Research Article

Analysis of clinical texts for prediction of COVID-19 using Bag of Words and Artificial Neural Networks

Dr. Aruna Bhat[1], Garvit Arora[2], Ashwin Joshi[3], Gaurab Bhushan Singh[4]
[1,2,3,4] Department of Computer Science , Delhi Technological University Delhi, India
Email: aruna.bhat@dtu.ac.in, garvitarora_2k17co119@dtu.ac.in, ashwinjoshi_2k17co81@dtu.ac.in , gaurabsingh55@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: In recent times, the number of COVID-19 cases worldwide has increased significantly, which led to overwhelming work for healthcare workers and forced many countries to go under complete lockdown. Thus, it is essential to find ways to detect and control coronavirus. Machine learning has shown promising results in the various medical fields by analyzing clinical data, so it is crucial to explore new ways to detect COVID-19. It is challenging to test everyone, so a controlled and automated system to detect the COVID-19 is needed nowadays. These days, an enormous amount of data related to COVID-19 is available. In this work, we propose an artificial neural network and bag of words model-based approach for classifying clinical reports of patients into four classes of the virus. The features were extracted using various techniques like Term frequency/Inverse document frequency (TF/IDF) and report length, which is then passed through a robust neural network classifier. We trained and tested different neural network models to find suitable architecture. The model showed better performance than all the classical and ensemble machine learning algorithms with an accuracy of 97.2%. It offers excellent potential for the early detection of COVID-19 and thus helps control the pandemic.

Keywords: Artificial Neural Network, COVID-19, Deep Learning

1. Introduction

1.1 General Introduction

W.H.O pronounced the COVID-19 outbreak as a major pandemic and referenced that the transmission of this deadly virus is through the respiratory organs' tract when a healthy and sound individual interacts with the infected person. The infected individual shows indications in two to fourteen days, depending upon the hatching duration of the SARS (the Severe Acute Respiratory Syndrome) and MERS.

As per W.H.O, the indications of COVID-19 cases are mild fever, cough, and breath shortness. In some cases, tiredness may occur. Many people with different illnesses like diabetes and many coronary diseases are more defenseless and more prone to this deadly virus.

The person having symptoms is analyzed based on the indications or symptoms, and his comprehensive history of the movement is tracked. Abnormal signs must be noticed distinctly of the person having symptoms and indications.

Regularly washing hands with soap for twenty seconds and maintaining social distancing of around one meter is very beneficial, and chances of getting influenced by this deadly virus are reduced majorly [1].

If you sneeze in public, just cover the nose and mouth with a tissue that is also expandable and avoid this tissue coming in contact with the nose and mouth. An airborne infection named SARS that showed up in the early twentieth century, majorly in China and covered approximately twenty-six nations by having eight thousand cases in the same year also moved from one individual to another. The primary symptoms and indications of SARS are mild fever, dry cough, looseness of the bowels.

A disease ARDS (extreme respiratory organ pain condition) primarily affects the lungs and has symptoms like skin becoming blue, pale, and weak [2].

Right now, the discovery of Covid infection 2019 (COVID-19) is one of the fundamental difficulties on the planet, given the quick spread of the sickness. "Fig. 1" shows the worldwide data regarding coronavirus. Ongoing measurements demonstrate that the quantity of individuals determined to have COVID-19 is expanding dramatically, with more than 10.6 million affirmed cases; the infection is spreading to numerous nations over the world.

1.2 Sigificance of Machine Learning and deep Learning amidst COVID-19

Aside from clinical techniques, machine and deep learning provide a great deal of help in recognizing sickness and detecting it with the assistance of pictures and clinical textual reports. They can also be used majorly for the distinguishable identification and detection of Covid-19.

Machine learning gives an energizing exhibit of apparatuses that are adaptable enough to permit their arrangement in any pandemic phase. With the vast amount of information being created while considering a virus cycle, Deep Learning considers examination and quick distinguishing proof of examples. The adaptability, capacity to adjust dependent on another comprehension of the illness cycle, personal growth as and when new information opens up, and the absence of human bias in the methodology of investigation makes Machine learning an exceptionally flexible novel apparatus for overseeing novel diseases [3].

