Machine learning approaches based on clinical text data to detecting COVID-19

Shaik Sameeruddin

VIT-AP University, Amaravati, India

Abstract :

Advances in technology have a rapid effect on every area of life, whether in the medical field or any other field. By analyzing the data, Artificial intelligence has demonstrated promising results in health care through its decision making. In no time had COVID-19 affected more than 100 countries. People across the globe are vulnerable to future consequences. Development of a control system that detects coronavirus is imperative.

Diagnosis of illness with the help of various AI tools can be one of the solutions to control the current havoc. In this paper we used classical and ensemble machine learning algorithms to classify textual clinical reports into four classes. Feature engineering was carried out using techniques such as Term Frequency / Inverse Document Frequency (TF / IDF), Word Bag (BOW), and Report Length. These features were provided to the classification of traditional and ensemble machine learning. By having 96.2 percent test accuracy, logistic regression and Multinomial Naïve Bayes showed better results than other ML algorithms. For greater accuracy, recurrent neural networks can be used in the future.

Introduction

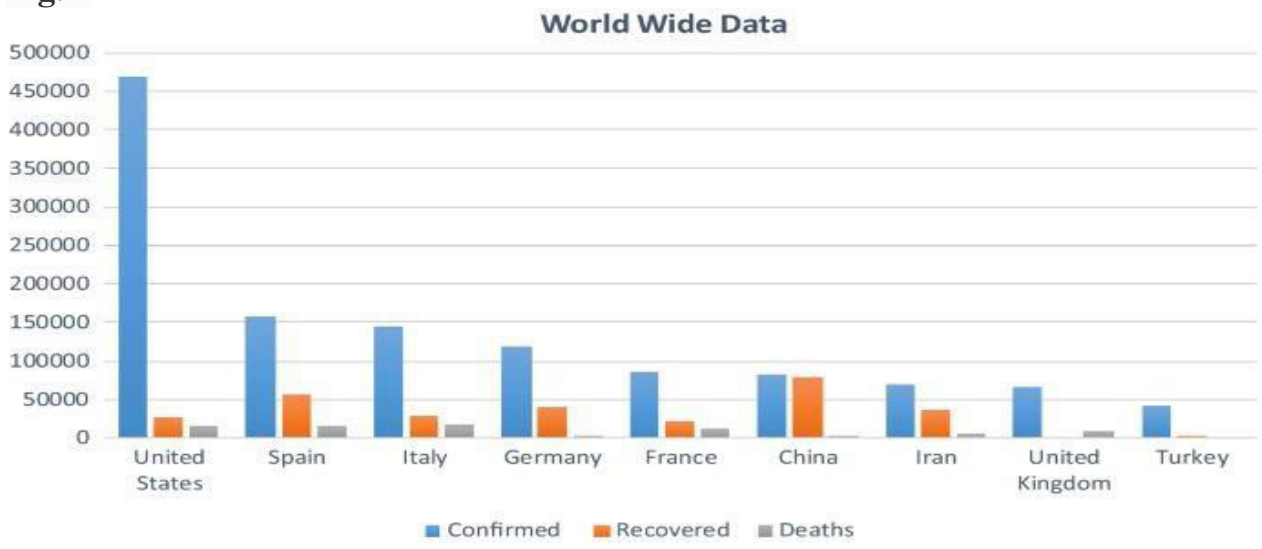
The novel coronavirus appeared in China's Wuhan city[1] in December 2019, and was reported to the World Health Organization (W.H.O) on December 31, 2019. The virus created a global threat and was named on 11th February 2020 by W.H.O as COVID-19[1]. The COVID-19 is a family of viruses including SARS,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

The signs and symptoms of mild to moderate cases, according to W.H.O, are dry cough, fatigue, and fever while, as in severe cases, dyspnea (shortness of breath), fever and fatigue may occur [3, 4]. Individuals with other diseases such as asthma, diabetes, and heart disease are more vulnerable to the virus and can become seriously ill. The person is symptom-based diagnoses and his travel history. Vital signs of a client having symptoms are keenly observed. As of April 10th, 2020, no specific treatment has been discovered and patients are being treated symptomatically. Drugs such as hydroxychloroquine, antipyretic, antivirals are used to treat the symptoms. There is currently no such vaccine being developed to prevent this deadly disease, and we may be taking some precautions to prevent it.

By regularly washing hands with soap for 20 s, and avoiding close contact with others by keeping a distance of about 1 m, this virus may reduce the chances of becoming affected.SARS is an airborne disease that appeared in China in 2003 and affected 26 countries with 8 K cases in the same year and was transferred from one person to another. SARS fever, cold, diarrhea, shivering, malaise, myalgia, and dyspnea are signs and symptoms of. ARDS (acute respiratory distress syndrome) is characterized by a rapid onset of inflammation in the lungs that results in respiratory failure, and its signs and symptoms are bluish skin color, fatigue, and shortness of breath. ARDS is diagnosed with a ratio of PaO2 / FiO2 under 300 mm Hg.Close to 1.6 million confirmed coronavirus cases are detected around the globe by 10 April 2020. Nearly 97 K people died, and 364 K people recovered from this deadly virus[5]. Figure 1 shows coronavirus data from around the world.

