Covid-19 Risk prediction

Predicting Covid Severity based on Patients' Personal Records

Mario Ghaly
CSCE
AUC
Sohag
mariomamdouh@aucegypt.edu

Moneer Zaki
CSCE
AUC
5th settlement
moneerzaki@aucegypt.edu

ABSTRACT

The COVID-19 pandemic has presented unprecedented challenges to global healthcare systems, necessitating the development of innovative approaches to managing patient care and allocate medical resources effectively. Among these challenges is predicting the severity of COVID-19 in its patients upon diagnosis since it is considered a critical task for improving clinical outcomes and optimizing healthcare delivery. Thus, this project searches how to manifest machine learning(ML) techniques and algorithms to predict COVID-19 severity based on patients' medical data. Leveraging a dataset from Kaggle, which includes demographic information, medical history, symptoms, and laboratory test results, we aim to develop a predictive model that can classify patients according to their risk of experiencing severe outcomes. First, this report investigates prior literature reviews targeting the same issue assessing the strengths and limitations of existing ML-based approaches tried on this problem. Our project is going to include data preprocessing, feature engineering, and the evaluation of several ML algorithms to identify the most effective and accurate model for predicting coronavirus severity. We also provide an in-depth analysis of possible datasets for our problem, discussing their applicability, limitations, and how they support our research objectives. Finally, this report includes our project purpose, proposed solution, and future applications. In the context of COVID-19, not only does this project contribute to the goal of enhancing patient care post the pandemic, but also offers insights into the application of ML in addressing complex health challenges.

INTRODUCTION

In the wake of the COVID-19 pandemic, the global healthcare community has been thrust into the forefront of an unprecedented battle against a virus that has shown a wide range of clinical outcomes in infected individuals. From asymptomatic carriers to patients requiring immediate intensive care, the variability in how individuals respond to the SARS-CoV-2 virus underscores the urgent need for innovative tools that can predict patient outcomes early and accurately. This project is motivated by the potential of machine learning (ML) to revolutionize our approach to predicting COVID-19 severity among patients. By harnessing the power of ML algorithms to analyze vast datasets comprising personal information and clinical data, we aim to develop a predictive model that could significantly impact treatment decisions and improve patient management. The goal is to not only enhance the precision of medical interventions but also to optimize resource allocation within overwhelmed healthcare systems, ultimately contributing to better patient outcomes and a more resilient public health infrastructure in the face of this ongoing global health crisis.

MOTIVATION

The COVID-19 pandemic emerged in late 2019 and has since negatively influenced global health, economic, and social challenges. There have been millions of victims due to this virus and society's lack of data and information about it. However, a critical aspect of managing the pandemic involves effectively predicting the severity of COVID-19 in patients, which can significantly impact treatment decisions, resource allocation, and patient outcomes. The variability in COVID-19 outcomes, ranging from asymptomatic cases to severe respiratory failure necessitating intensive care, underscores the need for precise predictive tools. Thus, we intend in this project to design a sophisticated ML model that, given a COVID-19 patient's medical data, can predict how severe the disease is.

PROBLEM SPECIFICATION

The project aims to develop a predictive model using machine learning to forecast the severity of COVID-19 in patients based on their personal and clinical data. This entails identifying key predictors of disease severity from a range of variables, including demographic information, medical history, symptoms, and laboratory test results. The ultimate goal is to enable healthcare providers to make more informed decisions regarding patient care and resource allocation, thus improving patient outcomes and healthcare efficiency during the pandemic. The model's effectiveness will be evaluated based on its accuracy, reliability, and the comparative performance of different machine learning algorithms applied to this task.

LITERATURE REVIEW

Machine learning approaches to predict COVID-19 severity[1]

In this context, machine learning (ML) and artificial intelligence (AI) have emerged as powerful tools for analyzing complex datasets and extracting meaningful patterns that can predict disease outcomes. This literature review focuses on existing attempts to predict COVID-19 severity using machine learning-based approaches, highlighting the methodologies, features used, and their performance.