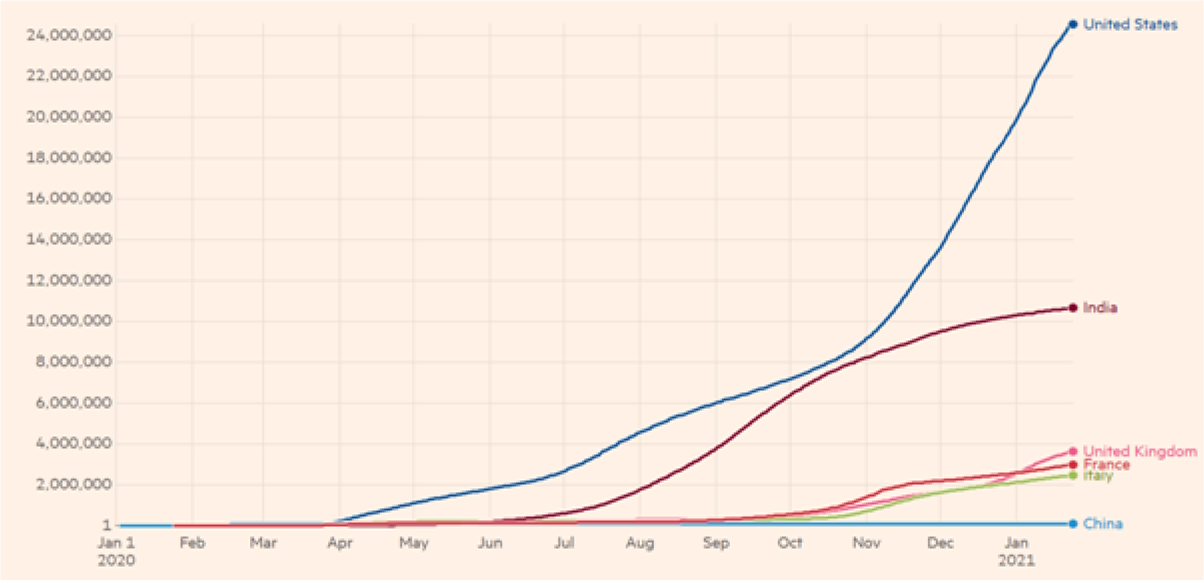

Figure 1. Worldwide Coronavirus cases as of 24th, January 2020

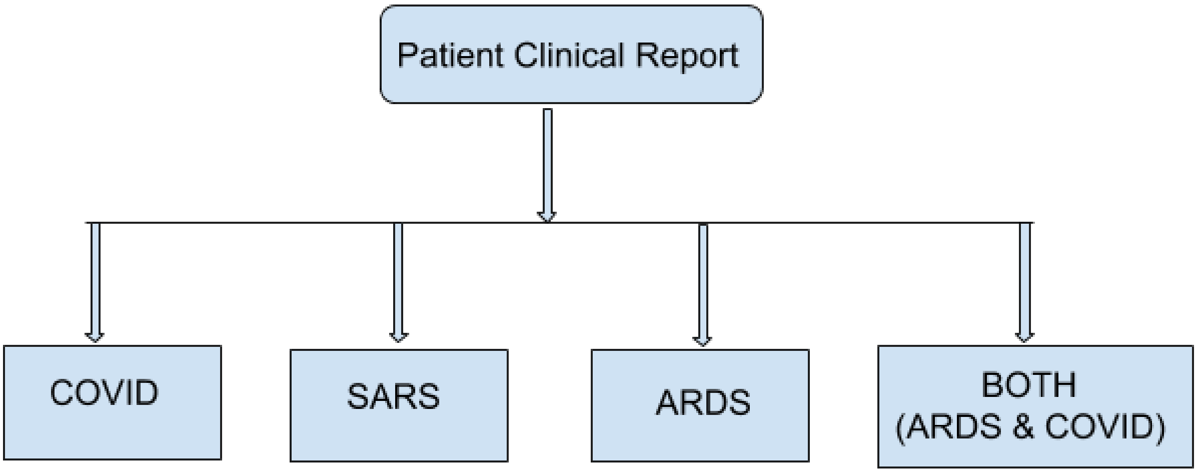

Figure **2.** Classification of Patient clinical Report

1.3 Using Deep Learning to classify Patient reports into various categories:

Natural language processing and Machine learning can be used to develop models for classifying textual data into multiple classes. But for better accuracy, we have used neural networks to classify patient reports into the following classes:

1) COVID
2) SARS
3) ARDS
4) Both (COVID, ARDS).

We have clinical reports as text, and our main motive is to classify the patient report into one of the four unique types of classes. So, in the end, we can detect whether the person is suffering from COVID or not.

