the data. The reason for less error rate is that the proposed hybrid model gives better results with unstructured data and efficiently deliver high quality results in prediction process.

Model	R ²	_	MSE	MAE	RMSE
		$R^2_{adjusted}$			
LR	0.39	0.21	480,922,814.51	17,016.08	21,929.95
LSSO	0.29	0.08	1,462,144,344.82	30,705.27	38,237.99
SVM	0.24	0.02	13,121,148,615.72	106,739.82	114,547.58
ES	0.99	0.99	5,970,634.07	1827.85	2443.48
Proposed model	0.998	0.998	4,980,987.90	1678.9	1899.90

 Table 9 Future forecasting for recovery rates

Models	Predicted variables	RMSE	MAE	MAPE
ARIMA	Confirmed Deaths	194.76	70.94	2.08
	Recovered	2324.9	1350.86	2.76
		2178.86	1305	1.21
LSTM	Confirmed Deaths	1167.56	979.03	0.43
	Recovered	992.84	866.66	1.9
		1670.84	1366.66	2.4
SLSTM	Confirmed Deaths	274.22	920.02	0.3
SLSTM	Recovered	309.12	278.29	0.6
		1125.47	864	1.8
PROPHET	Confirmed Deaths	9970.33	7231.1	2.8
	Recovered	1843.71	1389.5	3.7
		9370.83	7633.49	2.7
Our proposed model	Confirmed Deaths	190.32	67.900	0.29
* *	Recovered	1190	250.98	0.54
		1003.45	789.89	1.00

Table 10 Performance analysis of future forecasting on third dataset collected from Indian government website

4.7 Performance analysis on fourth dataset

The performance analysis on fourth dataset is explained in this subsection. Here, the proposed method is compared with RNN, and LSTM methods to show the efficiency of the system. Generally, overfitting can be occurred due to the increase of validation loss on deep and machine learning algorithms. In order to avoid this, the number of epochs are set as high as possible to avoid overfitting. In this research, epoch is set as 50 and the analysis is taken for the 50 epochs on existing methods with different activation function. Table 11 lists the simulation results on various parameters.

From the table analysis, it is shown that the proposed model is compared with different existing methods like RNN and LSTM with different activation functions. In this analysis, it is shown that the proposed design consumes less error rate when compared to other existing methods with different activation function. In this research, the proposed hybrid model performs the classification with ReLU-Inception activation function and achieves less classification error rate on different parameters. The performance values of proposed methodology is highlighted in Table 11.

4.8 Discussion of the results and limitation of the research

This section briefly discusses the results which are obtained from the result analysis section and the limitation of the research briefly. In this research paper, the performance of proposed methodology is evaluated with 2 different datasets by means of different performance parameters. For dataset 1, we have taken precision, recall, F1 score and accuracy as performance parameters also, it is compared with existing machine learning algorithms like LR, MNB, SVM, DT, Bagging, Adaboost, RF, and SGB for disease prediction using time-series data. In terms of accuracy calculation, the proposed hybrid model consumes 97.56% prediction accuracy which is comparably higher than the existing classical machine learning algorithms. Also, the proposed hybrid model produces better results in other parameters like precision, recall and F1 score. This is due to the advantages of deep learning algorithm over the machine learning methods.

Algorithm names	MAE	MSE	RMSE	MAPE%
RNN with Sigmoid function (50 epochs)	2.121	1.477	1.917	97.34
RNN with Tanh function (100 epochs)	3.814	8.766	7.625	93.2
LSTM with ReLU function (70 epochs)	0.544	1.604	0.726	99.27
Proposed model (50 epochs)	0.458	1.247	0.689	99.56

Table 11 Simulation results of proposed and existing methods with different performance parameters

The second dataset utilized in this research paper is about the future forecasting prediction of confirmed, deaths and recovered cases against some machine learning and deep learning algorithms. For second dataset, proposed hybrid model is evaluated by means of different error parameters like RMSE, R², R² adjusted, MSE, and MAE with different existing algorithms like LR, LSSO, SVM, and ES models. Overall, the proposed model outperforms all the existing methods by means of different error parameters. With third dataset, different kind of error parameters are taken to evaluate the proposed hybrid model with existing deep learning algorithms like ARIMA, LSTM, SLSTM and PROPHET. In this analysis, the proposed hybrid deep learning model exceeds all other existing deep learning methods. This is due to the unstructured data which efficiently delivers high quality results with less error rate.

