

# **Covid 19 Vaccine Analysis Project**

**BATCH MEMBERS:**

**M.POTHINI**

**S.SHANOFAR BEGUM**

**P.SUBAMOZHI**

**R.VIJAYASHREE**


Technology advancements have a rapid effect on every field of life, be it medical field or any other field. Artificial intelligence has shown the promising results in health care through its decision making by analysing the data. COVID-19 has affected more than 100 countries in a matter of no time. People all over the world are vulnerable to its consequences in future. It is imperative to develop a control system that will detect the coronavirus. One of the solution to control the current havoc can be the diagnosis of disease with the help of various AI tools. In this paper, we classified textual clinical reports into four classes by using classical and ensemble machine learning algorithms. Feature engineering was performed using techniques like Term frequency/inverse document frequency (TF/IDF), Bag of words (BOW) and report length. These features were supplied to traditional and ensemble machine learning classifiers. Logistic regression and Multinomial Naïve Bayes showed better results than other ML algorithms by having 96.2% testing accuracy. In future recurrent neural network can be used for better accuracy.

**Keywords:** Artificial intelligence, COVID-19, Imperative, Machine learning, Ensemble

## Introduction

---

In December 2019, the novel coronavirus appeared in the Wuhan city of China [1] and was reported to the World Health Organization (W.H.O) on 31st December 2019. The virus created a global threat and was named as COVID-19 by W.H.O on 11th February 2020 [1]. The COVID-19 is the family of viruses including SARS, ARDS. W.H.O declared this outbreak as a public health emergency [2] and mentioned the following; the virus is being transmitted via the respiratory tract when a healthy person comes in contact with the infected person. The virus may transmit between persons through other roots which are currently unclear. The infected person shows symptoms within 2–14 days, depending on the incubation period of the middle east respiratory syndrome (MERS) and the severe acute respiratory syndrome (SARS). According to W.H.O the signs and symptoms of mild to moderate cases are dry cough, fatigue

 Feedback

and fever while as in severe cases dyspnea (shortness of breath), Fever and tiredness may occur [3, 4]. The persons having other diseases like asthma, diabetes, and heart disease are more vulnerable to the virus and may become severely ill. The person is diagnoses based on symptoms and his travel history. Vital signs are being observed keenly of the client having symptoms. No specific treatment has been discovered as on 10th April 2020, and patients are being treated symptomatically. The drugs like hydroxychloriquine, antipyretic, anti-virals are used for the symptomatic treatment. Currently, no such vaccine is developed for preventing this deadly disease, and we may take some precautions to prevent this disease. By washing hands regularly with soap for 20 s and avoiding close contact with others by keeping the distance of about 1 m may reduce the chances of getting affected by this virus. While sneezing, Covering the mouth and nose with the help of disposable tissue and avoiding the contact with the nose, ear and mouth can help in its prevention. SARS is an airborne disease that appeared in 2003 in China and affected 26 countries by having 8 K cases in the same year and transferred from person to person. The signs and symptoms of SARS are fever, cold, diarrhoea, shivering, malaise, myalgia and dyspnea. The ARDS (acute respiratory distress syndrome) is characterized by rapid onset of inflammation in lungs which leads to respiratory failure and its signs and symptoms are bluish skin colour, fatigue and shortness of breath. ARDS is diagnosed by  $\text{PaO}_2/\text{FiO}_2$  ratio of less than 300 mm Hg. Till 10th of April 2020, almost 1.6 million confirmed cases of coronavirus are detected around the globe. Almost 97 K persons have died and 364 K persons have recovered from this deadly virus [5]. Figure 1 shows the worldwide data regarding coronavirus. Since no drug or vaccine is made for curing the COVID-19. Various paramedical companies have claimed of developing a vaccine for this virus. Less testing has also given rise to this disease as we lack the medical resources due to pandemic. Since thousands and thousands are being tested positive day by day around the globe, it is not possible to test all the persons who show symptoms.