2. Related work

The best way to control the current pandemic situation is to test more and find more patients to save others from contact with these patients. The test by which corona is confirmed is called Reverse transcription-polymerase chain reaction (RT-PCR). This test gives accurate results, but it is very time-consuming and requires more resources to conduct. Countries which are not having many resources are not able to conduct such tests in large number. In this situation, machine learning techniques can be helpful to find coronavirus in a human. Many researchers have worked on this problem and have found impressive results to find coronavirus in individuals using different techniques. Khanday et al. used various supervised and ensemble machine learning techniques to classify clinical textual data into four classes, namely COVID, SARS, ARDS, or both ARDS and COVID and achieved excellent results. We plan to enhance this further using more feature engineering and the concept of deep learning. Hughes et al [4] . suggested a novel method for using CNNs and embeddings for the classification of textual medical data. By using their methods, we can learn and automatically classify clinical text data into various classes. Jiang et al. tried [5] to build an artificial intelligence framework to find how much life-threatening coronavirus can become for an individual. The main aim that researchers wanted to accomplish is to identify the combinations of medical features that help prediction of patients at higher risk of having more dangerous symptoms of covid-19. The prediction model will use the historical data to learn and identify the patient with more potential to develop ARDS(acute respiratory distress syndrome). This is a life-threatening disease that gets developed due to covid-19. The study was conducted in Wenzhou, China. The artificial intelligence model identifies the features that could result in the ARDS development in patients after some time. The increase in the level of alanine aminotransferase, which is the liver enzyme, myalgias, and an increase in red blood cell level are some medical features that were found to be most predictive. The models achieved an overall 70% to 80% accuracy. In feature engineering, some algorithms are used to reduce feature space's dimension to a reduced set of more helpful features in predictive analysis. Iwendi et al. [6] try to incorporate and patient's travel dates and their medical data to detect if a person is suffering from covid-19 or not. The study was done to predict the covid-19 in patients using symptoms of covid-19 in people; the delay is done in reporting in nearby hospitals and by considering their travel history. The researchers processed both medical data and travel data. They have compared multiple algorithms that they could use, and after doing analysis, they found that boosted random forest was the best method to detect covid-19. They also performed an exercise to fine-tune the hyper-parameters of the method they used to improve the accuracy. The paper shows a way to further improve the research paper's work by making a more accurate predictive model that makes predictions based on demographics, travel, and other health data, including medical image data. Lo´pez-U´ beda et al. [7] uses textual radiological reports and detect if a person is having covid-19. The textual radiological report is also important and relevant information that can be used to detect covid-19 in a patient. The paper uses Natural Language Processing(NLP) and shows that NLP techniques could be of very much use to doctors to make decisions regarding coronavirus and answer the question if a person is suffering from coronavirus. The paper proposes a text classification system that uses information from different resources. The paper focuses on finding coronavirus in a human by analyzing the data written in a radiology report. To conduct further tests using the proposed model, researchers used around three hundred reports of radiological scans. This data is of people with a suspicion of covid-19. Researchers in this paper have applied machine learning algorithms and used entity chunking so as to train their made classification system. Two types of information as input are taken: the radiological report text and covid-19 related anomalies extracted from SNOMED-CT. The best results were achieved from SVM, with baseline results achieving an accuracy of 85%.

2.1 Limitations of existing work

In many of the research papers that we have referred to, a lot of work has already been done related to detecting covid -19 in a person using different types of image data such as X-ray reports, other types of scans of the chest of a patient, and also using radiological text reports. In one of the above research papers, researchers have also predicted the possibility of having covid – 19 based on the patient's travel history. Such methods, although help a lot in detecting whether a patient has covid-19 or not, but we cannot ignore a person's clinical text data to predict whether they are suffering from covid – 19. Clinical text data is very important data that can contain more details and thus can be very useful to accurately predict if a person is suffering from covid – 19. We decided to work on clinical text data because less research has been done to explore the use of clinical text data as compared to using X-rays or other image data in the detection of covid – 19. The work done by Khandat et al. uses
the naive Bayes classifier that worked the best. Still, there are specific issues with naive Bayes classifiers. It uses examples from the clinical textual data to learn what is in a class but doesn't learn any information about what isn't learning, which is important too.

3. Methodology

In this section, we have discussed the methodology used in the process of writing this paper. "Fig. 3" shows the methodology.The step by step process that was followed is described here. First, we collected data, and then from that data, we collected relevant information. We then preprocessed the text collected. After preprocessing was done, we did feature extraction, then classification was done using the neural networks. We have discussed the various subsections of the methodology in detail below.