Because there is no drug or vaccine to cure COVID-19. Various paramedical companies have claimed of developing a vaccine for this virus. This disease has also been caused by less testing, as we lack the medical resources due to the pandemic. Since thousands and thousands of people around the globe are being tested positive day by day, it is not possible to test all of the people who show symptoms.

Fig. 1



Besides clinical procedures, with the help of image and textual data, machine learning provides much support in identifying the disease. Novel coronavirus can be identified using machine learning. It can also predict the nature of the virus anywhere in the world. Machine learning does, however, require a huge amount of data to classify or predict diseases. In order to classify the text or image into different categories, supervised machine learning algorithms require annotated data. A huge amount of progress has been made in this area since the last decade to resolve some of the critical projects. The recent pandemic has drawn many researchers around the world to solve this issue.

Data provided by John Hopkins University in the form of X-ray images and various researchers build a machine learning model that classifies or does not classify X-ray images into COVID 19. The metadata of these images is given since the latest data published by Johns Hopkins. The data in this paper consists of clinical reports in the form of a text, we classify that text into four different categories of diseases, so that it can help to detect coronavirus from earlier clinical symptoms. We used supervised techniques of machine learning to classify the text into four different categories: COVID, SARS, ARDS, and both (COVID, ARDS); We also use classification techniques for the ensemble learning. Section 2 provides a survey of the literature concerning the proposed work.. Sects discusses the framework for detecting coronavirus from the clinical text data. 3 and 4 provide the experimental results of the Framework and Sect proposed. 5 concludes our work.

Related work

Machine learning and natural language processing use large data-based models to recognize patterns, explain them, and predict. In recent years, NLP has gained considerable interest, mostly in the field of text analytics, classification is one of the main tasks in text mining and can be performed using different algorithms[6]. Kumar et al.[7] conducted a SWOT analysis of the various text classification algorithms supervised and unsupervised for the mining of unstructured data. The different text classification applications are sentiment analysis, fraud detection and spam detection etc.

Opinion mining is used primarily for election, advertising, business etc. Verma et al.[8], using the lexicon-based dictionary, analyzed the Sentiments of Indian government projects. Machine learning has changed the diagnostic perspective by delivering great results for diseases such as diabetes and epilepsy. Chakraborti et al.[9] detected epilepsy using machine learning approaches, electroencephalogram (EEG) signals are used with artificial neural networks (ANN) to detect normal and epileptic conditions.
The resulting diagnosis of diabetes by Sarwar et al.[10] using machine learning and ensemble learning techniques indicated that the ensemble technique provided 98.60% accuracy.
Diagnosing and predicting COVID-19 can be beneficial for those purposes. COVID-19's firm and accurate diagnosis can save millions of lives and can produce massive amounts of data on which to train a machine learning (ML) model. In this regard, ML may provide useful input, particularly in making clinical text-based diagnoses, radiography images etc. Machine learning and deep learning can replace humans by making an accurate diagnosis, according to Bullock et al.[11]. Perfect diagnosis can save time for radiologists and can be cost-effective than standard COVID-19 testing.

The machine learning model can be trained with X-rays and computed tomography (CT) scans.In this regard, there are several initiatives underway. Wang and Wong[12] have developed COVID-Net, a deep, coevolutionary neural network that can diagnose COVID-19 from chest x-ray images. Once the COVID-19 is detected in a person, the question is whether this person will be affected, and how intensively. Not all positive patients with COVID-19 will need rigorous care. Being able to forecast who will be more severely affected can help in directing aid and planning the allocation and utilization of medical resources.