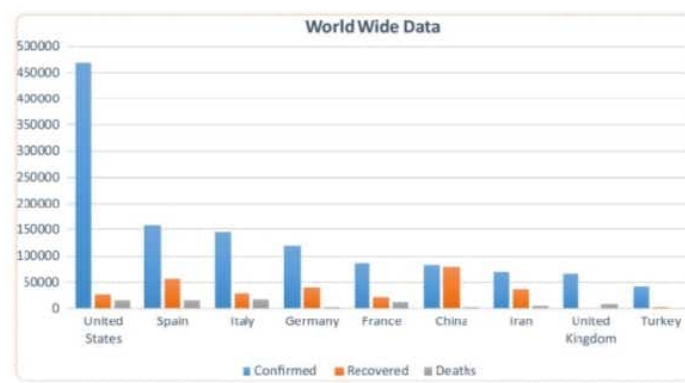


Fig. 1

Worldwide coronavirus as of 10th April 2020

Apart from clinical procedures, machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. From the past decade, a huge amount of progress is being made in this area for resolving some critical projects. Recent pandemic has attracted many researchers around the globe to solve this problem. Data provided

by John Hopkins University in the form of X-ray images and various researchers build a model of machine learning that classifies X-ray image into COVID-19 or not. Since the latest data published by Johns Hopkins gives the metadata of these images. The data consists of clinical reports in the form of text in this paper, we are classifying that text into four different categories of diseases such that it can help in detecting coronavirus from earlier clinical symptoms. We used supervised machine learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS). We are also using ensemble learning techniques for classification. Section 2 gives the literature survey regarding the proposed work. The framework for detecting coronavirus from clinical text data is being discussed in Sects. 3 and 4 gives the experimental results of the proposed framework and Sect. 5 concludes our work.

## Related work

Machine learning and natural language processing use big data-based models for pattern recognition, explanation, and prediction. NLP has gained much interest in recent years, mostly in the field of text analytics. Classification is one of the major task in text mining and can be performed using different algorithms [6]. Kumar et al. [7] performed a SWOT analysis of various supervised and unsupervised text classification algorithms for mining the unstructured data. The various applications of text classification are sentiment analysis, fraud detection, and spam detection etc. Opinion mining is majorly being used for elections, advertisement, business etc. Verma et al. [8] analysed Sentiments of Indian government projects with the help of the lexicon-based dictionary. The machine learning has changed the perspective of diagnosis by giving great results to diseases like diabetes and epilepsy. Chakraborti et al. [9] detected epilepsy using machine learning approaches, electroencephalogram (EEG) signals are used for detecting normal and epileptic conditions using artificial neural networks (ANN). Sarwar et al. [10] diagnosis diabetes using machine learning and ensemble learning techniques result indicated that ensemble technique assured accuracy of 98.60%. These purposes can be beneficial to diagnose and predict COVID-19. Firm and exact diagnosis of COVID-19 can save millions of lives and can produce a massive amount of data on which a machine learning (ML) models can be trained. ML may provide useful input in this regard, in particular in making diagnoses based on clinical text, radiography Images etc. According to Bullock et al. [11], Machine learning and deep learning can replace humans by giving an accurate diagnosis. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19. X-rays and computed tomography (CT) scans can be used for training the machine learning model. Several initiatives are underway in this regard. Wang and Wong [12] devel-

oped COVID-Net, which is a deep convolutional neural network, which can diagnose COVID-19 from chest radiography images. Once the COVID-19 is detected in a person, the question is whether and how intensively that person will be affected. Not all COVID-19 positive patients will need rigorous attention. Being able to prognosis who will be affected more severely can help in directing assistance and planning medical resource allocation and utilization. Yan et al. [13] used machine learning to develop a prognostic prediction algorithm to predict the mortality risk of a person that has been infected, using data from (only) 29 patients at Tongji Hospital in Wuhan, China. Jiang et al. [14] proposed a machine learning model that can predict a person affected with COVID-19 and has the possibility to develop acute respiratory distress syndrome (ARDS). The proposed model resulted in 80% of accuracy. The samples of 53 patients were used for training their model and are restricted to two Chinese hospitals. ML can be used to diagnose COVID-19 which needs a lot of research effort but is not yet widely operational. Since less work is being done on diagnosis and predicting using text, we used machine learning and ensemble learning models to classify the clinical reports into four categories of viruses.

