United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

Machine Learning-Based Approaches For Detecting COVID-19 Using Clinical Text Data

M. Nandhini[1] , G. Bhanu Sruthi[2] , R. D. S. Dikshitha[3] , K .Naga Durga[4]

[1,2,3,4]UG Scholars, Dept. Of Computer Science and Engineering, R.K. College of Engineering, Vijayawada, India

mahantinandhini@gmail.com , gbhanusruthi@gmail.com , repalledijendhrasivadeekshitha@gmail.com, nagadurga@gmail.com

ABSTRACT - 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 then 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 frequence/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 Bayes showed better result 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

I.INTRODUCTION

In December 2019, the novel coronavirus appeared in the Wuhan city of China [1] and was reported to the World Health Organizatin (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]. COVID-19 is 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 track 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).

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 hug 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.

II.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] developed 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.

III.METHODOLOGIES

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 the Sentiments of Indian government projects with the help of the lexicon-based dictionary. 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, and 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] developed 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.

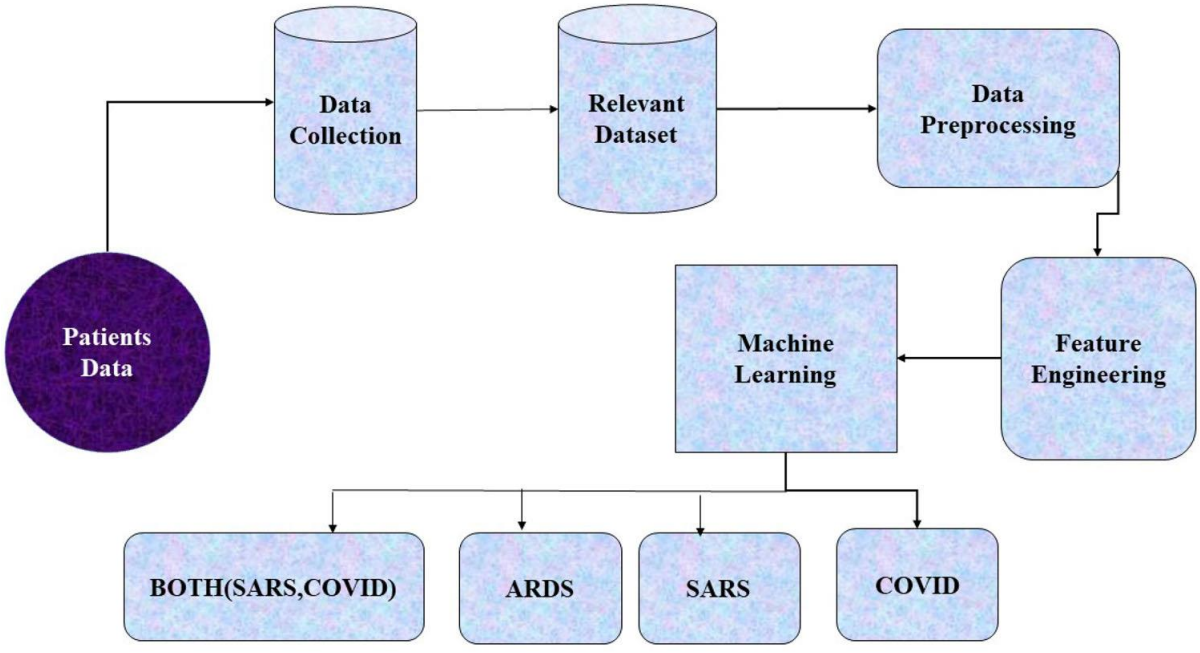


Figure 1: Methodology

3.1 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.1 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_O2, extubated, temperature, pO2_saturation,

leukocyte_count, neutrophil count, lymphocyte count, view,modality, date, location, folder, filename, DOI, URL.License. Clinical notes and other notes.

3.2 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 2 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, SARS and Both (COVID, ARDS). Figure 2 shows the different classes in which clinical text is being categorized and corresponding report length.

3.3 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 2 shows the main steps in preprocessing.

