Analysis of COVID-19 Spread using Random Forest Classifier

Sakshi Mhatre
Department of Computer Science
Virginia Tech
Falls Church, VA
sakshimm@vt.edu

Sai Chandan Reddy Koncha
Department of Computer Science
Virginia Tech
Falls Church, VA
saichandanreddyk@vt.edu

Mahita Selvaraj
Department of Electrical and Computer Engineering
Virginia Tech
Falls Church, VA
smahita22@vt.edu

Aniruddha Hore
Department of Computer Science
Virginia Tech
Falls Church, VA
aniruddhah@vt.edu

Abstract

The COVID-19 pandemic has significantly impacted global health and socio-economic systems, prompting the need for advanced analytical tools to understand the patterns and dynamics of its spread. This project employs Random Forest algorithm to explore distinctive groupings and patterns in the transmission of COVID-19 across various regions or countries. By analyzing epidemiological data, the study aims to uncover hidden structures and relationships that may provide valuable insights into the factors influencing the virus's dissemination. The results of this analysis could contribute to more effective public health interventions and targeted strategies for controlling the spread of the virus.

Keywords: COVID-19, Random Forest Classifier, Epidemiology, Machine Learning, Public Health, SARS-CoV-2, Epidemiological Data, Spread Patterns, Socio-economic Impact, Geographical Entities.

ACM Reference Format:
Sakshi Mhatre, Sai Chandan Reddy Koncha, Mahita Selvaraj, and Aniruddha Hore. 2024. Analysis of COVID-19 Spread using Random Forest Classifier. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The emergence of the novel coronavirus, SARS-CoV-2, and the ensuing COVID-19 pandemic have placed unprecedented demands on global health systems and prompted a crucial need for innovative analytical methodologies. In response to the ongoing efforts to control and understand the spread of the virus, this project adopts advanced machine learning techniques, specifically the Random Forest Classifier, to unravel intricate patterns in the transmission dynamics of COVID-19.

This project concentrates on predictive modeling using a Random Forest Classifier. By leveraging this ensemble learning algorithm, we aim to discern essential features influencing the likelihood of individuals contracting COVID-19 based on symptoms and demographic factors. The Random Forest Classifier, known for its adaptability to complex datasets and robust predictive capabilities, provides a valuable tool for identifying and interpreting patterns in epidemiological data.

This study contributes to the broader discourse on employing machine learning in public health by demonstrating the effectiveness of the Random Forest Classifier in predicting COVID-19 cases. Through a meticulous analysis of symptoms and demographic information, our approach not only enhances diagnostic accuracy but also provides valuable insights for healthcare professionals and policymakers. In the face of the global health crisis, the integration of machine learning applications into public health strategies emerges as a crucial component in the fight against the COVID-19 pandemic

2 Related Research

2.1 Machine Learning in Epidemiology

Machine learning (ML) is rapidly transforming the field of epidemiology, the study of diseases and their distribution across populations. This powerful tool offers exciting opportunities to enhance our understanding of disease patterns, predict outbreaks, and develop effective interventions [3]. ML algorithms can handle large and complex datasets, identifying subtle patterns and relationships that might be missed

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

by traditional methods. ML can also predict the future course of an outbreak, allowing for better resource allocation and preparedness measures [7]. It can identify individuals at high risk of contracting a disease, enabling targeted interventions and prevention strategies and analyzing data from various sources, like social media and electronic health records, to detect early signs of outbreaks [1].

2.2 Demographic Factors and Disease Spread

2.2.1 Age. Increased susceptibility: Younger and older individuals tend to be more susceptible to certain diseases due to weaker immune systems. Transmission patterns: Children often play a vital role in transmitting respiratory illnesses due to their close contact with others. Differential disease severity: Age can also affect the severity of illness, with older populations often experiencing more severe complications [8].

2.2.2 Sex. Biological factors: Sex differences in hormonal and immune responses can influence susceptibility and disease progression. Occupational differences: Gender-specific roles and occupations can lead to different exposure risks and behaviors. Healthcare access: Access to preventive care and treatment may vary based on sex and social norms [6].

2.2.3 Race/Ethnicity. Social determinants of health: Factors like poverty, housing conditions, and access to healthcare can contribute to racial disparities in disease burden. Cultural practices: Cultural beliefs and practices can influence exposure risks and help-seeking behaviors. Genetic variations: Genetic predisposition to certain diseases may vary across different racial and ethnic groups [4].

2.2.4 Socioeconomic Status. Living conditions: Crowded housing and limited access to sanitation can increase transmission risks. Occupational exposures: Certain jobs may involve higher exposure to disease-causing agents. Access to healthcare: Individuals with lower socioeconomic status often face barriers to accessing preventive care and timely treatment [5].

