COVID 19 ANALYSIS

ABSTRACT

The coronavirus is considered this century's most disruptive catastrophe and global concern. This disease has prompted extreme social, psychological and economic impacts affecting millions of people around the globe. COVID-19 is transmitted from one infected person's body to another through respiratory droplets. This virus proliferates when people breathe in air-contaminated space with droplets and microscopic airborne particles. This research aims to analyze automatic COVID-19 detection using machine learning techniques to build an intelligent web application. The dataset has been preprocessed by dropping null values, feature engineering, and synthetic oversampling (SMOTE) techniques. Next, we trained and evaluated different classifiers, i.e., logistic regression, random forest, decision tree, k-nearest neighbor, support vector machine (SVM), ensemble models (adaptive boosting and extreme gradient boosting) and deep learning (artificial neural network, convolutional neural network and long short-term memory) techniques. Explainable AI with the LIME framework has been applied to interpret the prediction results. The hybrid CNN-LSTM algorithm with the SMOTE approach performed better than the other models on the employed open-source dataset obtained from the Israeli Ministry of Health website, with 96.34% accuracy and a 0.98 F1 score. Finally, this model was chosen to deploy the proposed prediction system to a website, where users may acquire an instantaneous COVID-19 prognosis based on their symptoms.

INTRODUCTION

Coronaviruses are a group of diverse viruses with a wide range of variants that differ in several ways (Bo et al., 2021). In 1965, scientists discovered the first strain of the coronavirus that infected humans. This strain caused common colds in the host body. After more than a decade, the researchers discovered a group of viruses found in humans and animals named after their appearance, which resembled a crown. Hence, the word ‘corona’ in the name of the virus derives from a Latin word that means ‘crown.’ So far, scientists contend that up to seven coronaviruses can infect people. One such severe acute respiratory syndrome-related virus was discovered for the first time in 2003 in southern China, and it spread rapidly in approximately 30 different countries. Specialists confirmed that SARS-CoV-2 originated in bats. In the wet market of Wuhan, people would visit to purchase fish and fresh meat as the animals were slaughtered in the same place. It is believed that this is where the contamination spread to humans. The congested and

crowded environment is prone to facilitating cross-contamination and the swapping of genes between various animals, which may have resulted in viruses undergoing significant mutations, potentially infecting humans and propagating the infection in a rapid and devastating manner.
identify the presence of COVID-19 in individual patients. The SVM algorithm achieved the highest level of accuracy at 98.38%, but the XGBoost model was determined to be the best model since it had a higher recall value (99.26%) without significantly influencing the other evaluation criteria (accuracy level of 97.71%). Tiwari et al. (Tiwari, Bhati, Al-Turjman & Nagpal, 2022) examined the coronavirus infection trend, treatment and mortality rates by employing classical machine learning techniques. The authors discovered that the naive Bayes framework forecasted the COVID-19 disease with the highest accuracy. Rai and coauthors (Rai et al., 2022) predicted the death rate of COVID-19 patients employing majority rule-based ensemble techniques. Multivariate imputation, synthetic oversampling and feature selection approaches were used in this work. The XGBoost model attained the best accuracy and F1 coefficient of 86.9% and 71.6%, respectively.

Some authors have tried to predict coronavirus inspection by using both machine learning and deep learning techniques. The use of the occlusion technique in COVID-19 detection was discussed in Udawat, Santani and Agrawal (2021). A content-adaptive progressive occlusion analysis (CAPAO) algorithm was used to perform the analysis. Accuracies ranging between 78.33% and 98.33%. By using a combination of a wrapper feature selection technique and machine learning and deep learning classifiers, Turabieh et al. (H. Turabieh & Karaa, 2021) initiated the automatic forecasting of COVID-19. CNN with the BGA approach achieved an 80% accuracy in predicting COVID-19. Cobre and researchers (Cobre et al., 2021) anticipated coronavirus disease positivity and severity by applying various machine learning and neural network models. The authors employed artificial neural network, DT and KNN models to accurately categorize negative and positive instances with more than 84% accuracy

Proposed system.

The open-source dataset used in this work is acquired from the Department of Health, Israel (Zoabi, Deri-Rozov & Shomron, 2021). The dataset contains individual results of different symptoms, basic information about the patients, and the COVID-19 test results of 2742,596 patients. There are three types of information available in the dataset. It contains basic information about the patient, i.e., the test date, gender, and if the patient's age is over 60, indicators that denote the symptoms of the patient, including information on five symptoms in the dataset, i.e., cough, fever, shortness of breath, sore throat, and headache and the COVID-19 test result and whether the patient recently came into contact with a COVID-19 patient a...

