Turkish Journal of Computer and Mathematics Education Vol.14 No.03 (2023),598-603
Research Article

MACHINE LEARNING BASED APPROACHES FOR DETECTING COVID-19 USING CLINICAL TEXT DATA

Venu M [1], Revathy P [2], Naveen Kumar A [3]
[1] Assistant Professor, Department of Computer Science and Engineering
[2] Assistant Professor, Department of Computer Science and Engineering
[3]Assistant Professor, Department of Computer Science and Engineering
[1] Narsimha Reddy Engineering College,Kompally,Hyderabad,India.
[2] Narsimha Reddy Engineering College,Kompally,Hyderabad,India.
[3] Narsimha Reddy Engineering College,Kompally,Hyderabad,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 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: COVID-19, ML, High accuracy, AI.

INTRODUCTION

The main objective of this project is 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.

The outbreak of the novel coronavirus disease 2019 (COVID-19) in late 2019 has posed unprecedented challenges to healthcare systems worldwide. As the pandemic continues to evolve, early and accurate detection of COVID-19 cases remains crucial for effective disease management, resource allocation, and public health interventions. Machine learning (ML) techniques have emerged as powerful tools in the battle against COVID-19, particularly when applied to clinical text data.

Clinical text data encompass a wide range of information, including electronic health records (EHRs), radiology reports, medical notes, and

patient narratives. This rich source of unstructured data contains valuable insights that can aid in the timely identification and monitoring of COVID-19 cases. In this context, ML-based approaches have demonstrated their potential in assisting healthcare professionals, researchers, and policymakers by automating the detection and analysis of COVID-19-related information embedded in clinical text data.

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.

PROPOSED SYSTEM

The 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.

WORKING METHODOLOGY

This paper aims to provide an overview of the application of machine learning-based approaches for detecting COVID-19 using clinical text data. We will explore the following key aspects:

Importance of Clinical Text Data: Clinical text data play a pivotal role in healthcare as they capture a patient's medical history, symptoms, treatments, and outcomes. Leveraging this textual information for COVID-19 detection offers a holistic view of the patient's condition, aiding in more accurate diagnoses.

Challenges in COVID-19 Detection: We will discuss the challenges in diagnosing COVID-19, including the variability in symptoms, the need for rapid testing, and the potential for asymptomatic carriers. These challenges underscore the importance of ML-based approaches that can handle diverse and evolving clinical scenarios.

Machine Learning Techniques: We will delve into various ML techniques employed for COVID-19 detection, such as natural language processing (NLP), deep learning, and ensemble methods. These techniques enable the extraction of meaningful patterns and insights from

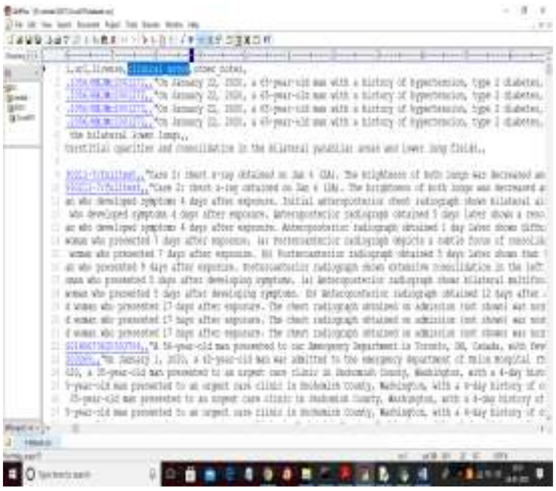
clinical text data, facilitating early diagnosis and decision-making.

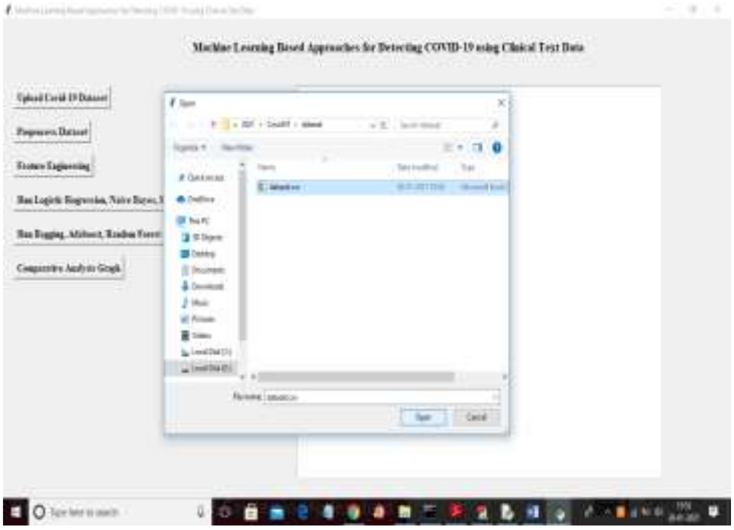
Data Sources: The availability and quality of clinical text data are essential factors in developing effective ML models. We will explore the sources of clinical text data, including EHRs, radiology reports, and social media, and discuss the advantages and limitations of each.