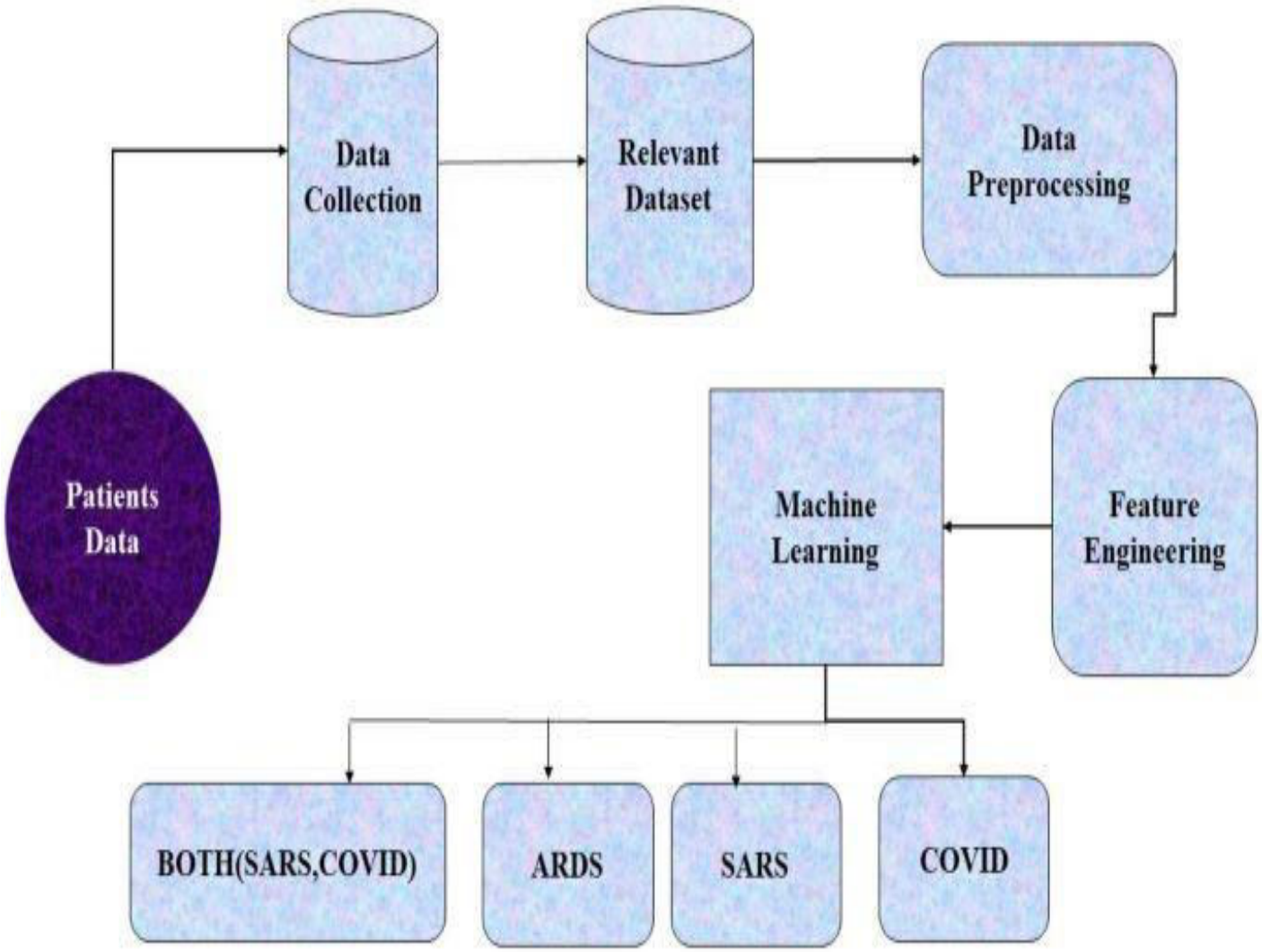
Yan et al.[13] used machine learning to develop a forecasting algorithm for predicting the mortality risk of an infected person; Using data from (only) 29 Tongji Hospital patients in Wuhan, China; Jiang et al.[14] proposed a machine learning model capable of predicting a person affected by COVID-19 and having the potential for developing acute respiratory distress syndrome (ARDS). The model proposed led to 80 percent accuracy. 53 patient samples have been used to train their model and are limited to two Chinese hospitals.

ML can be used to diagnose COVID-19, which requires a great deal of research effort but is still not widely operational. As less work is done on diagnosis and prediction using text, we have used machine learning and ensemble learning models to classify clinical reports into four categories of viruses

Methodology

The methodology proposed is comprised of steps 2.1 to 2.5. Data collection is carried out in steps 2.1 and 2.2 defines data refining, 2.3 provides an overview of pre-processing and 2.4 provides a mechanism for extracting the function. Traditional machine learning algorithms are discussed in E, and 2,5 provides an overview of the ensemble's machine learning algorithms. A visual representation of the methodology proposed is shown in Fig. 2. And these are discussed below.

Fig. 2


Data Collection
Relevant Dataset
Data Preprocessing
Patients Data
Machine Learning
Feature Engineering
BOTH(SARS,COVID)
ARDS
SARS
COVID

Data collection

The coronavirus pandemic was declared as a health emergency by W.H.O. Researchers and hospitals are giving the data on this pandemic an open access. We collected data from an open-source data repository GitHub. Footnote1 In which data about 212 patients are stored showing symptoms of corona and other viruses. Data consists of approximately 24 attributes, namely patient I d, offset, sex, age, discovery, survival, intubation, went ICU, needed supplemental O2, extubation, temperature, pO2 saturation, leukocyte count, neutrophil count, Number of lymphocytes, view, model, date , location, folder, filename, DOI, URL. Clinical notes, and additional notes.