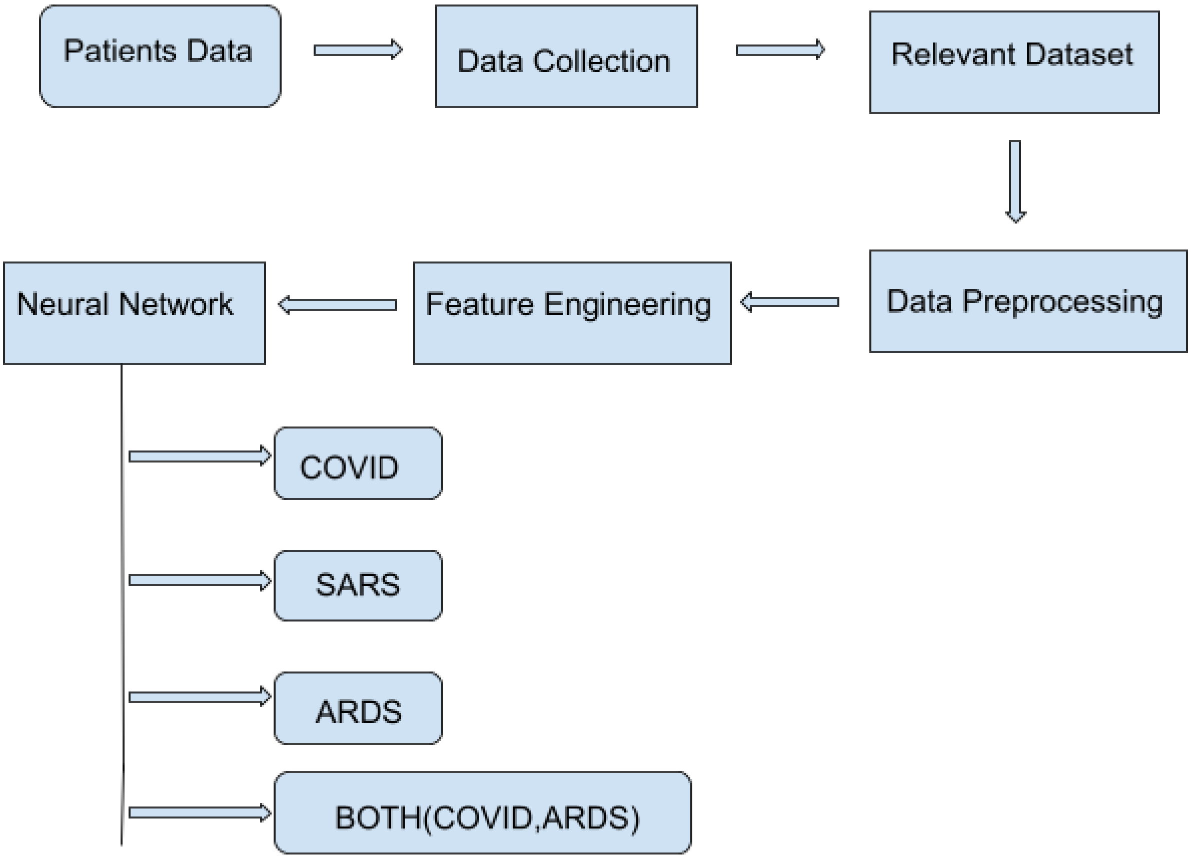

Figure 3. Methodology

3.1 Data collection

There is a lot of research going on the coronavirus pandemic. Since the coronavirus started in China, a lot of data has been made available to the public by many hospitals and different organisations such as the World Health Organization. We have used the data available in a Github repository which is freely available. The data is a text data of 892 patients. These patients who have shown symptoms of coronavirus or some other virus such as SARS, ARDS or patients having both coronavirus and ARDS. There are 24 attributes in the data set. The attributes that are in the data set are patient id, offset, sex, age, finding, RTPCR positive, survival, intubated, intubation present, whether a patient went to ICU or Not, in ICU, needed supplemental oxygen, extubated, temperature, PO2 saturation, leukocyte count, neutrophil count, lymphocyte count, view, modality, date, location, folder, file name, DOI, URL, license, clinical notes, other notes.

3.2 Extracting Relevant dataset

In our data set, we have around twenty-four attributes. All the attributes of the data set are not useful for us to work on. So, in this step, we extracted data from those attributes that we thought would be necessary for our work. In this paper, we focus on the classification of disease using the clinical text data, so we used the clinical notes that were present in the data set for the further classification process.

In this step, we extracted the clinical notes of all the patients that were available in the data set. Then we labelled all the clinical text reports to their corresponding classes. In our data set of the clinical text data, we used four classes, in which we classified the data. The four classes used are COVID, ARDS, SARS, and Both (COVID, ARDS).