3.4 Feature engineering

From the preprocessed clinical reports, various features are extracted as per the semantics and are converted into 1 https://github.com/Akibkhanday/Meta-data-of-Coronavirus. Int. j. inf. tecnol. (September 2020) 12(3):731–739 733123

IV. RESULTS AND DISCUSSION

We used a windows system with 4 GB Ram and 2.3 GHz processors for performing this work. Sciket 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 Naı¨ve 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.

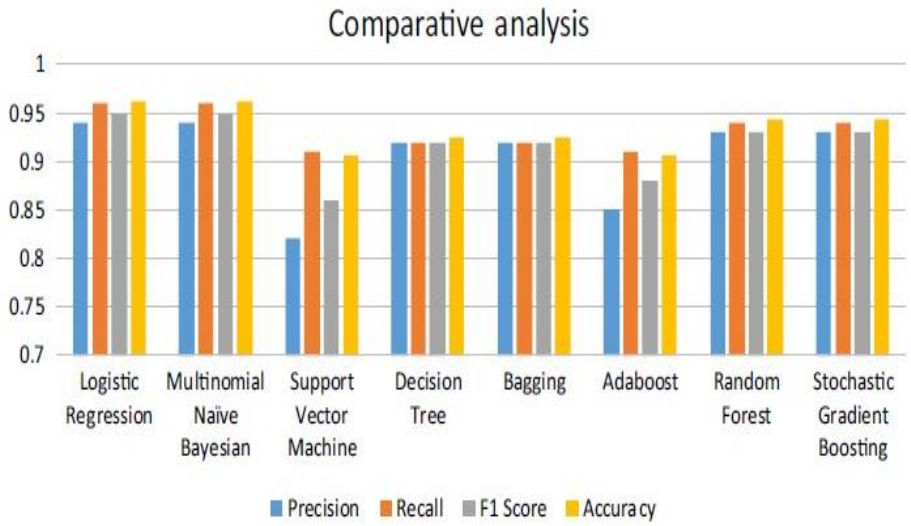
Table 1 Comparative analysis of traditional machine learning algorithms

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



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 severe challenge in tackling the deadly virus, our work will somehow help the community by analysing the clinical reports and take necessary actions. Also, it was analyzed

that the COVID-19 patients report length is much smaller than other classes and it ranges from 125 characters to 350 characters.

V. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

COVID-19 has shocked the world due to its non-availability of vaccine or drug. Various researchers are working to conquer 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 Nai¨ve Bayesian classifier gives excellent results by having 94% precision, 96% recall, 95% f1 score and accuracy 96.2%.

5.2 Future Scope

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.

VI.REFERENCES

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china. Nature 44(59):265–269
2. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape. com/viewarticle/924596
3. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513
4. World health organization: https://www.who.int/new-room/g-adetail/ q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020
5. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/Template:2019%E2%80%9320_coronavirus_pan demic_data. Accessed 10 Apr 2020
6. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., "Face Recognition Techniques: A Critical Review" 2018
7. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0072-1
8. Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte.C6612.098319
9. Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy. Int J Inf Technol 10:257–263. https://doi.org/10.1007/s41870-018-0088-1
10. Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0270-5
11. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against COVID-19. https://arxiv.org/abs/2003.11336v1

12. Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. https://arxiv.org/abs/2003.09871
13. Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, Amd A, Luo BJ, Cao Z, Xu H, Yuan Y (2020) Prediction of criticality in patients with severe covid-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv. https://doi.org/10.1101/2020.02.27. 20028027
14. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Compu Mater Contin 63(1):537–551
15. Description of Logistic Regression Algorithm. https://machine learningmastery.com/logistic-regression-for-machine-learning/. Accessed 15 May 2019
16. Description of Multinomial Naı¨ve Bayes Algorithm https://www. 3pillarglobal.com/insights/document-classification-using-multi nomial-naive-bayes-classifier. Accessed 15 May 2019
17. Khanday AMUD, Khan QR, Rabani ST. SVMBPI: support vector machine based propaganda identification. SN Appl. Sci. (accepted)
18. Description of Decision Tree Algorithm: https://dataspirant.com/ 2017/01/30/how_decision_tree_algorithm_works/. Accessed 10 July 2019