Relevant dataset

Since our work concerns text mining, clinical notes and findings have been extracted. Clinical notes consist of text whilst the finding of the attribute consists of the corresponding text label. Around 212 reports have been used, and their length has been calculated. We only consider those reports which are written in English.

The length distribution of clinical reports written in English is given in Figure 3. The clinical reports are labeled according to their classes. We have four classes in our dataset: COVID, ARDS, SARS, and Both (COVID, ARDS). Figure 4 shows the different classes in which clinical text is categorised, and the report's corresponding length.

Fig. 3

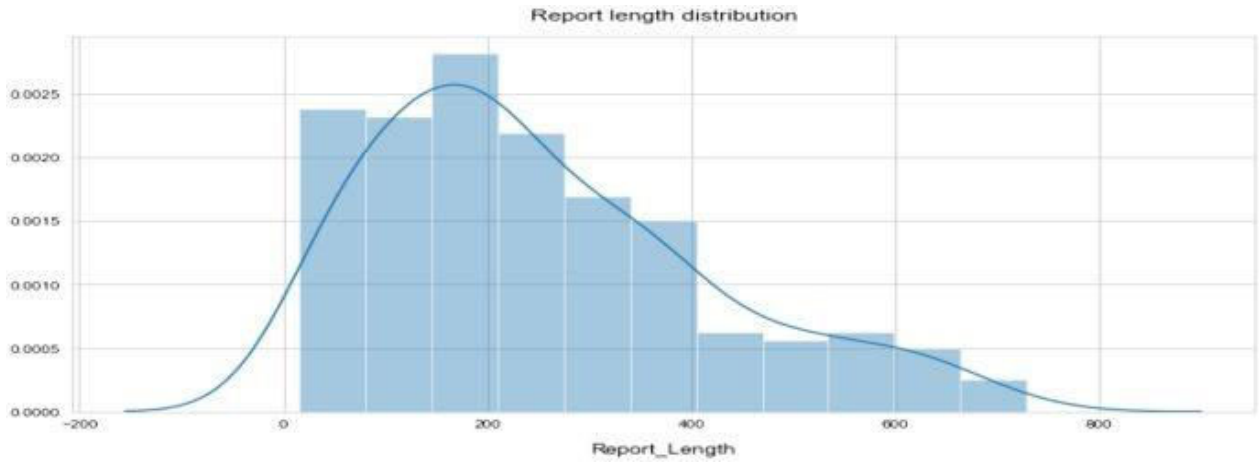
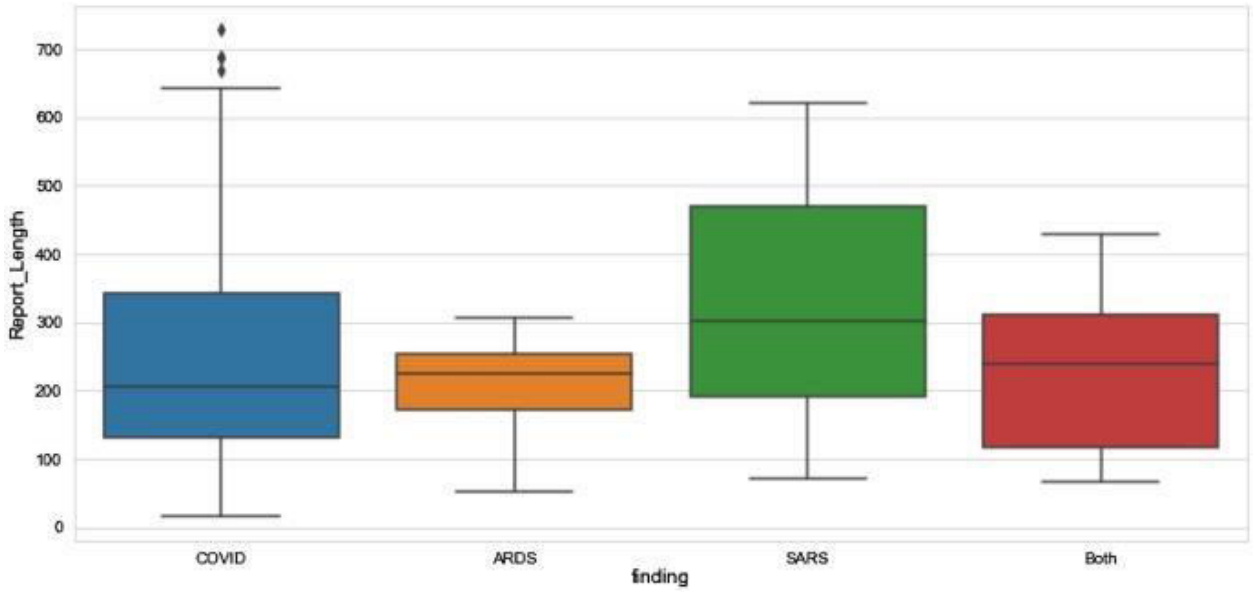


Fig. 4



Preprocessing

The text is unstructured, so it needed to be refined to allow for machine learning. During this phase various steps are followed; the text is cleaned up by removing unnecessary text.