## Methodology

The proposed methodology consists of 2.1 to 2.5 steps. In step 2.1 data collection is being performed and 2.2 defines the refining of data, 2.3 gives an overview of preprocessing, 2.4 provides a mechanism for feature extraction. In E traditional machine learning algorithms are discussed, and 2.5 gives an overview of ensemble machine learning algorithms. The visual representation of the proposed methodology is shown in Fig. 2. and are being discussed below.

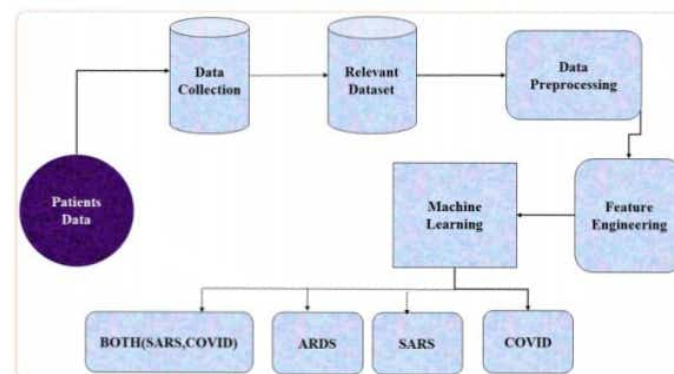


Fig. 2

Methodology

## Data collection

As W.H.O declared Coronavirus pandemic as Health Emergency. The researchers and hospitals give open access to the data regarding this pandemic. We have collected from an open-source data repository GitHub.<sup>1</sup> In which about 212 patients data is stored which have shown symptoms of corona virus and other viruses. Data consists of about 24 attributes namely patient id, offset, sex, age, finding, survival, intubated, went\_icu, needed\_supplemental\_O<sub>2</sub>, extubated, temperature, pO<sub>2</sub> saturation, leukocyte\_count, neutrophil count, lymphocyte count, view, modality, date, location, folder, filename, DOI, URL. License. Clinical notes and other notes.

## Relevant dataset

Since our work is regarding text mining so we extracted clinical notes and findings. Clinical notes consist of text while as the attribute finding consist label of the corresponding text. About 212 reports were used and their length was calculated. We consider only those reports that are written in the English language. Figure 3 gives the length distribution of clinical reports that are written in English. The clinical reports are labelled to their corresponding classes. In our dataset, we have four classes COVID, ARDS

reports are labelled to their corresponding classes. In our dataset, we have four classes COVID, ARDS, SARS and Both (COVID, ARDS). Figure 4 shows the different classes in which clinical text is being categorized and corresponding report length.

---

[Fig. 3](#)

Clinical report length

[Fig. 4](#)

Different classed with their report length

## Preprocessing

The text is unstructured so it needed to be refined such that machine learning can be done. Various steps are being followed in this phase; the text is being cleaned by removing unnecessary text. Punctuation and lemmatisation are being done such that the data is refined in a better way. Stopwords, symbols, Url's, links are removed such that classification can be achieved with better accuracy. Figure 5 shows the main steps in preprocessing.

[Fig. 5](#)

Preprocessed data set

## Feature engineering

From the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values. We use TF/IDF technique for extracting relevant features. Bag of words was also taken into consideration, unigrams, bigrams were also extracted. We identified 40 relevant features by which the classification can be achieved. These features are shown in Fig. 6. By giving the corresponding weight to the feature and the same input is being supplied to machine learning algorithms.