3.3 Preprocessing the text extracted

After we were done with extracting the clinical text data from the data set, then the next step was to do the preprocessing of the text data. The text data in the clinical text was not suitable for use in our machine learning model directly. To make our text data in a form that was easy to analyse and predict its class it was necessary to clean the data and to change its structure so that it could be used in a machine learning model.

We used various methods to preprocess our data. Steps of pre-processing the clinical text data are as follows:

(1) The first step was to remove the statements in the clinical text data that were of no use in finding the disease of the patient.

(2) Then tokenization of the sentences in clinical text data was done to split the sentences in the clinical text data into words because the meaning of sentences could be easily understood by analysing the words present in the text data. So it was necessary to split text data into words.

(3) The next step of preprocessing was to remove the stop words in the clinical text data. Stop words are words such as: is, am, a, the, are, and, or etc. which do not contribute to the meaning of the sentence. It is safe to remove such words as removing them does not change the meaning of the sentences, and removing such words reduces the data that needs to be processed.

(4) After removing stop words, we did lemmatization of the words in the clinical text data. Lemmatization is needed to reduce the inflected words in clinical text data. Lemmatization is done to find the root word, which is called the lemma of the inflected word. For example, for words like run, runs and running their lemma is run so all such words in the text whose lemma is run will be reduced to run. We also considered stemming in this step of preprocessing of the clinical text data but after analyzing the results, we preferred lemmatization as it was giving a slight improvement in the result because of the property that lemmatization returns the actual words of the language.

3.4 Feature engineering

The next step after preprocessing and cleaning the clinical text data was feature extraction from the clinical text data. The machine learning model that we had to use to classify the clinical text data and predict the disease of patients could not understand the text data. So we could not work on text data directly. We had to use some feature extraction techniques so that we could convert the text data into a numerical form that could be processed by our machine learning algorithm.

To extract features from the clinical text data, we used different feature engineering techniques. We used the bag of words model to find features in the clinical text data. We created a vocabulary of all unique words from the clinical text data. We also considered bigrams for creating vocabulary. Then we created a matrix of features by keeping each word in a separate column, and each row corresponds to a record of the clinical report. We also used the TF-IDF model on bigrams as a vocabulary to create a matrix of features. The feature matrix created in this step was given as input to the machine learning model created in the next step to classify that the clinical record of a patient belongs to which disease class.

3.5 Classification using Artificial Neural Network model

In this step, we used a machine learning model to predict the class of the clinical text data. In this step, we used the vector created in the previous step as input to a neural network model. The neural network classified the input into one of the four classes as coronavirus, SARS, ARDS, or both coronavirus and ARDS.

An artificial neural network is based on the functioning of neural networks in the human brain. In an artificial neural network, we have a collection of multiple layers of units called neurons. Each layer takes input from the previous layer of neurons and gives its output as input to the layer of neurons that are after it in the whole arrangement of the neural network. Each connection between two neurons which is called an edge has a weight attached to it. This weight increases or decreases during the learning process. If the weight of an edge is increased, then it increases the signal strength in that edge. Neurons have a threshold so that if the signals are stronger than a threshold, then only the signal is passed to the subsequent neuron.

Let’s suppose that there are n inputs x1,x2,..., xn given to the input layer of an artificial neural network. The input xi is then passed to all the neutrons available in the hidden layer. When it is passed to hidden layers, it gets multiplied to some weight wi. More the weight for a particular input the more will be its influence in the neural network and the prediction done by the neural network. Inside the neuron, the weighted input gets summed together.

$$X \cdot W = x_1 w_1 + x_2 w_2 + .. + x_n w_n \qquad (1)$$

Then a bias value is added to this sum.

$$Z = b + \sum_{i=1}^{n} (x_i \ \cdot w_i) \qquad (2)$$

This Z value is then passed to the non-linear activation function. These are used so as to increase a non-linearity in the output of a neuron. An activation function that is normally used is the sigmoid function. This sigmoid function takes the Z value above and gives a number between 0 and 1. The equation of the sigmoid function is as follows :

$$\gamma = \sigma(Z) = \frac{1}{1 + e^{-z}} \qquad (3)$$

After this value is calculated then a loss function is used to find out how better is the output that we get from our neural network. For a classification problem, the loss function that is generally used is the cross-entropy function. For multi-class classification problems, it is given by :

$$C = -\sum_{i=1}^{M} t_{i,o} log\left(p_{i,o}\right) \qquad (4)$$

Where ti = 1 if class label i is correct for observation o, p is the probability that observation o is of class i and M is the number of classes.

After the loss is calculated the backward propagation method and optimisation are used to modify the value of weights used in the neural network so that we can get a better and accurate model. Here we used a three-layer architecture with the input layer having 150 neurons and Hidden layer having 120 neurons and the output layer having four neurons.