Punctuation and lemmatization are done in such a way that the data is better refined. Stopwords, symbols, Url's, links are removed so classification can be achieved with greater precision. Figure 5 shows key preprocessing steps.

Fig. 5

Clinical_notes	Finding	Report_Length	Punctuation	Lemmatisation	Stop_Word Removal
infiltrate in the upper lobe	COVID	45	infiltrate in the upper	infiltrate in the upper lobe	infiltrate upper lobe leave lung
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
diffuse infiltrates in the bil	COVID	48	diffuse infiltrates in th	diffuse infiltrate in the bila	diffuse infiltrate bilateral lower lungs
progressive diffuse inters	COVID	115	progressive diffuse int	progressive diffuse interst	progressive diffuse interstitial opacities consolidation
Severe ARDS. Person is in	ARDS	53	severe ards person is	severe ards person be intu	severe ards person intubate og place
Case 2: chest x-ray obtain	COVID	563	case 2 chest x-ray obt	case 2 chest x-ray obtain o	case 2 chest x-ray obtain jan 6 (2a) brightness lungs
Case 2: chest x-ray obtain	COVID	563	case 2 chest x-ray obt	case 2 chest x-ray obtain o	case 2 chest x-ray obtain jan 6 (2a) brightness lungs
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 29-year-old wor	SARS	378	sars in a 29-year-old v	sars in a 29-year-old wom	sars 29-year-old woman present 7 days exposure ()
SARS in a 29-year-old wor	SARS	378	sars in a 29-year-old v	sars in a 29-year-old wom	sars 29-year-old woman present 7 days exposure ()
SARS in a 42-year-old wor	SARS	145	sars in a 42-year-old v	sars in a 42-year-old wom	sars 42-year-old woman present 9 days exposure p

Feature engineering

Different features are extracted from the pre-processed clinical reports as per semantics, and converted into probabilistic values. We use the Technique TF/IDF to extract relevant features. Word bags were also taken into consideration, extracting unigrams, bigrams, too. We identified 40 relevant features which could be used to achieve the classification. These characteristics are shown in Fig. 6. It is provided to machine learning algorithms by giving the corresponding weight to the feature and the same input.

Fig. 6

lungs	chest	patient	multiple	peripheral	bilateral	lower	lung	leave	image	lob	opacities	ct	right	lobe	air	pneum	glass opacities	history
0.379	0	0	0	0	0.34539	0.373	0.379	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.612	0	0	0	0	0	0	0	0	0.45	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.31916	0.241498	0.13287	0	0	0.24	0	0	0.119823	0	0	0	0	0.16	0.256674173	0.223
0	0	0	0	0	0.45603	0	0	0	0.573	0	0	0	0	0	0	0	0	0
0.342	0	0	0	0	0	0.336	0.342	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.50247	0.298721	0	0	0	0	0	0	0	0	0	0	0	0	0.393524911	0
0	0	0	0.26141	0	0	0	0	0	0	0	0	0.3	0.357	0.3	0.3	0	0	0
0	0	0	0	0	0	0	0	0	0	0.41	0	0	0	0	0	0	0	0
0	0	0.2225	0	0.340237	0	0	0	0.2	0	0.17	0	0	0	0	0	0	0	0.314
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.481
0	0	0	0.28544	0	0	0	0	0	0	0	0	0.3	0.39	0.4	0.4	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.317
0	0	0	0	0	0	0	0	0	0	0	0	0.4	0.282	0.3	0.3	0	0	0.238

Machine learning classification

Classification is conducted to classify the given text into four different virus types. The four virus classes, COVID (a person with coronavirus), ARDS, SARS, and both (a person with both corona and ARDS virus). Different supervised algorithms of machine learning are used to classify the text into those categories. For this task, we're used machine learning algorithms such as support vector machine (SVM), multinomial Naïve Bayes (MNB), logistic regression, decision tree, random forest, bagging, Adaboost and stochastic gradient boosting.

Traditional machine learning algorithms

Logistic regression

This algorithm predicts the class of numerical variables based on their label relation[15]. As shown in Figure, the 40 features selected in feature engineering with values are represented as a table and provided as an input. 6.