Model Performance and Validation: Evaluating the performance of ML models for COVID-19 detection is critical. We will examine different metrics and validation techniques used to assess the accuracy, sensitivity, specificity, and generalizability of these models.

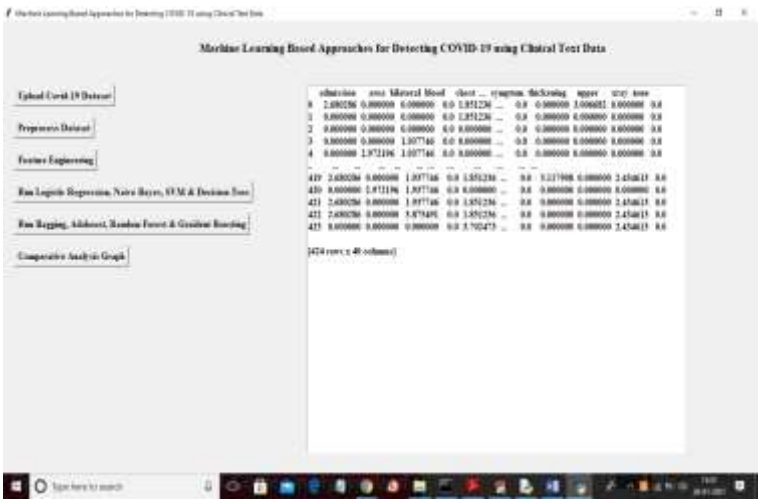
Ethical and Privacy Considerations: The use of clinical text data in ML applications raises important ethical and privacy concerns. We will discuss the challenges related to patient data protection, informed consent, and the responsible use of AI in healthcare.

Future Directions and Implications: Finally, we will highlight the potential future directions in this field, including the integration of ML-based approaches into clinical practice, the development of standardized datasets, and the implications for public health policy.






Machine Learning Based Approaches for Detecting COVID-19 using Clinical Text Data
Upload Covid 19 Dataset
Preprocess Dataset
Feature Engineering
Run Logistic Regression, Naive Bayes, SVM & Decision Tree
Run Bagging, Adaboost, Random Forest & Gradient Boosting
Comparative Analysis Graph


Machine Learning Based Approaches for Detecting COVID-19 using Clinical Text Data
Upload Covid 19 Dataset
Preprocess Dataset
Feature Engineering
Run Logistic Regression, Naive Bayes, SVM & Decision Tree
Run Bagging, Adaboost, Random Forest & Gradient Boosting
Comparative Analysis Graph