2.3 Ensemble Learning in Healthcare

Ensemble learning has emerged as a powerful tool in healthcare, offering significant potential to improve the accuracy and reliability of diagnoses, predictions, and treatment decisions. This approach combines multiple models, leveraging their collective strengths to overcome individual weaknesses and achieve superior performance. the benefits of ensemble learning can be improved prediction accuracy, reduced risk of overfitting, enhanced robustness, improved interpretability.

3 Methodology

3.1 Data Collection

Obtained a dataset containing information on COVID-19 cases, deaths, recoveries, testing rates, and other relevant metrics for different regions or countries.

3.2 Data Preprocessing

The obtained dataset was preprocessed by cleaning, handling the missing values and normalizing the numerical features while selecting the appropriate features for analysis using random forest classifier.

3.3 Exploratory Data Analysis

Conducted exploratory data analysis to understand the distribution of COVID-19 metrics and identified potential trends or outliers.

3.4 Feature Scaling

Categorical variables are encoded, and necessary transformations are applied to prepare the data for machine learning models. Additionally, demographic features such as 'Sex' and 'Known-contact' are appropriately mapped to numerical values for compatibility

3.5 Model Selection and Training

Random Forest classification algorithm was employed to train our predictive model. The dataset is split into training and testing sets, and hyper parameter tuning was performed using GridSearchCV to enhance the model's performance [2].

3.6 Visualization

The performance of the Random Forest model was evaluated using various metrics, including accuracy, precision, recall, F1 score, and the area under the Receiver Operating Characteristic (ROC) curve. Confusion matrices provide insights into the model's ability to correctly classify COVID-19 cases.

Cough symptoms: A significant number of individuals who tested negative also reported no cough symptoms. However, among those who reported having a cough, the number of positive cases is relatively higher.

The following are what we can deduce from the visualized plots.

Fever: Most individuals did not report having a fever. Among those who did, the number of positive cases is higher than negative cases. Sore throat: The majority did not experience this symptom. Shortness of breath: This symptom was also not common among the individuals. Headache: A notable number of individuals reported not having headaches. Age 60 and above: A considerable number of individuals were below 60 years of age. Sex: The number of females and males in the dataset is almost evenly distributed. Known

contact: Most individuals did not have a known contact with a confirmed case.

4 Experiments

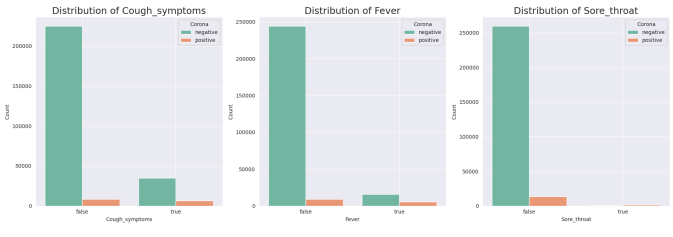

Figure 1. Cough Symptoms, Fever, Sore Throats

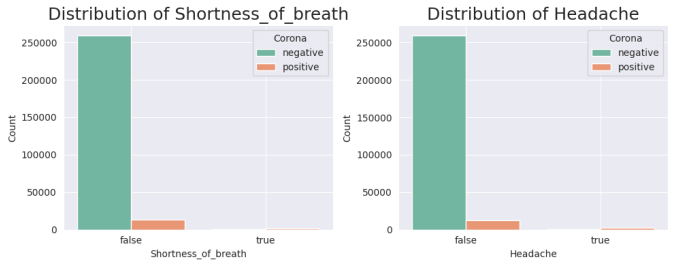

Figure 2. Distribution of Shortness of Breath and Headache

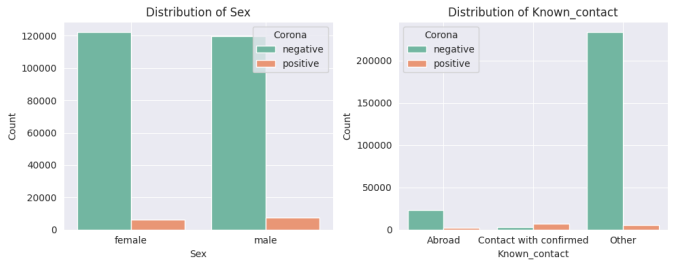

Figure 3. Distribution of Sex and Known Contact

5 Discussions

With hyper parameter training, the random forest classifier achieves a high accuracy (98.93), precision (98.92), and recall (97.68). This indicates that the model is very good at predicting COVID-19, with a low false positive rate and a low false negative rate. The F1 score (98.30) is also very high. This indicates that the model has a good balance between precision and recall and is very effective overall at predicting COVID-19.