Generally, the algorithm calculates the probability of class membership. We have four classes in here. $Y \in \{0,\ldots\ldots\ldots\ldots\ldots,3\}$. With the help of Eq, the posterior probabilities can be calculated. help of Eq. 1.

$$P(y=k|x) = \frac{exp^{\varphi^T \theta_k}}{1+\sum_{k=1}^{3} exp^{\varphi^T \theta_k}} \forall k = 1,\ldots,3$$

$$P(y=0|x) = \frac{exp^{\varphi^T \theta_k}}{1+\sum_{k=1}^{3} exp^{\varphi^T \theta_k}} \quad (1)$$

Multinomial Naïve Bayes

Using Bayes rule[https://link.springer.com/article/10.1007/s41870-020-00495-9#ref-CR1616], MNB computes class probabilities of a given text. In our problem, let C denote the set of classes we have four classes C = 0, 1, 2, and 3. In addition, N is the set of features we have N = 40 here (40 features are taken using TF / IDF) as shown in Fig. 6. Then by using Bayes rule MNB assigns test text ti to the class that has the highest probability P(c|ti) 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 textual clinical data to the total number of textual clinical data labeled as class c. P(ti) is the probability of obtaining a report of clinical text such as ti in class c and is calculated as:

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

Where one is the word/terminus 'n' count in our clinical text report ti and P(wn|c) is the word/term 'n' probability given in class c. The latter probability is calculated from the training data with:

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

Where Fxc is the word/term 'x' count in all Class c clinical training reports. To avoid the problem of zero frequency Laplace estimator which assigns value one to each word count is used.

Support vector machine (SVM)

Support vector machine (SVM) is a supervised algorithm in machine learning to classify text into different categories[17]. It takes the given label to 'n' number of features for the particular text. Here we took 40 nature unigram and bigram features, as the dataset is small.

Here the training set's data points are (yk, xk)n1, where n is the number of traits taken. The 40 features selected with values in feature engineering are represented in table form and are supplied as an input, as shown in Fig. 6. The main aim of SVM is to build an Eq-shaped classifier. 3.

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

Where αk = real positive constant. B = true constant.

$$\psi(x, x_k) = \begin{cases} x_k^T x : LinearSVM \\ \left(x_k^T x + 1\right) d : PolynomialSVM with Degree d \\ exp(-||x - x_k||_2^2/\sigma^2 : RBFSVM \end{cases}$$

Where k, σ is constant.

The classifier is constructed on the assumption of:

ωT φ(xk) + b ≥ 1 , ifyk = +1 ωT

φ(xk) + b ≤ −1 , ifyk = −1

Which is equivalent to Eq. 4:

Yk [ωT φ(xk) + b ≤ −1 , if yk = −1 , k = 1,….,n (4)

Where φ(.)=nonlinear function maps input spaces into larger dimensional space.

It forms the hyperplane by which classification is carried out. The hyperplane distinguishes the four classes (COVID, ARDS, SARS, and

Both) in order to introduce a new variable k. Equation 5 is the Hyperplane equation:

$$yk[\omega T\ \varphi(\ xk) + b] \geq 1 - \xi k,\ k = 1,\ldots..,n$$
$$\xi k \geq 0\ ,\ k = 1,\ldots\ldots,n \qquad (5)$$

Decision trees

An alternative classification approach, it partitions the input space into regions and independently classifies each region[18]. The 40 features selected with values in feature engineering are represented in table form and are supplied as an input, as shown in Fig.
6. It divides the space according to the inputs recursively, and classifies at the bottom of the tree. The nodes on the leaf classify the text into four classes. A vital function which is known as the splitting criterion must be considered when building a decision tree. The function defines how to split data so as to maximise performance.

$$IGR(EX,a)=IG/IV \qquad (6)$$

Where IG = Gaining information. IV = Information which is intrinsic. The gain of information is calculated using entropy, as shown below:

$$IG(Ex,a)=H(Ex)-\sum v \in values(a)(|\{x \in Ex|value(x,a)=v\}||Ex|.H(\{x \in Ex|value(x,a)=v\}))$$

Where Ex = set of training examples and x x rangeEx define the value of a particular example x for a function. H = entropy and characteristics a =.

It calculates the intrinsic value of information by:

$$IV(Ex,a)=-\sum v \in values(a)|\{x \in Ex|VALUE(X,A)=v\}||Ex|.LOG2((\{X \in Ex|value(x,a)=v\})|Ex|)$$

Ensemble machine learning techniques

Bagging

An ensemble machine learning algorithm which improves the performance of other machine learning algorithms for classification and regression[19]. Bagging algorithm helps to prevent overfitting. Let size 'n' be given to a training set X, by sampling uniformly 'm' new training sets Xi are generated with replacements each having size 'n.' The 40 features selected with values in feature engineering are represented in table form and are supplied as an input, as shown in Fig. (6). Some observations could repeat in each Xi due to replacements. If m'=n then set Xi for large n to have a fraction (1 − 1 / e) of the unique X examples, the rest are duplicates.