The classification is performed to classify the given text into four different types of viruses. The four classes of viruses, COVID ( a person having coronavirus), ARDS, SARS and both (consists a person that is having both corona virus as well as ARDS). Various supervised machine learning algorithms are being used to classify the text into these categories. The machine learning algorithms like support vector machine (SVM), multinomial Naïve Bayes (MNB), logistic regression, decision tree, random forest, bagging, Adaboost and stochastic gradient boosting were used for performing this task.

#### Traditional machine learning algorithms

*Logistic regression* This algorithm predicts the class of numerical variable based on its relationship with the label [15]. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig. 6. The algorithm generally calculates the class membership probability. Here we have four classes.  $y \in \{0, \dots, 3\}$ . The posterior probabilities can be calculated with the help of Eq. 1.

$$\begin{aligned} P(y = k|x) &= \frac{\exp^{\varphi^T \theta_k}}{1 + \sum_{k=1}^3 \exp^{\varphi^T \theta_k}} \forall k = 1, \dots, 3 \\ P(y = 0|x) &= \frac{\exp^{\varphi^T \theta_0}}{1 + \sum_{k=1}^3 \exp^{\varphi^T \theta_k}} \end{aligned} \quad 1$$

*Multinomial Naïve Bayes* MNB computes class probabilities of a given text by using Bayes rule [16]. Let C denote the set of classes in our problem we have four classes  $C = 0, 1, 2$  and  $3$ . Moreover, N is the set of features here we have  $N = 40$  (40 features are taken using TF/IDF) as shown in Fig. 6. Then MNB assigns test text  $t_i$  to the class that has the highest probability  $P(c|t_i)$  by using Bayes rule shown in Eq. 2:

$$P(c|t_i) = \frac{P(c)P(t_i|c)}{P(t_i)}, \quad c \in C \quad 2$$

$P(c)$  can be calculated by dividing the number of clinical textual data that is labelled as class  $c$  to the total number of clinical textual data.  $P(t_i|c)$  is the probability of obtaining a clinical text report like  $t_i$  in-class  $c$  and is calculated as:

$$\begin{aligned} P(t_i|c) &= \left( \sum_n f_{ni} \right)! \\ &\prod_n \frac{P(w_n|c)^{f_{ni}}}{f_{ni}!} \end{aligned}$$

where  $f_{ni}$  is the count of word/term 'n' in our clinical text report  $t_i$  and  $P(w_n|c)$  is the probability of word/term 'n' given in class  $c$ . From the training data, the latter probability is calculated by:

$$P(w_n|c) = \frac{1 + F_{nc}}{N + \sum_{x=1}^N F_{xc}}$$

where  $F_{xc}$  is the count of word/term 'x' in all the clinical training reports belonging to the class c. for avoiding zero-frequency problem Laplace estimator is used which assigns value one to each word's count.

*Support vector machine (SVM)* Support vector machine (SVM) is a supervised machine learning algorithm for classifying text into different categories [17]. It takes 'n' number of features for the particular text with the given label. Here we have taken 40 features that are of nature unigram and bigram as the dataset is small. Here the data points of the training set are  $(y_k, x_k)_1^n$ , where n is the number of features taken. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig. 6. The main aim of SVM is to construct a classifier in the form of Eq. 3.

$$y(x) = \text{sign} \left[ \sum_{k=1}^n \alpha_k y_k \psi(x, x_k) + b \right] \quad 3$$

where  $\alpha_k$  = positive real constant. b = real constant.

$$\psi(x, x_k) = \begin{cases} x_k^T x : \text{Linear SVM} \\ (x_k^T, x, +, 1)^d : \text{Polynomial SVM with Degree } d \\ \exp(-||x - x_k||_2^2 / \sigma^2) : \text{RBF SVM} \end{cases}$$

where k,  $\sigma$  are constants.

The classifier is built by assuming the following:

$$\omega^T \varphi(x_k) + b \geq 1, \text{ if } y_k = +1$$

$$\omega^T \varphi(x_k) + b \leq -1, \text{ if } y_k = -1$$

Which is equivalent to Eq. 4:

$$y_k [\omega^T \varphi(x_k) + b] \leq -1, \text{ if } y_k = -1, k = 1, \dots, n \quad 4$$

where  $\varphi(\cdot)$  = nonlinear function that maps input spaces into more significant dimensional space.