The model has a very low false positive rate (1.08). This means that the model is very good at avoiding identifying healthy individuals as having COVID-19. The model has a very low false negative rate (2.32). This means that the model is very good at identifying individuals who have COVID-19, even if they have mild symptoms. The model is better at identifying individuals with COVID-19 who have had known contact with a confirmed case (98.33 recall). This suggests that the model is able to learn from the contact history information to better identify individuals who are at high risk of infection.

Confusion Matrix: [[77217 837] [1835 2522]]

Overall, the confusion matrix suggests that the random forest classifier is a very accurate and reliable tool for predicting COVID-19 diagnosis. It has a low false positive rate, a low false negative rate, and is better at identifying individuals with COVID-19 who have had known contact with a confirmed case.

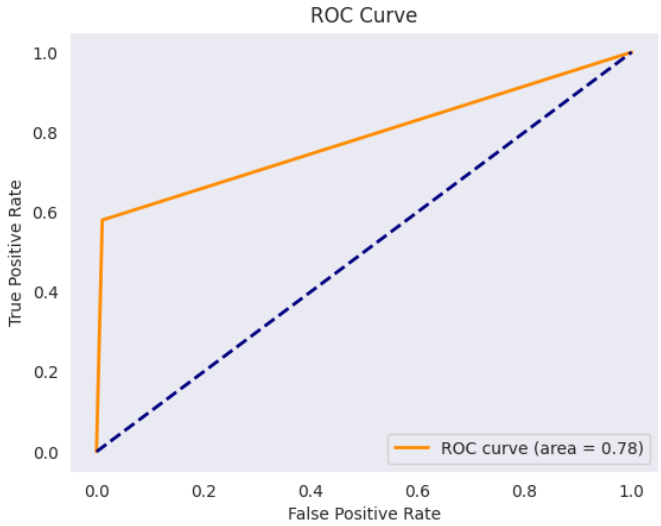

Figure 4. ROC Curve

6 Conclusion

It can be concluded that Cough, fever, and sore throat are the most prominent symptoms associated with COVID-19 diagnosis. These symptoms hold the highest importance in predicting the disease, suggesting their crucial role in identification and screening efforts. Shortness of breath and headache also contribute significantly to the prediction model, although to a lesser extent. These features provide additional information and enhance the model's accuracy. Known contact with a confirmed case acts as a moderately important predictor. This highlights the importance of contact tracing and isolation measures in mitigating the spread of the virus.

The overall analysis suggests that the random forest classifier is a powerful tool for predicting COVID-19 based on clinical symptoms and contact history. Its high accuracy, precision, recall, and ROC demonstrate its effectiveness in identifying cases and aiding in disease control efforts. Further research and potential integration into healthcare systems

can leverage this model's strengths to improve diagnosis, resource allocation, and ultimately, public health outcomes.

7 Author Contributions

Code: Sakshi Mhatre
Code: Sai Chandan Reddy Koncha
Report: Mahita Selvaraj
Report: Anirudhha Hore
link to code: https://github.com/Sakshi0700/Analysis-of-COVID-19-Spread-using-Random-Forest-Classifier

References

[1] Daniel M. Bean, Zeljko Kraljevic, Thomas Searle, Rebecca Bendayan, Anthony Pickles, Amos Folarin, et al. Treatment with ace-inhibitors is associated with less severe disease with sars-covid-19 infection in a multi-site uk acute hospital trust. 2020. Preprint. https://doi.org/10.1101/2020.04.07.20056788.

[2] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[3] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 4:170127, 2017.

[4] Shruti Magesh, Daniel John, Wei Tse Li, Yuxiang Li, Aidan Mattingly-App, Sharad Jain, Eric Y Chang, and Weg M Ongkeko. Disparities in covid-19 outcomes by race, ethnicity, and socioeconomic status: a systematic review and meta-analysis. *JAMA network open*, 4(11):e2134147–e2134147, 2021.

[5] Gonzalo E Mena, Pamela P Martinez, Ayesha S Mahmud, Pablo A Marquet, Caroline O Buckee, and Mauricio Santillana. Socioeconomic status determines covid-19 incidence and related mortality in santiago, chile. *Science*, 372(6545):eabg5298, 2021.

[6] Mateus D Pinheiro, Anderson P Rodrigues, and Fernando M Coelho. Gender differences in covid-19: cardiovascular risk factors, inflammation, and immunity. *Journal of Clinical Medicine Research*, 7(12):993–1004, 2015.

[7] Atul Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nigam Hajaj, Moritz Hardt, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[8] Z. Shahid, R. Kalayanamitra, B. McClafferty, D. Kepko, D. Ramgobin, R. Patel, and R. Jain. Covid-19 and older adults: what we know. *Journal of the American Geriatrics Society*, 69(3):491–500, 2021.