4. Results and discussions

In the clinical textual data, we used all the standard text cleaning algorithms for cleaning our text, such as removing the punctuation marks, removing stop words, and using the inbuilt scikit libraries in python. Then by carefully applying various computations such as measuring the frequency of all the words we picked around 150 most relevant features from our dataset which we provided as input to our neural network model. We also split 80% of the data into a training set and 20% of the data into the testing set to further check our results. We have accumulated data of 892 patients from various sources along with the clinical text data and divided it into four classes. After feature engineering of this data, we used artificial neural networks for classification. The first layer is the input layer, which contains 150 neurons, and the output layer contains four neurons, so the value of the number of neurons in the hidden layer should be between these values [18]. The activation function used for the hidden layer is Sigmoid, and Softmax is used as the activation function in the output layer, which gives the probability of each class, and then we took the maximum out of those four. We used 120 neurons in the hidden layer as it gave the best results. During training, with the help of back-propagation, we try to adjust the weights such that the model can predict the results correctly. We tried various combinations of hidden layers and the different number of iterations as we also have to be careful not to overfit the data. So, whenever our model performed very well with the training data but did poorly with the validation data, we discarded those results. We also used grid search for hyperparameter tuning to obtain the best results. Further, we used the concept of Early Stopping and model checkpoint to find the best model. We considered two types of validation for the validation of our model: A test set of 178 clinical text data chosen randomly and that wasn't used in training before and 5-fold cross-validation. Table 1 shows that the model achieved the best accuracy of 97.28% on the testing set using 120 neurons in the hidden layer and a decent accuracy with varying numbers of neurons in the hidden layer. The Artificial neural network Architecture of the model can be seen in "Fig. 6". "Fig. 4" shows the graph for accuracy as a function of the number of epochs, and it shows that the model performs well on both the training and the testing data. Since it's a text classification model, we could have got a better result if we had received more clinical text data regarding these deadly diseases. For our research, we used a system with 16GB RAM, 2.3 GHz processor, and Windows operating system, and Python for implementation.

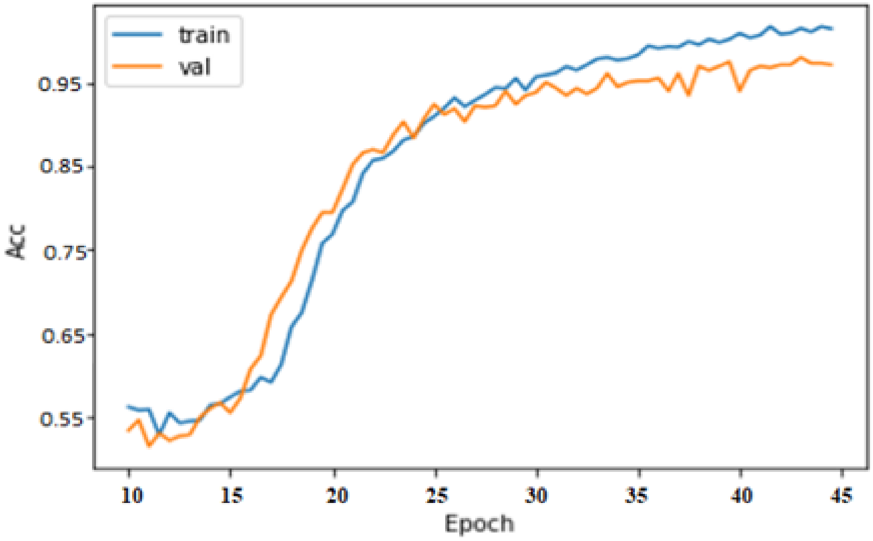

Figure 4. Accuracy vs Number of Epochs.