AdaBoost

AdaBoostThis learning algorithm ensemble works with those weighted instances of the dataset[20]. The 40 features selected with values in feature engineering are represented in a table form and provided as an input, as shown in Fig 6. It starts with having equal weights for every observation and uses weighted data to train a weak learning algorithm.

By performing this, we produce a weak classifier. Choose a coefficient α according to the performance of this weak learning classifier. On misclassified points the weights improve and the weights of the correctly classified points decrease.Then again execute the weak learning algorithms to get a weak classifier for the new weighted data.
Repeating that procedure leads to an AdaBoost classifier being developed.

Random forest classifier

Ensemble machine learning algorithm used to classify and works like a tree of decisions. To train the random forest algorithm, the bootstrap aggregation technique is used. The overall prediction can be made by an averaging of all the individual regression trees. The majority vote is taken in case of Classification trees.

This algorithm uses a modified algorithm for tree learning that selects and splits each learning process by a subset of random features[21]. The 40 features selected with values in feature engineering are represented in table form and are supplied as an input, as shown in Fig. 6.The algorithm creates a forest from a subset of randomly selected data with the help of various decision trees, and summarises the votes for the decision trees to decide the final class of the object.

Stochastic gradient boosting

This algorithm allows for greedy growth of trees from training dataset samples. The 40 features selected with values in feature engineering are represented in a table form and provided as an input, as shown in Fig 6. This is used to reduce the gradient-boosting correlation between the trees. At each iteration a subsample of the training data is drawn at random without replacing the complete training data set. Instead of the full sample the randomly selected subsample is then used to fit the base learner[22].

Results and discussion

To perform this work we used a windows system with 8 GB Ram and 3.2 GHz processors. Sciket learning tool is used to perform classification of machine learning with the help of various libraries such as NLTK, STOPWORDS etc. to improve the accuracy of all the machine learning algorithms pipeline. Once the statistical computation was performed, deeper insights were achieved on the data. The data is split into a ratio of 70:30, where 70 percent of data is used to train the model and 30 percent is used to test the model.

We have clinical text reports of 212 patients labelling in four classes.We split our initial data set into separate training and test subsets to explore the generalisation of our model from training data to unseen data, and reducing overfitting possibilities and the tenfold cross-validation strategy was implemented for all algorithms, and this process was repeated five times independently to avoid the sampling bias introduced by random partitioning of the data set into the cross-validation.

Table 1 provides a comparative analysis of all the classical methods of machine learning used to perform that task.Table 2 presents a comparative analysis of all classical machine learning methods and the Ensemble learning methods used to perform the task of classifying the clinical text into four classes. Results showed that logistical regression and Multinomial Naïve Bayes Algorithm show better results than all other algorithms by 94 percent accuracy, 96 percent recall, 95 percent F1 score and 96.2 percent accuracy with other algorithms such as random forest, and 94.3 percent accuracy with gradient boosting also showed good results.

The comparative analysis visualised for all the algorithms used in our work is shown in Fig. 7. The COVID-19 data is least available, since we all know.We experimented with it in two stages to get the model 's true accuracy. We took 75 percent of the available data in the first stage and it shows less accuracy compared to the stage at which entire data was used for experimentation. So we can conclude that if more data are provided to these algorithms, there is a chance of performance improvement. As we face a serious challenge in tackling the deadly virus, our work will somehow help the community by analysing the clinical reports and taking the necessary action. It has also been analysed that the length of reporting for COVID-19 patients is much smaller than other classes and ranges from 125 characters to 350 characters.