The hyperplane is formed by which classification is performed. The hyperplane differentiates the four classes (COVID, ARDS, SARS and Both) for this a new variable  $\xi_k$  is introduced. Equation 5 is the equation for hyperplane:

$$\begin{aligned}
y_k [\omega^T \varphi(x_k) + b] &\geq 1 - \xi_k, \\
k &= 1, \dots, n \\
\xi_k &\geq 0, \quad k = 1, \dots, n
\end{aligned}$$

5

*Decision trees* An alternative approach for classification it partitions the input space into regions and classifies every region independently [18]. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig. 6. It splits the space recursively according to the inputs and classifies at the bottom of the tree. The leaf nodes classify the text into four classes. While building a decision tree, a vital function needs to be considered which is known as the splitting criterion. The function defines how data should be split in order to maximise performance. We have used information gain ratio which is equal to information gain to the intrinsic information and is shown in Eq. 6.

$$IGR(EX, a) = IG / IV$$

6

where IG = information gain. IV = intrinsic information.

Information gain is calculated with the help of entropy as shown below:

$$\begin{aligned}
IG(E, x, a) &= H(E, x) - \\
&\sum_{v \in \text{values}(a)} \left( \frac{|\{x \in Ex | \text{value}(x, a) = v\}|}{|E, x|} \right) \\
&\cdot H(\{x \in Ex | \text{value}(x, a) = v\})
\end{aligned}$$

where Ex = set of training examples and  $x \in Ex$  which defines the value of a specific example x for feature a. H = entropy and a = features.

Intrinsic information value is calculated by:

$$\begin{aligned}
IV(E, x, a) &= - \\
&\sum_{v \in \text{values}(a)} \frac{|\{x \in Ex | \text{value}(x, a) = v\}|}{|E, x|} \\
&\cdot \log_2 \left( \frac{|\{x \in Ex | \text{value}(x, a) = v\}|}{|Ex|} \right)
\end{aligned}$$

Ensemble machine learning techniques

*Bagging* An ensemble machine learning algorithm which improves the performance of other classification and regression machine learning algorithms [19]. Bagging algorithm helps in avoiding overfitting. Let a training set X of size 'n' is given, by sampling uniformly 'm' new training sets Xi are generated each having size 'n' with replacements. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig. 6. Due



to replacements, some observations could repeat in each  $X_i$ . If  $m' = n$  then set  $X_i$  for large  $n$  is expected to have a fraction  $(1 - 1/e)$  of the unique examples of  $X$ , the rest being duplicates. This sample is known as a bootstrap sample. The  $m$  models are fitted using  $m$  bootstrap samples and combined by voting.

**AdaBoost** This ensemble learning algorithm works with those instances of the dataset, which are weighted [20]. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig 6. It starts with having equal weights for each observation and trains a weak learning algorithm using weighted data. By performing this, a weak classifier is being produced. Depending on the performance of this weak learning classifier choose a coefficient  $\alpha$ . On misclassified points improve weights and decrease weights of the correctly classified points. Then the weak learning algorithms are again executed to get a weak classifier for the new weighted data. Repeating this procedure leads to develop an AdaBoost classifier.

---

**Random forest classifier** Ensemble machine learning algorithm used for classification and works like a decision tree. The bootstrap aggregating technique is used for training the random forest algorithm. By averaging predictions of all the individual regression trees, the overall prediction can be made. In case for classification trees, the majority vote is taken. This algorithm uses a modified tree learning algorithm which selects and split each learning process by a subset of random features [21]. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig. 6. With the help of various decision trees, the algorithm creates a forest from a subset of randomly selected data and sums up the votes for the decision trees to decide the final class of the object.

