International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Machine Learning Based Approaches for Detecting Covid19

Attluri Kavya Sree[1], Ganja Pravinya[2], Geddam Roja Blessy[3], K Srilatha[4]

Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

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 analyzing 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 solutions 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: Classification, COVID-19, Decision tree, Disease prediction, Logistic regression, Machine learning, Random Forest.

1. Introduction

1.1 About Paper

The virus created a global threat and was named as COVID 19 by WHO on 11th February 2020.The COVID 19 is the family of the viruses including sars ,ards .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 corona virus .It can also forecast nature of the virus across the globe. Supervised Machine learning algorithm needs annotated data for classifying text or image into different categories. In this project we will collect data, we will consider only the relevant data by using data preprocessing and feature engineering and after that by using different classifiers we calculate metrics like accuracy, f1 measure, recall, precision. And we visualize them.

1.2 Objectives of the Paper

The main objective of this project is to detect the corona virus. To help in controlling the current havoc with the help of ML algorithm. To classify textual clinical reports into four categories COVID, SARS, ARDS and both (COVID, ARDS).

1.3 Scope of the Paper

Hospitals, Patient can get results of detection of covid-19 effectively. This is really helpful. Clinics, this project can be useful in clinics also where you can collect clinical data and can perform computations on those data and then after doing all necessary actions, we can visualize the results. Diagnostic centers this project can be useful even in diagnostic centers also. Because this project can help with diagnosis of the disease.

2. Literature Survey

2.1 Existing System

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 tasks in text mining and can be performed using different algorithms. 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. Limitations of Existing System include It is imperative to develop a control system that will detect the coronavirus. People all over the world are vulnerable to its consequences in future

2.2 Proposed System

Proposed a machine learning model that can predict a person affected with COVID-19 and has the possibility to develop acute respiratory 8 distress syndrome (ARDS). The proposed model resulted in 80% of accuracy. The samples of 53patients 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. Advantages of Proposed System include, we can predict the person affected with corona virus using machine learning model. So that we can avoid the extreme problem in future.

3. Proposed Architecture

3.1 Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

3.2 Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

3.3 SYSTEM DESIGN

3.3.1 Introduction

System Design is the process of art of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. One could see it as the application of the systems theory to product development. There is some overlap and synergy with the disciplines of systems analysis, systems architecture and systems engineering.

3.3.2 Architecture

The main objective of this project is, It is imperative to develop a control system that will detect the coronavirus. One of the solutions is to control the current havoc can be the diagnosis of disease with the help of various AI tools. 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 analyzing 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 solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools.

3.3.3 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 tasks in text mining and can be performed using different algorithms. 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. analyzed 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. 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. 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., 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 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 that will be affected more severely can help in directing assistance and planning medical resource allocation and utilization. Yan et al. used machine learning to develop a prognostic prediction algorithm to 21 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. 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 23 predicting using text, we used machine learning and ensemble learning models to classify the clinical reports into four categories of viruses.

4. Implementation

4.1 Algorithm

Step1: Data Collection -We will collect samples of patients from various sources,

Step 2: Data Preprocessing-Then we will consider data which is relevant means we will remove unwanted data. And we visualize number of cases of COVID, ARDS, BOTH, SARS with the help of box plot.

Step 3: Feature Engineering-We will convert the textual data into a specific format so that we can do training and testing.

Step 4: Classification-We will use ML algorithms to categorize input into four categories. And we also calculate metrics like accuracy, f1 measure, recall, precision.

4.2 Implementation

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

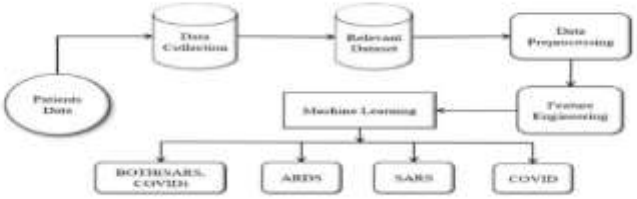

Fig. 1. Data Flow Diagram of COVID- 19 Diagnoses

This Project contains information Server and Data collected from WHO (World Health Organization) as a Dataset. Selecting and Uploading the Dataset file. Extract all the text data and remove all stop words. Preprocess all stop words are removed. Feature Engineering the data to apply TF-IDF features. Running all the traditional algorithms to calculate the accuracy. Running the classical algorithms. We analyze and compare, measuring the metrics which is displayed as a graph.