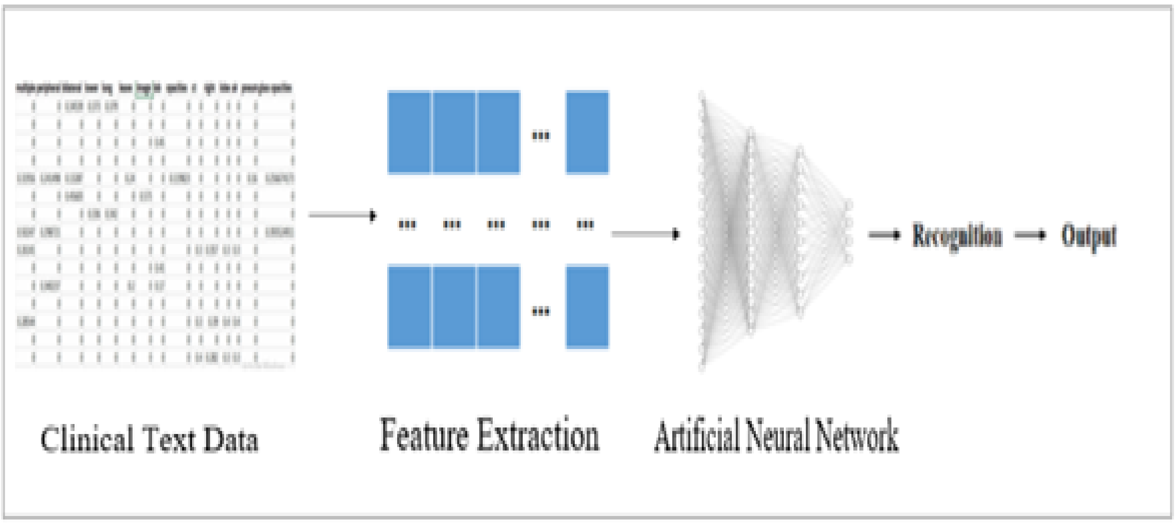

Figure 5. Flow of data

Table 1. Neural Network Performance with number of hidden neurons in hidden layer

Number of hidden Neurons	Iterations	Training Time	Performance	Gradient	Results	
					Cross Entropy Error	Percentages of Error(%E)
20	59	0:00:01	0.0186	0.0245	0.4985	10.3385
30	57	0:00:01	0.00433	0.0150	0.2016	6.6511
40	51	0:00:01	0.00327	0.0133	0.1679	5.338
50	53	0:00:01	0.00244	0.00869	0.4530	4.7614
60	55	0:00:01	0.00368	0.00940	0.1585	4.3567
70	53	0:00:01	0.00632	0.0313	0.1966	3.8792
80	49	0:00:01	0.00353	0.0277	0.1758	3.4552
90	47	0:00:01	0.00311	0.0195	0.1499	3.3667
100	45	0:00:01	0.00203	0.00840	0.2570	3.1254
110	41	0:00:01	0.00257	0.00842	0.2175	2.9467
120	43	0:00:01	0.00201	0.00593	0.1957	2.7277
130	46	0:00:02	0.00187	0.00381	0.1825	3.0012