**Stochastic gradient boosting** This algorithm allows trees to be greedily created from samples of the training dataset. The 40 features that have been selected in feature engineering with values are represented in the form of a table and are supplied as an input, as shown in Fig 6. This is used for reducing the correlation between the trees in gradient boosting. At each iteration, a subsample of the training data is drawn at random without replacements from the full training dataset. The randomly selected subsample is then used instead of the full sample to fit the base learner [22].

## Results and discussion

---

We used a windows system with 4 GB Ram and 2.3 GHz processors for performing this work. Scikit learn tool is being used for performing machine learning classification with the help of various libraries like NLTK, STOPWORDS etc. for improving the accuracy of all the machine learning algorithms pipeline is being used. After performing the statistical computation, deeper insights about the data were achieved. The data is being split into 70:30 ratio where 70% data is being used for training the model and 30% is used for testing the model. We have clinical text reports of 212 patients that are labelled into four classes. The classification was done using machine learning algorithms by supplying them features that were extracted in the feature engineering step. In order to explore the generalization of our model from training data to unseen data and reduce the possibility of overfitting, we split our initial dataset into separate training and test subsets. The tenfold cross-validation strategy was conducted for all algorithms, and this process was repeated five times independently to avoid the sampling bias introduced by randomly partitioning the dataset in the cross-validation. Table 1 gives a comparative analysis of all the classical machine learning methods that are used for performing this task. Table 2 gives a comparative analysis of all the classical machine learning and Ensemble learning methods that are used for performing the task of classifying the clinical text into four classes. The results showed that logistic regression and Multinomial Naïve Bayes Algorithm shows better result than all other algorithms by having precision 94%, recall 96%, F1 score 95% and accuracy 96.2% other algorithms like random forest, gradient boosting also showed good results by having accuracy 94.3% respectively. The visualized comparative analysis of all the algorithms that are used in our work is shown in Fig. 7. Since we all know, the COVID-19 data is least available. To get the real accuracy of the model we experimented it in two stages. In the first stage, we took 75% of the available data and it shows less accuracy as compared to the stage in which whole data was used for experimentation. So we can conclude that if more data is supplied to these algorithms, there are chances of improvement in performance. As we are facing a se-

Table 1

Comparative analysis of traditional machine learning algorithms

Algorithm	Precision	Recall	F1 score	Accuracy (%)
Logistic regression	0.94	0.96	0.95	96.2
Multinomial Naïve Bayesian	0.94	0.96	0.95	96.2
Support vector machine	0.82	0.91	0.86	90.6
Decision tree	0.92	0.92	0.92	92.5

Table 2

Shows the comparative analysis of classical as well as ensemble machine learning algorithms

Algorithm	Precision	Recall	F1 score	Accuracy (%)
Logistic regression	0.94	0.96	0.95	96.2
Multinomial Naïve Bayesian	0.94	0.96	0.95	96.2
Support vector machine	0.82	0.91	0.86	90.6
Decision tree	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
Random forest	0.93	0.94	0.93	94.3
Stochastic gradient boosting	0.93	0.94	0.93	94.3

Fig. 7

Comparative analysis of machine learning and ensemble learning algorithms

## Conclusion

COVID-19 has shocked the world due to its non-availability of vaccine or drug. Various researchers are working for conquering this deadly virus. We used 212 clinical reports which are labelled in four classes namely COVID, SARS, ARDS and both (COVID, ARDS). Various features like TF/IDF, bag of words are being extracted from these clinical reports. The machine learning algorithms are used for classifying clinical reports into four different classes. After performing classification, it was revealed that logistic regression and multinomial Naïve Bayesian classifier gives excellent results by having 94%

precision, 96% recall, 95% f1 score and accuracy 96.2%. Various other machine learning algorithms that showed better results were random forest, stochastic gradient boosting, decision trees and boosting. The efficiency of models can be improved by increasing the amount of data. Also, the disease can be classified on the gender-based such that we can get information about whether male are affected more or females. More feature engineering is needed for better results and deep learning approach can be used in future.