COVID-19 test results in terms of the features in the dataset.
the dataset is highly imbalanced as there is a higher number of negative cases and a significantly smaller number of positive cases with a ratio of 9.3:1.0. The linear dependence of various features of the used dataset has been measured with the Pearson .
Correlation of various features of the used dataset.
COVID-19 data collection involved gathering various types of information to track and understand the pandemic. Key aspects of data collection included:

INOVATION

1. Case Data: Collecting data on the number of confirmed COVID-19 cases, including information on age, gender, and location of infected individuals.

2. Testing Data: Recording data on the number of tests conducted, test results, and testing locations.

3. Hospitalization Data: Tracking the number of COVID-19 patients admitted to hospitals, their conditions, and the availability of hospital resources.

4. Mortality Data: Recording data on COVID-19-related deaths, including age, gender, and underlying conditions.

5. Contact Tracing Data: Collecting information about individuals who may have been exposed to the virus, which is critical for containment.

6. Vaccination Data: Monitoring the number of people vaccinated, vaccine types, and vaccination locations.

7. Genomic Sequencing: Conducting genomic sequencing of the virus to monitor mutations and variations.

8. Public Health Measures: Gathering data on the implementation and effectiveness of measures like lockdowns, mask mandates, and social distancing.

9. Data Sources: Data was collected from various sources, including healthcare providers, testing centers, contact tracing apps, and government agencies.

10. Reporting and Visualization: Data was often reported through dashboards, reports, and visualizations to help inform the public and guide decision-making.

Effective data collection and analysis were crucial in managing the COVID-19 pandemic and making informed decisions to mitigate its impact.

CAUSES

The analysis of the causes and contributing factors to the spread of COVID-19 involves various factors. While the primary cause of the pandemic is the introduction of the novel coronavirus (SARS-CoV-2) into the human population, several factors have contributed to its rapid transmission and impact:

1. Human-to-Human Transmission: COVID-19 primarily spreads through respiratory droplets when an infected person coughs, sneezes, or talks. Close contact with infected individuals is a major factor in transmission.

2. Asymptomatic and Presymptomatic Spread: People who are infected with the virus but show no symptoms (asymptomatic) or who have not yet developed symptoms (presymptomatic) can unknowingly spread the virus.

3. Variants: The emergence of new variants of the virus has impacted its transmission and potential resistance to immunity developed through vaccination or prior infection.

4. Global Travel: International travel allowed the virus to spread rapidly from its original epicenter in Wuhan, China, to other parts of the world.

5. Public Health Measures: The effectiveness of public health measures, such as mask-wearing, social distancing, and lockdowns, in curbing the virus's spread varies depending on adherence and enforcement.

6. Healthcare Infrastructure: The readiness and capacity of healthcare systems have played a role in managing and treating COVID-19 cases.

7. Vaccination Rates: The availability and distribution of COVID-19 vaccines have a direct impact on the spread and severity of the virus.

8. Public Behavior: Individual behavior, compliance with guidelines, and vaccine acceptance influence the spread of the virus.

9. Variability in Government Responses: The response to the pandemic, including testing, contact tracing, and quarantine measures, has varied between countries and regions.

It's important to note that the situation and causes of COVID-19 continue to evolve as more is learned about the virus and its variants. Analysis of these causes helps inform public health strategies to control the spread of the disease.

IMPACT

The impact of COVID-19 analysis has been profound and has influenced various aspects of society and healthcare. Here are some key impacts:

1. Public Health Response: Analysis of COVID-19 data has informed public health measures, such as social distancing, mask mandates, and lockdowns, to curb the spread of the virus and protect public health.

2. Vaccination Strategy: Ongoing analysis of vaccine efficacy and distribution data has shaped vaccination strategies, ensuring that vaccines are distributed to those at highest risk first.

3. Resource Allocation: Data analysis has guided the allocation of healthcare resources, including ventilators, ICU beds, and medical supplies, to regions with high case numbers.

4. Economic Impact: The analysis of COVID-19's economic impact has influenced government stimulus packages and financial relief measures to support businesses and individuals affected by the pandemic.

5. Travel and Border Policies: Data analysis has been crucial in determining travel restrictions, quarantine requirements, and

border policies to prevent the spread of the virus across regions and countries.
6. Remote Work and Education: The analysis of case data has led to the widespread adoption of remote work and online education to reduce the risk of transmission.

7. Mental Health and Well-being: Data analysis has highlighted the mental health impacts of the pandemic, leading to increased focus on mental health support and services.

8. Research and Vaccine Development: Analysis of the virus's genetic sequence has informed vaccine development efforts, leading to the rapid development and distribution of COVID-19 vaccines.

9. Scientific Collaboration: The pandemic prompted extensive global scientific collaboration, accelerating research and data sharing.

10. Future Pandemic Preparedness: COVID-19 analysis has underscored the importance of pandemic preparedness and the need for improved surveillance, early warning systems, and research.