1. Deep Learning Approaches

A significant portion of the research has focused on deep learning techniques, particularly using chest X-ray and CT images to predict the severity of COVID-19 infections. Convolutional Neural Networks (CNNs) are at the forefront of these efforts due to their ability to extract intricate patterns from image data. For instance, studies leveraging pre-trained models such as CheXNet, ResNet, and DenseNet have shown promising results in identifying features indicative of COVID-19 severity from imaging data. One such study utilized a pre-trained DenseNet model to assess severity based on lung involvement and opacity scores, achieving high accuracy in severity classification. These approaches benefit from the rich, complex features extracted from images, which correlate with disease severity, such as lung opacities and involvement extent.

2. Hybrid and Handcrafted Feature Approaches

Aside from purely deep learning-based methods, some studies have explored hybrid approaches combining handcrafted and deep learning-derived features. These studies manually extract specific features from images before using machine learning algorithms for classification or regression tasks. For example, combining PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination) for feature selection before classification has been shown to improve model performance significantly. Such methods blend the

interpretability of handcrafted features with the robustness of deep-learned features, providing a comprehensive view of the data that enhances prediction accuracy.

3. Machine Learning Algorithms

Classical machine learning algorithms have also been applied to predict COVID-19 severity, often using clinical and demographic data. Algorithms like Support Vector Machines (SVM), Random Forests, and Gradient Boosting (e.g., XGBoost) have been employed to classify patients into different severity categories based on features such as age, pre-existing conditions, and laboratory results. These studies highlight the potential of machine learning models to utilize readily available clinical data for early severity prediction, which is crucial for patient triage and resource allocation.

4. Performance Metrics

The performance of these machine learning models is typically evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). For instance, hybrid models using PCA and RFE feature selection followed by XGBoost classification have achieved accuracy rates as high as 97%, with similar excellence in precision, recall, and F1-score. Such high performance underscores the potential of machine learning models in effectively predicting COVID-19 severity and aiding in the management of hospital resources.

5. Conclusion

The application of machine learning in predicting COVID-19 severity has shown considerable promise, with deep learning models utilizing imaging data and classical ML algorithms leveraging clinical and demographic data demonstrating high accuracy and performance. The integration of handcrafted and deep-learned features offers a balanced approach that maximizes the predictive power of the models. As the pandemic evolves, these predictive models play a crucial role in guiding clinical decisions, optimizing resource allocation, and improving patient outcomes. Future research may focus on refining these models, incorporating more diverse data sources, and improving their interpretability and generalizability across different populations and clinical settings.

DATASETS

Now, for the analysis of the existing datasets to accommodate our project purposes, we will analyze the following 2 datasets:

First: COVID-19 Dataset[2]

1. Overview

The first dataset, “COVID-19 patient's symptoms, status, and medical history,” is fully available on Kaggle. It is provided by the Mexican Government, and it is very relevant to the purpose of our study as it classifies the seriousness of COVID-19 illness according to the patient’s medical problems. Moreover, the boolean features of this dataset are of integer type, where 1 means "yes", 2 means "no", and values as 97 and 99 are missing data.

2. Limitations

A. Most of the features are binary rather than a scale for the severity of the patient’s prior diseases, which reduces the accuracy of the outcomes to an extent, as not all these medical issues are measured in reality as a yes or no.

B. Also, there is a feature indicating whether the patient has another disease or not. This is a limitation as the other disease may not be a reliable feature since we do not know what this disease is. In other words, in some instances, this disease could be of a high impact on the illness

severity, while in other cases, it could be irrelevant.

3. #Instances

There are 1,048,576 unique patients.

4. #Features

There are essentially 21 features. Examples of these features are

A. Age: an integer representing the age of the patient when he did the coronavirus test.
B. Asthma: An integer equal to 1 or 2 for whether the patient has asthma or not. There are many other features like Asthma, but for other diseases such as cardiovascular diseases
C. Date Died: an integer indicating the date of death or 9999-99-99 if the patient is still alive.