The fourth dataset is about the time series prediction of COVID-19 cases using deep learning algorithms. The proposed model is compared with deep learning algorithms like RNN and LSTM with different activation function to show the efficiency of proposed methodology also, the error calculation parameters like MSE, RMSE, MAE and MAPE are analysed to show that the proposed algorithm achieves less error rate while performing prediction process. Moreover, the proposed methodology is considered as one of the best frameworks which can help the prevention and control of corona virus spread. By this proposed hybrid deep learning model, we can predict the epidemic patterns of several infectious diseases easily but with limited prediction rate. Moreover, this method is not suitable for live forecasting and these kinds of problems are assumed as the limitation of this research. The limitation can be exceeded with the help of replacing deep learning models with more effective models. Depending on the current scenario, the meta-heuristic models are applied in the deep learning models to predict a future forecasting data as the disease is spreading very quickly. So, this can be applied in the future studies.

5 Conclusion

Due to the unavailability of drugs and vaccines, COVID-19 has shocked a whole world and it creates more economic problems and several researchers research about this deadly virus. So, hybrid deep learning model is developed in this research to predict a mortality rate, recovery rate and disease prediction in patients using two different datasets. Initially, different kind of pre-processing methods are utilized to remove the unwanted words from the dataset. Then, several features like BoWs, and TF/IDF features are being extracted from these clinical data. After that, MO algorithm is utilized to select optimal features from the set of features. Finally, hybrid deep learning algorithm is proposed to forecast the mortality, and recovery rate from forecasting dataset and predict positive cases in another dataset. The simulation result analysis proved that proposed hybrid model is better than another existing machine learning algorithm

by means of different parameters. In first dataset, the proposed method achieves 97.56% accuracy which is 1.40% greater than LR and MNB, 3.39% higher than RF and SGB as well as 5.32% higher than DT and Bagging techniques. Additionally, when compared to existing machine learning and deep learning algorithms, the proposed hybrid deep learning methodology consumes less error rate in second, third and fourth datasets with different error calculation parameters. When compared to existing machine learning models, the simulation result indicates that proposed hybrid deep learning method is valuable in COVID-19 diagnosis and future mortality prediction analysis. This study forecasts the predictions to take timely actions and make decisions to reduce COVID-19 crisis. In future, we have planned to explore the prediction methodology using the updated dataset for forecasting and real-time live forecasting will be considered as another one future work.

Authors contributions All authors read and approved the final manuscript.

Funding No funding is provided for the preparation of manuscript.

Data availability statement Data sharing is not applicable to this article.

Declarations

Conflict of interest Authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate All the authors involved have agreed to participate in this submitted article.

Consent to publish All the authors involved in this manuscript give full consent for publication of this submitted article.