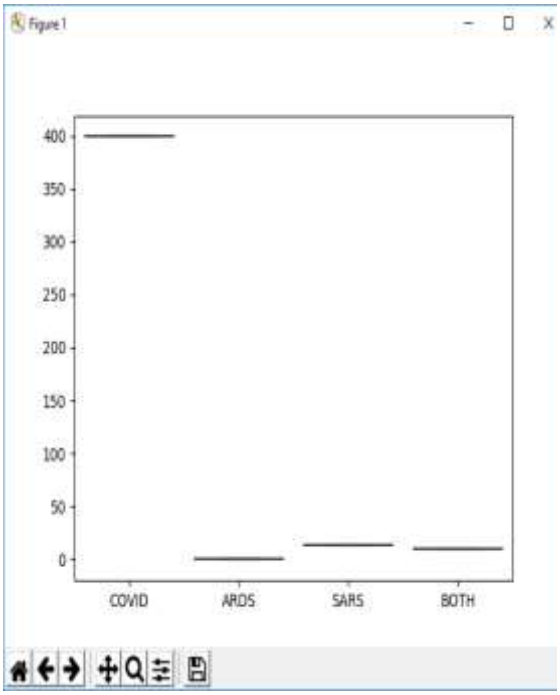
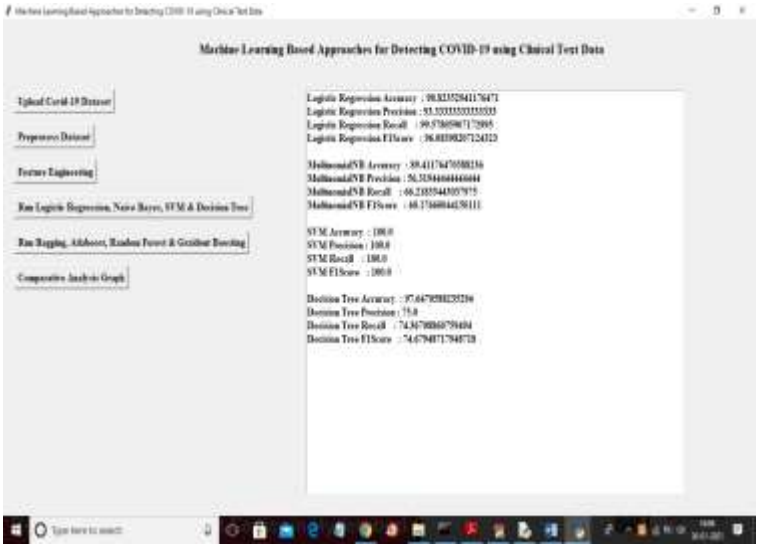
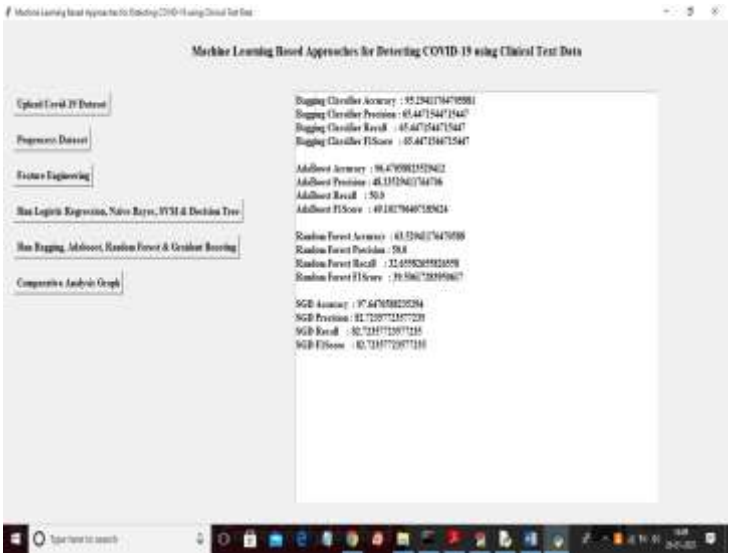
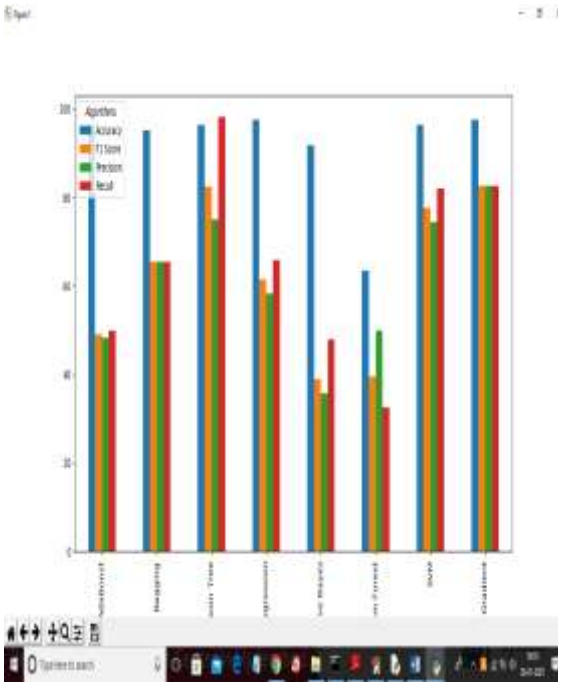


Figure 1
400
350
300
250
200
150
100
50
0
COVID
ARDS
SARS
BOTH


Machine Learning Based Approaches for Detecting COVID-19 using Clinical Text Data
Upload Covid 19 Dataset
Preprocess Dataset
Feature Engineering
Run Logistic Regression, Naive Bayes, SVM & Decision Tree
Run Bagging, Adaboost, Random Forest & Gradient Boosting
Comparative Analysis Graph


Machine Learning Based Approaches for Detecting COVID-19 using Clinical Text Data
Upload Covid 19 Dataset
Preprocess Dataset
Feature Engineering
Run Logistic Regression, Naive Bayes, SVM & Decision Tree
Run Bagging, Adaboost, Random Forest & Gradient Boosting
Comparative Analysis Graph



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. 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 getinformation 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.

REFERANCES

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