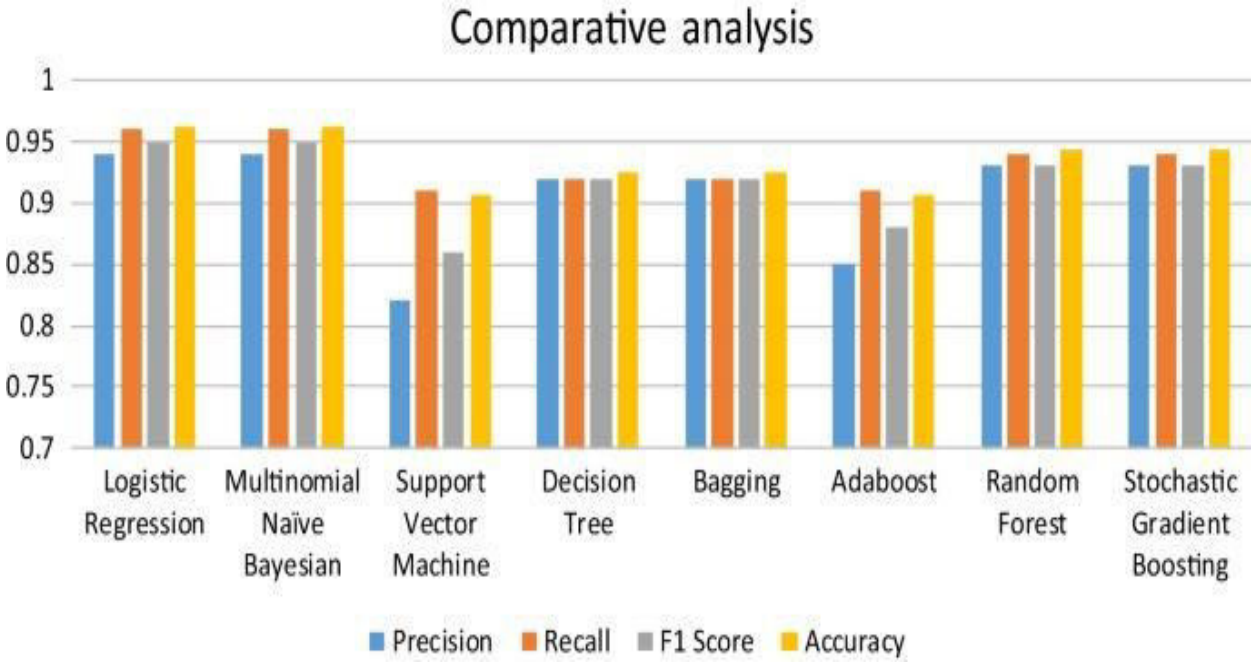
Table 1 Comparative analysis of traditional algorithms regarding machine learning

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 1 Comparative analysis of traditional algorithms regarding machine learning

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



Conclusion

COVID-19 has shocked the world because of its vaccine or drug inaccessibility. Different researchers are working to conquer that deadly virus. We used 212 clinical reports labelled as COVID, SARS, ARDS and both (COVID, ARDS) in four classes. From these clinical reports are extracted various features such as TF / IDF, word bag. The algorithms of machine learning are used to classify the clinical reports into four different classes.After classification it was revealed that logistical regression and multinomial Naïve Bayesian classifier deliver excellent results with 94% accuracy, 96% recall, 95% f1 score and 96.2 percent accuracy.

The random forest, stochastic gradient boosting, decision trees and boosting were several other machine learning algorithms that showed better results. Model efficiency can be improved by increasing the data volume. The disease can also be classified on a gender-based basis, so we can get information about whether more males or females are affected. For better results, more feature engineering is needed and a deep learning approach may be used in future.

References

[1] http://www.emro.who.int/health-topics/corona-virus/about-covid-19.html, http://www.emro.who.int/health-topics/corona-virus/news.html ; 2020

[2] https://www.deccanherald.com/national/coronavirus-india-update-state-wise-tot al-number-of -confirmed-cases-deaths-on-july-15-861334.html ; 2020

[3] Shreshth Tulia, Shikhar Tuli, Rakesh Tulic and Sukhpal Singh Gill. Predicting COVID-19 Pandemic growth and trend using cloud computing and machine learning; May 2020

[4] https://delhifightscorona.in/ ; Delhi Fights Corona » COVID-19 Response Updates from the Delhi Government; 2020

[5] Pal, R., Sekh, A. A., Kar, S., and Prasad, D. K. Country-based neural network Wise risk prediction of COVID-19. arXiv preprint arXiv:2004.00959 (2020)

[6] Uhlig, S., Nichani, K., Uhlig, C., and Simon, K. Pandemic Projections Model COVID-19 through the combination of epidemiological, statistical and neural networking approaches. medRxiv (2020)

[7] Wu, Y., Yang, Y., H., and Saitoh, M. Deep learning for epidemiological predictions. In Proceedings of 41st International ACM SIGIR Conference on Information Retrieval Research and Development (2018), pp. 1085–1088..

[8] C. A. Oyeleye , Grace Oladele, Oluwaseun Alade,Bello O. A, Ismail Adeyemo, Abiodun Emmanuel and Adedeji oluyinka. Modified genetic algorithm for solving nurse scheduling problem, IRJCS, ISSN: 2393-9842, Issue 04, Volume 07, April 2020.

[9] Disease transmission on Animals ; https://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=191484 ; April2017

[10] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Delhi#Timeline ; 2020

[11] Networkx documentation ; https://networkx.github.io/ ; September 2017

[12] Chollet F et al. Keras. https://github.com/fchollet/keras ; 2015.

[13] Ah ChungTsoi,AndrewBack.Discrete time recurrent neural network architectures: A unifying review,https://doi.org/10.1016/S0925-2312(97)00161-6;1997.

[14] Sequence Tagging models Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF, https:/arxiv.org / abs/1508.01991;2015

[15] https://www.boente.eti.br/fuzzy/ebook-fuzzy-mitchell.pdf ; An Introduction to Genetic Algorithms Mitchell Melanie, A Bradford Book The MIT Press Cambridge, Massachusetts • London, England Fifth printing, 1999