5.Result

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig. 2. An Application Window

Fig. 3. An Uploaded Dataset

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

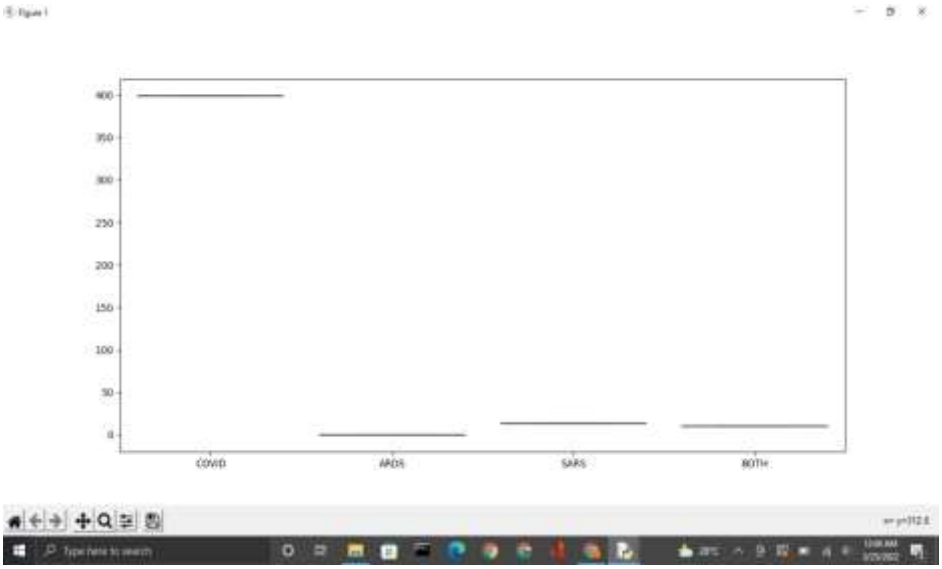

Fig. 4. Preprocess Dataset

Fig. 5. Feature Engineering

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

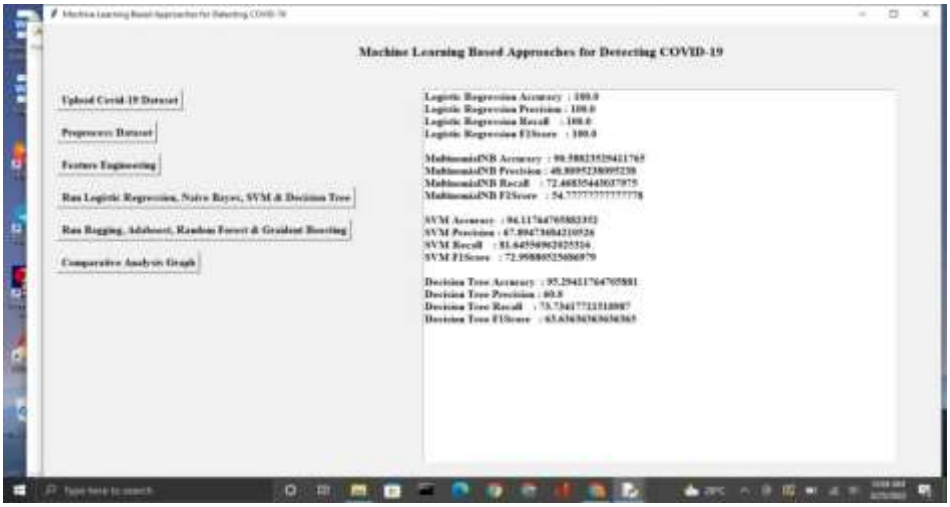

Fig. 6. Run Logistic Regression, Naive Bayes, SVM & Decision Tree

Fig. 7. Run Bagging, Adaboost, Random Forest & Gradient Boosting

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

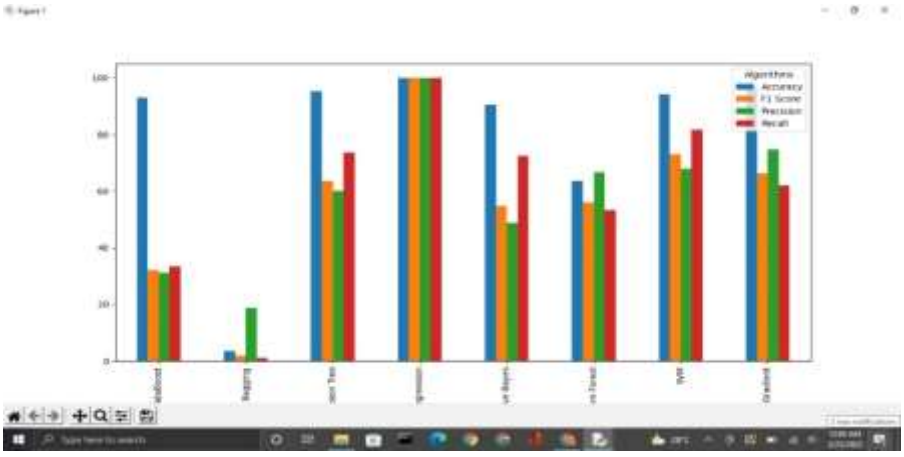

Fig. 8. Comparative Analysis Graph

6. 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 labeled in four classes namely COVID, SARS, ARDS and both (COVID, ARDS). Various features like TF/IDF, bag of words is 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 Naive 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 is affected more or female. More feature engineering is needed for better results and deep learning approach can be used in future.

7. Future Scope

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 is affected more or females. More feature engineering is needed for better results and deep learning approach can be used in future. In future recurrent neural network can be used for better accuracy.

8. References

1. 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
1. 2.Description.of.Logistic..Regression..Algorithmhttps://machine..learningmastery.com/logisticregression-for-machine-learning/. Accessed 15 May 2019
2. 3 Description of Multinomial Naive Bayes Algorithm https://www. 3pillarglobal.com/insights/document-classification-using-multi nomial-naive-bayes-classifier. Accessed 15 May 2019
3. Khanday AMUD, Khan QR, Rabani ST. SVMBPI: support vector machine-based propaganda identification. SN Appl. Sci. (accepted)
4. 5.Description.of.Decision..Tree..Algorithm:..https://dataspirant.com/..2017/01/30/how_decision_tree_algorithm_works/. Accessed 10 July 2019
5. Description of Boosting Algorithm: https://towardsdatascience. com/boosting. Accessed 10 July 2019