5. Label

The label in this dataset is named “classification.” It has 7 values, where 1-3 indicates that the patient was diagnosed with COVID in different degrees. 4 or higher means that the patient is not a carrier of COVID or that the test is inconclusive

Second: Covid-19 Case Surveillance Public Use Dataset[3]

1. Overview

COVID-19 Case Surveillance Public Use Data provided by the Centers for Disease Control and Prevention (CDC) includes individual-level data reported from various U.S. states, territories, and autonomous reporting entities. This dataset has been compiled since April 5, 2020, when COVID-19 was added to the Nationally Notifiable Condition List. The data is classified as "immediately notifiable, urgent (within 24 hours)" by the Council of State and Territorial Epidemiologists (CSTE). The dataset contains de-identified information, including demographic characteristics, exposure history, disease severity indicators, outcomes, clinical data, laboratory diagnostic test results, and comorbidities of COVID-19 cases. These data elements are collected by jurisdictions and voluntarily shared with the CDC for surveillance purposes. Updates to the dataset are expected weekly, and it is provided under the CC0 Public Domain license, indicating it can be freely used and shared. Researchers and public health officials can utilize this dataset for various analyses, including demographic trends of COVID-19 cases and deaths. The dataset is well-documented and maintained by the CDC, providing reliable and clean data for research and public health applications..

2. Limitations

A. The Number of columns, features, are not enough to make a high-accuracy prediction for Covid-19 prediction. There are only 11 features with only binary values which will make prediction even harder.
B. There is no classification feature, which would be super difficult to predict the exact severity of the Covid-19 case. The prediction will be only if the patient is a Covid-19 carrier or not.

3. #Instances

There are 1,048,575 instances Which is a very good number to train our model on. However, the number of features is still a problem.

4. #Features

There are 11 unique features. Examples of these features are

A. Sex: might be a very effective feature to predict Covid-severity.

B. Race and ethnicity: would be super important because of the physical and chronological natures of different races than others which would help in predicting specific severity rate for a specific person.

C. ICU: indicates whether the patient was admitted to an intensive care unit or not.

5. Label

A "Yes-No" prediction on the patient if a COVID-19 carrier or not, which could be implicitly deducted.

PROJECT INTENTIONS

Our project aims to predict how severe the coronavirus is on a COVID-19 patient based on their medical records -- since research has shown the relationship between the coronavirus and other diseases, especially those related to the respiratory and cardiovascular systems. Our project will categorize the riskiness of COVID-19 for these patients into one of 5 degrees ranging from mild to severe.

ATTEMPTED SOLUTIONS

Our attempted solutions are going to be an extension of what has been done in prior literature reviews mentioned in this report. This is a classification problem for which we plan to develop different machine-learning models and analyze them to find the most accurate model for the dataset given. For example, the KNN algorithm would be one of the models explored to solve our classification problem. We will train the various models based on the following 20 features:

1. Sex
2. Age
3. Patient Care Type(Returned Home or hospitalized)
4. Pneumonia: Presence of air sacs inflammation or not
5. Pregnancy: Current pregnancy status of the patient
6. Diabetes
7. Chronic Obstructive Pulmonary Disease(COPD)
8. Asthma: Current presence of Asthma
9. Inmsupr: Current presence of immunosuppression
10. Hypertension: Current presence of hypertension
11. Cardiovascular: Current presence of heart or blood vessel-related disease
12. Renal Chronic: Current presence of chronic renal disease
13. Obesity
14. Tobacco: its use status by the patient
15. Other Disease: Presence of other disease or not
16. USMR: Type of treated medical unit
17. Medical Unit: Type of institution providing care
18. Intubed: Connection to a ventilator
19. ICU: Admission to an Intensive Care Unit
20. Date Died: Patient's death date if not alive

Most of these features are binary, which would ease the preprocessing. However, we would need to deal with NULL values in a good manner.

APPLICATIONS

Applications for this model could greatly impact the healthcare system in detecting high-risk COVID-19 patients who may require intensive care or ventilators, for example. Furthermore, this predictive model could analyze medical records to identify population groups at a higher risk of COVID-19, which could help healthcare organizations in taking care of these groups and accommodate their health.

REFERENCES

[1] Abdel-Fattah Sayed, Safynaz, et al. "IEEE Xplore." *Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity*, IEEE, ieeexplore.ieee.org/Xplore/home.jsp. Accessed 18 Feb. 2024.

[2] Nizri, Meir. “Covid-19 Dataset.” *Kaggle*, 13 Nov. 2022, www.kaggle.com/datasets/meirnizri/covid19-dataset.

[3] Möbius. “Covid-19 Case Surveillance Public Use Dataset.” *Kaggle*, 21 Dec. 2020, www.kaggle.com/datasets/arashnic/covid19-case-surveillance-public-use-dataset.