References

- Achterberg MA, Prasse B, Ma L, Trajanovski S, Kitsak M, Van Mieghem P (2020 Oct 9) Comparing the accuracy of several network-based COVID-19 prediction algorithms. Int J Forecast
- Alakus TB, Turkoglu I (2020 Nov 1) Comparison of deep learning approaches to predict COVID-19 infection. Chaos Solit Fractals 140:110120
- Ali S, Zhou Y, Patterson M (2022 May 4) Efficient analysis of covid-19 clinical data using machine learning models. Med Biol Eng Comput 60:1–6
- Alyasseri ZA, Al-Betar MA, Doush IA, Awadallah MA, Abasi AK, Makhadmeh SN, Alomari OA, Abdulkareem KH, Adam A, Damasevicius R, Mohammed MA (2021) Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. Expert Syst 39(3):e12759
- Angadi S, Reddy RV (2018) Enhanced framework for sentiment analysis in text using distance based classification scheme. In 2018 second international conference on advances in electronics, computers and communications (ICAECC) 1-6
- Babukarthik RG, Adiga VA, Sambasivam G, Chandramohan D, Amudhavel J (2020 Sep 21) Prediction of covid-19 using genetic deep learning convolutional neural network (GDCNN). IEEE Access 8:177647– 177666
- Berahmand K, Bouyer A, Vasighi M (2018 Dec 7) Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes. IEEE Trans Comput Social Syst 5(4): 1021–1033

- Berahmand K, Nasiri E, Li Y (2021 Nov 1) Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. Comput Biol Med 138:104933
- Dabiri S, Heaslip K (2019) Developing a twitter-based traffic event detection model using deep learning architectures. Expert Syst Appl 118:425–439
- Devaraj J, Elavarasan RM, Pugazhendhi R, Shafiullah GM, Ganesan S, Jeysree AK, Khan IA, Hossain E (2021 Feb 1) Forecasting of COVID-19 cases using deep learning models: is it reliable and practically significant? Results in Physics 21:103817
- 11. Fang C, Bai S, Chen Q, Zhou Y, Xia L, Qin L, Gong S, Xie X, Zhou C, Tu D, Zhang C (2021 Aug 1) Deep learning for predicting COVID-19 malignant progression. Med Image Anal 72:102096
- Gianchandani N, Jaiswal A, Singh D, Kumar V, Kaur M (2020 Nov 16) Rapid COVID-19 diagnosis using ensemble deep transfer learning models from chest radiographic images. J Ambient Intell Humaniz Comput:1–3
- 13. González-Recio O, Jiménez-Montero JA, Alenda R (2013 Jan 1) The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. J Dairy Sci 96(1):614–624
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition 770–778
- Heap B, Bain M, Wobcke W, Krzywicki A, Schmeidl S (2017) Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems. arXiv preprint arXiv: 1709.05778
- Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B (2020 Dec 1) Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform 144:104284
- Huang J, Zhang L, Liu X, Wei Y, Liu C, Lian X, Huang Z, Chou J, Liu X, Li X, Yang K (2020 Nov 30) Global prediction system for COVID-19 pandemic. Sci Bull 65(22):1884–1887
- Huang CJ, Chen YH, Ma Y, Kuo PH (2020 Jan 1) Multiple-input deep convolutional neural network model for covid-19 forecasting in China. MedRxiv
- Khanday AM, Rabani ST, Khan QR, Rouf N, Din MM (2020 Sep) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731–739
- Khanday AM, Khan QR, Rabani ST (2021) SVMBPI: support vector machine-based propaganda identification. In: Cognitive informatics and soft computing, springer, pp 445–455
- Kumar A, Gupta PK, Srivastava A (2020) A review of modern technologies for tackling COVID-19 pandemic. Diabetes Metab Syndr: Clin Res Rev 14(4):569-573
- 22. Kumar V, Singh D, Kaur M, Damaševičius R (2021 May 26) Overview of current state of research on the application of artificial intelligence techniques for COVID-19. PeerJ Comput Sci 7:e564
- Lin W, Wu Z, Lin L, Wen A, Li J (2017 Aug 10) An ensemble random forest algorithm for insurance big data analysis. IEEE access 5:16568–16575
- Magesh S, Niveditha VR, Rajakumar PS, Natrayan L (2020 Aug 6) Pervasive computing in the context of COVID-19 prediction with AI-based algorithms. Int J Pervasive Comput Commun 16:477–487
- 25. Majeed T, Rashid R, Ali D, Asaad A (2020 Jan 1) Covid-19 detection using cnn transfer learning from x-ray images. medRxiv
- Mandal M, Jana S, Nandi SK, Khatua A, Adak S, Kar TK (2020 Jul 1) A model based study on the dynamics of COVID-19: prediction and control. Chaos Solit Fractals 136:109889
- Mehrpooya A, Saberi-Movahed F, Azizizadeh N, Rezaei-Ravari M, Eftekhari M, Tavassoly I (2021 Jan 1) High dimensionality reduction by matrix factorization for systems pharmacology. bioRxiv
- Mo A, Jarrah M, Alotaibi R (2022 Jan 11) Time series predicting of COVID-19 based on deep learning. Neurocomputing 468:335–344
- Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards TC Jr (2006 Nov 16) Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol Model 199(2):176–187
- 30. Orellana G, Arias B, Orellana M, Saquicela V, Baculima F and Piedra N (2018) A study on the impact of pre-processing techniques in Spanish and English text classification over short and large text documents. In 2018 international conference on information systems and computer science (INCISCOS) 277-283
- Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V (2020 Nov 1) A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images. Chaos Solit Fractals 140:110190
- Patel D, Kher V, Desai B, Lei X, Cen S, Nanda N, Gholamrezanezhad A, Duddalwar V, Varghese B, Oberai AA (2021 Feb 25) Machine learning based predictors for COVID-19 disease severity. Sci Rep 11(1): 1–7, 4673.
- Rajesh A, Pai H, Roy V, Samanta S, Ghosh S (2020) CoVID-19 prediction for India from the existing data and SIR (D) model study. medRxiv: 1–14

- Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, Choi GS (2020 May 25) COVID-19 future forecasting using supervised machine learning models. IEEE Access 8:101489–101499
- 35. Saberi-Movahed F, Mohammadifard M, Mehrpooya A, Rezaei-Ravari M, Berahmand K, Rostami M, Karami S, Najafzadeh M, Hajinezhad D, Jamshidi M, Abedi F (2021 Jan 1) Decoding clinical biomarker space of covid-19: exploring matrix factorization-based feature selection methods. medRxiv
- 36. Santosh KC (2020 Sep) COVID-19 prediction models and unexploited data. J Med Syst 44(9):1-4
- Sen S, Saha S, Chatterjee S, Mirjalili S, Sarkar R (2021 Apr 19) A bi-stage feature selection approach for COVID-19 prediction using chest CT images. Appl Intell 51:8985–9000
- Sethy PK, Behera SK (n.d.) Detection of coronavirus disease (covid-19) based on deep features and Support Vector Machine
- Shahraki A, Abbasi M, Haugen Ø (2020 Sep 1) Boosting algorithms for network intrusion detection: a comparative evaluation of real AdaBoost, gentle AdaBoost and modest AdaBoost. Eng Appl Artif Intell 94: 103770
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition 1–9
- Tabik S, Gómez-Ríos A, Martín-Rodríguez JL, Sevillano-García I, Rey-Area M, Charte D, Guirado E, Suárez JL, Luengo J, Valero-González MA, García-Villanova P (2020 Nov 10) COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. IEEE J Biomed Health Inform 24(12):3595–3605
- Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, Cai M, Yang J, Li Y, Meng X, Xu B (2021 Feb 24) A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). Eur Radiol 31:6096– 6104
- 43. Wieczorek M, Siłka J, Połap D, Woźniak M, Damaševičius R (2020 Dec 17) Real-time neural network based predictor for cov19 virus spread. PLoS One 15(12):e0243189
- 44. Zareie B, Roshani A, Mansournia MA, Rasouli MA, Moradi G (2020 Jan 1) A model for COVID-19 prediction in Iran based on China parameters. MedRxiv
- 45. Zervoudakis K, Tsafarakis SA (2020 Jul 1) Mayfly optimization algorithm. Comput Ind Eng 145:106559

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.