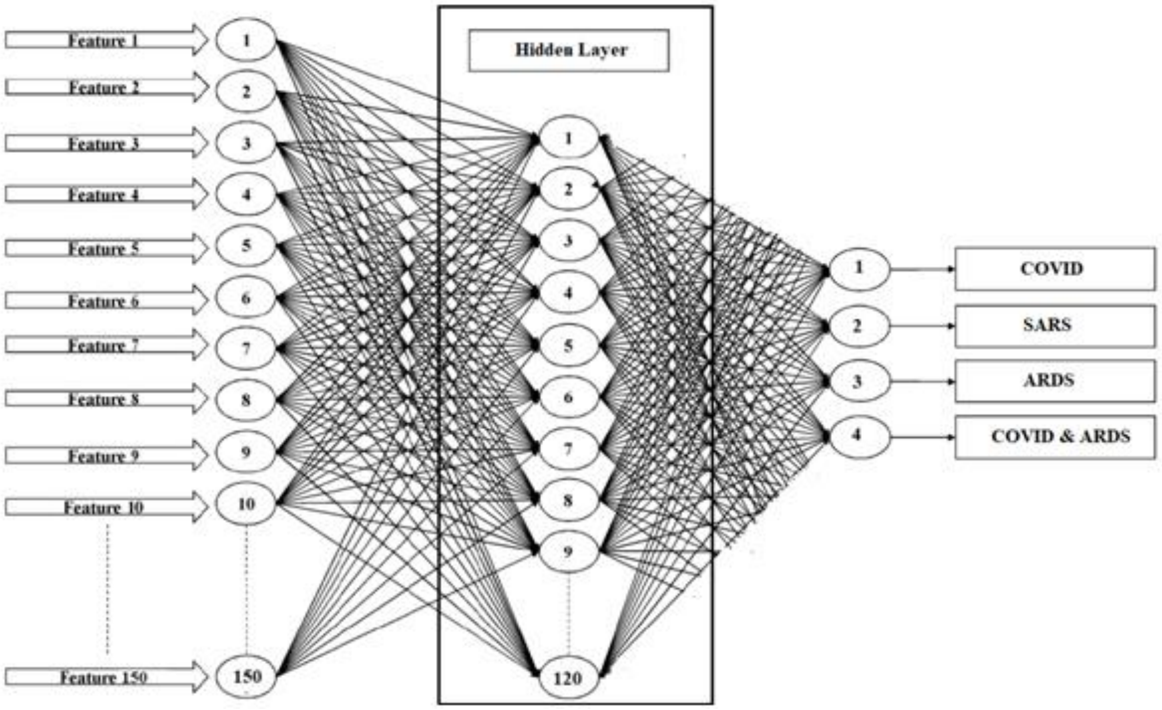

Figure 6. ANN architecture

5. Conclusion

Coronavirus has stunned all the population across the globe. Different specialists and researchers are working to overcome this lethal virus. As the cases are expanding quickly, there is a critical need for a computerized approach to distinguish the coronavirus. Thus, deep learning gives this genuinely necessary automated tool with accurate results. We propose a novel 3-layered artificial neural network architecture to classify clinical text into four classes: COVID, SARS, ARDS, Both(COVID & ARDS) by analyzing 892 clinical reports. Using these clinical reports, we trained our model. With the help of feature engineering, we extracted 150 critical features, and then, using artificial neural networks, we were able to achieve an accuracy of 97.28%. We can improve the efficiency of the model by training it on a bigger dataset. Furthermore, by doing gender-based classification, we can get information about whether males are affected more or females, which can also improve accuracy.

References

1. “COVID-19 Pandemic.” Wikipedia, Wikimedia Foundation, 1 Feb. 2021, en.wikipedia.org/wiki/COVID-19 pandemic.
2. Coronaviruses and Acute Respiratory Syndromes (COVID-19, et al. “Coronaviruses and Acute Respiratory Syndromes (COVID-19, MERS, and SARS) - Infectious Diseases.” MSD Manual Professional Edition, MSD Manuals, www.msdmanuals.com/professional/infectious-diseases/respiratoryviruses/
3. coronaviruses-and-acute-respiratory-syndromes-covid-19-mers-and-sars.
4. P, Suresh, et al. “Deep Learning Applications and Perspectives COVID 19.” Innovations in Information and Communication Technology Series, 2020, pp. 487–490., doi:10.46532/978-81-950008-1-4 106.
5. Qing, Li, et al. “A Novel Neural Network-Based Method for Medical Text Classification.” Future Internet, vol. 11, no. 12, 2019, p. 255., doi:10.3390/fi11120255.
6. ”Jiang, Xiangao, et al. ”Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity.” CMC: Computers,
7. Materials & Continua 63 (2020): 537-51.”

8. ”Iwendi, Celestine, et al. ”Covid-19 patient health prediction using boosted random forest algorithm.” Frontiers in public health 8 (2020): 357.”
9. “Lo´pez-U´ beda, Pilar, et al. ”COVID-19 detection in radiological text reports integrating entity recognition.” Computers in Biology and Medicine 127
10. (2020): 104066.”
11. Wu, Fan, et al. ”A new coronavirus associated with human respiratory disease in China.” Nature 579.7798 (2020): 265-269.
12. Gallegos, A. ”WHO declares public health emergency for novel coronavirus. Medscape Medical News.” (2020).
13. Chen, Nanshan, et al. ”Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.” The Lancet 395.10223 (2020): 507-513.
14. ”World health organization: https://www.who.int/new-room/g-adetail/q-a corronaviruses#:/text=symptoms.” ”Wikipedia coronavirus Pandemic data: https://en.m.wikipedia.org/wiki/Template:COVID-19 pandemic data.”
15. Kumar, Akshi, Vikrant Dabas, and Parul Hooda. ”Text classification algorithms for mining unstructured data: a SWOT analysis.” International Journal of Information Technology (2018): 1-11.
16. Chakraborti, Satarupa, et al. ”A machine learning-based method to detect epilepsy.” International Journal of Information Technology 10.3 (2018): 257-263.
17. Pearlmutter, Barak A. Learning state space trajectories in recurrent neural networks: A preliminary report. CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY PROJECT, 1988.
18. Robinson, Anthony J. ”An application of recurrent nets to phone probability estimation.” IEEE transactions on Neural Networks 5.2 (1994): 298-305.
19. Sundermeyer, Martin, Ralf Schl¨uter, and Hermann Ney. ”LSTM neural networks for language modeling.” Thirteenth annual conference of the international speech communication association. 2012.
20. Shin-ike, K. A two phase method for determining the number of neurons in the hidden layer of a 3-layer neural network. In Proceedings of the SICE Annual Conference 2010, Taipei, Taiwan, 18–21 August 2010; pp. 238–242.