Rover Publications

United International Journal of Engineering and Sciences (UIJES)

An International Peer-Reviewed (Refereed) Engineering and Science Journal Impact Factor:7.984(SJIF) Volume-4, Special Issue-3; ISSN: :2582-5887

# Covid-19 Detection

**K Siva Naga Raju, I Phanindra Babu, SK Ibraheem, B Sai Kumar, C Harsha Vardhan**
Computer Science and Engineering, JNTU Kakinada
RK College of Engineering
Vijayawada, India.
koraganjisivanagaraju@gmail.com

**DOI:10.53414/UIJES:2024.43.474**

**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 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 Naı¨ve 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 neural networks, convolutional neural networks, healthcare, infectious diseases.**

## I. 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 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 hydroxy chloriquine, 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 PaO2/FiO2 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.

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.

## II. 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 task in text mining and can be performed using different algorithms

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

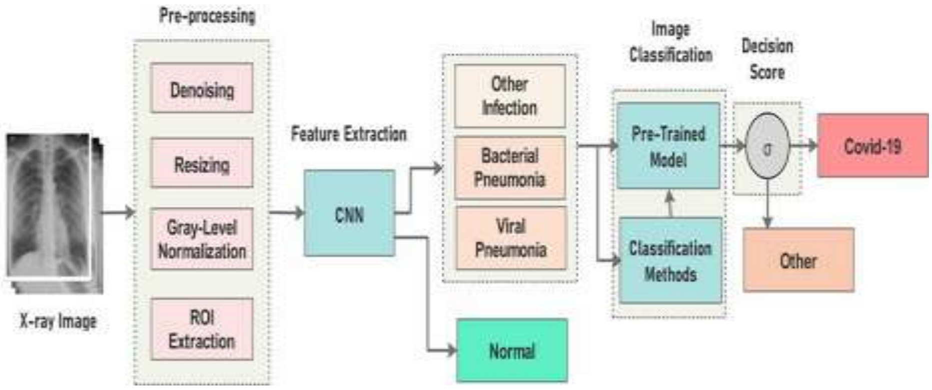

Fig. 1. Basic Architectural Diagram of COVID-19 Detection System

## III. Proposed System

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.

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.

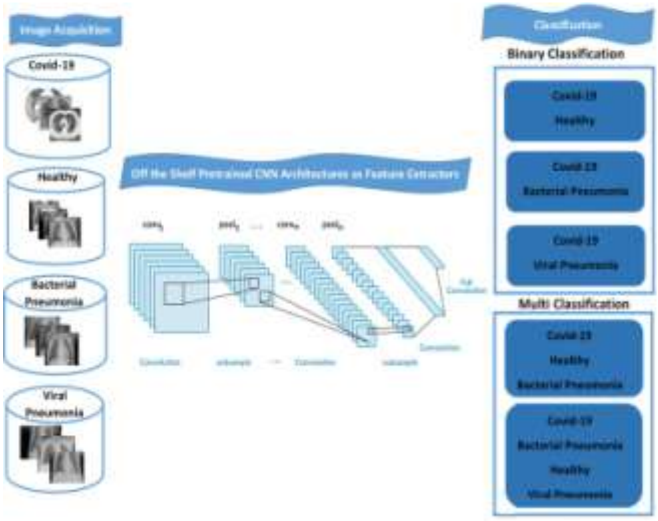

Fig. 2: Proposed system for detection and classification of COVID-19

.

## IV. Relevant Data Set

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, SARS and Both (COVID, ARDS). Figure 4 shows the different classes in which clinical text is being categorized and corresponding 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.

## V Technique Used or Algorithm Used

we are using traditional and classical machine learning algorithms to predict COVID-19 disease. In traditional algorithms we are using Logistic Regression, Naïve Bayes, SVM and Decision Tree and in classical algorithms we are using Bagging, AdaBoost, Random Forest and Stochastic Gradient Boosting classifier. In all algorithms Logistic Regression giving better performance.

## VI . Machine Learning Classification

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 Naı¨ve Bayes (MNB), logistic regression, decision tree, random forest, bagging, Adaboost and stochastic gradient boosting were used for performing this task

## VII. 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 Naı¨ve 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

## 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_pandemic_data. Accessed 10 Apr 2020