6. Description of Adaboost Algorithm: https://towardsdatascience. com/boosting-algorithmadaboost-b673719ee60c. Accessed 10 July 2019
7. Katuwal R, Suganthan PN (2018) Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, Arxiv:1802.01240v1
8. Friedman JH (2002) stochastic gradient boosting. Computer. Stat. Data Anal. 38(4):367–378. https://doi.org/10.1016/S0167- 9473(01)00065-2
9. 10.Wikipedia.coronavirus.Pandemic..data:..https://en.m.wikipedia.org/wiki/Template:2019%E2%80%9320_coronavirus_pan demic_data. Accessed 10 Apr 2020
10. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., ‘‘Face Recognition Techniques: A Critical Review’’ 2018
11. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured..data:..a..SWOT..analysis…Int..J..Inf..Technol.

12. 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

13. MVSPN Dr Y V S Sai Pragathi, Analysis and implementation of realtime stock prediction using reinforcement frameworks, Journal of Physics: Conference Series 2089 (1).2021
14. DYVSS Pragathi, MVSP Narasimham, Realtime stock prediction using Transformations and Modeling, Journal of Contemporary Issues in Business and Government 27 (3), 2129-2135, 2021
15. YVSSP M V S Phani Narasimham,Realtime Cost performance Improved Reservoir Simulator Service using ANN and Cloud Containers, International Journal of Innovative Technology and Exploring Engineering 9 …, 2020
16. P Yellanki, MVSP Narasimham,Secure Routing Protocol for VANETS using ECC, 2020 International Conference on Computer Science, Engineering and …, 2020
17. YVSSP Dumala Anveshini, Real-Time Emulation to Investigate the performance of LANMARK Routing Protocol in MANET, International Journal of Advanced Science and Technology 29 (3), 9395-9404, 2020
18. A Parveen, YVS Sai Pragathi,A study of routing protocols for energy conservation in manets, Advances in Decision Sciences, Image Processing, Security and Computer …, 2020
19. YVSSP M V S Phani Narasimham,Development of Realistic models of oil well by modeling porosity using modified ANFIS technique, International Journal on Computer Science and Engineering 11 (07), 34-39, 2019