The impact of COVID-19 analysis is ongoing and has been central to the global response to the pandemic, helping to adapt strategies and mitigate its effects.

VACCINATION

Analyzing COVID-19 vaccination efforts is crucial in understanding and managing the pandemic. Here are key aspects of COVID-19 vaccination analysis:

1. Vaccine Efficacy: Ongoing analysis assesses the effectiveness of different COVID-19 vaccines in preventing infection, severe disease, and transmission.

2. Vaccine Coverage: Monitoring the percentage of the population that has received one or both vaccine doses and the impact of vaccination on community immunity.

3. Vaccine Distribution: Analyzing the distribution of vaccines to ensure equitable access and identifying underserved or at-risk populations.

4. Booster Shots: Assessing the need for booster doses and analyzing their effectiveness in maintaining immunity.

5. Vaccine Hesitancy: Identifying factors contributing to vaccine hesitancy and strategies to increase vaccine acceptance.

6. Adverse Events: Monitoring and analyzing adverse events associated with vaccination to ensure safety.

7. Variants: Analyzing the impact of emerging variants on vaccine effectiveness and the need for updated vaccines.

8. Global Impact: Evaluating the distribution of vaccines worldwide and their role in reducing the global spread of the virus.

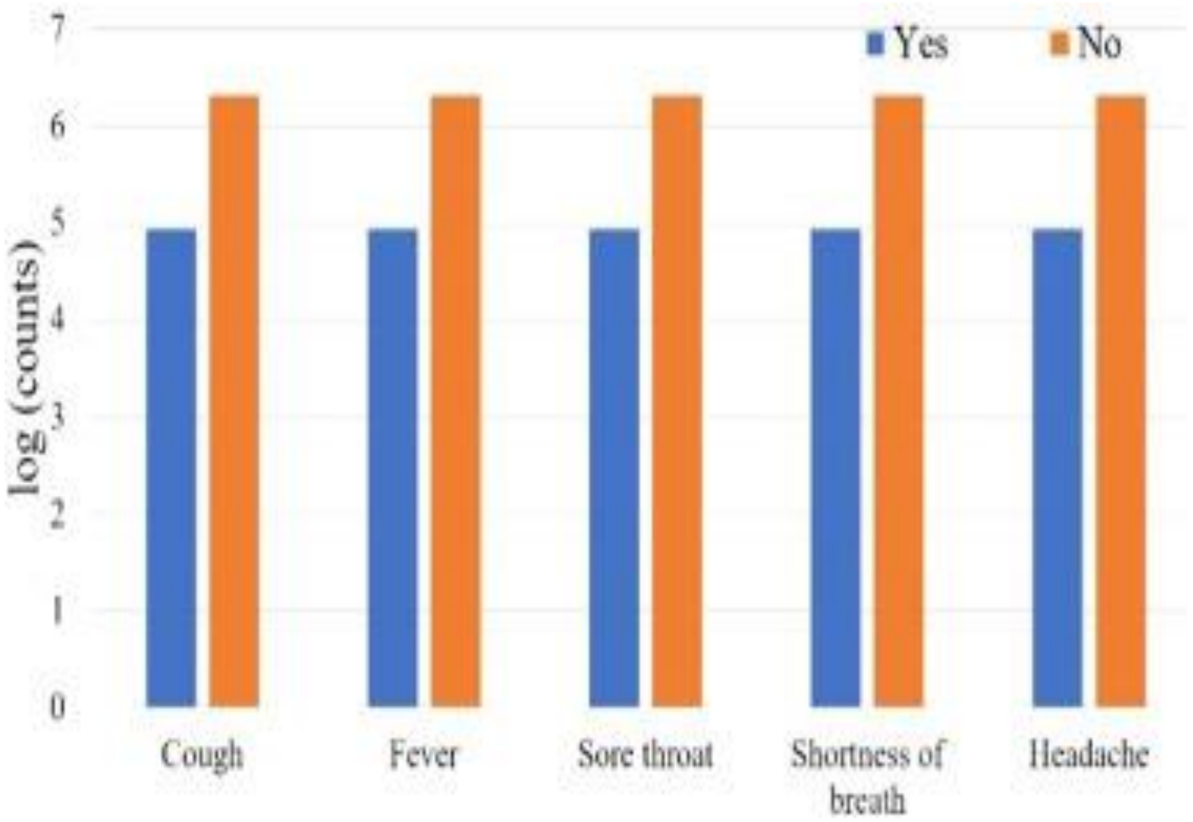
9. Public Health Measures: Analyzing how vaccination impacts the need for other public health measures, such as mask-wearing and social distancing.

10. Vaccine Equity: Ensuring fair distribution of vaccines to low- and middle-income countries and addressing global vaccine inequality.

11. Data Transparency: Maintaining transparency in vaccine data reporting to build trust and confidence in vaccination programs.

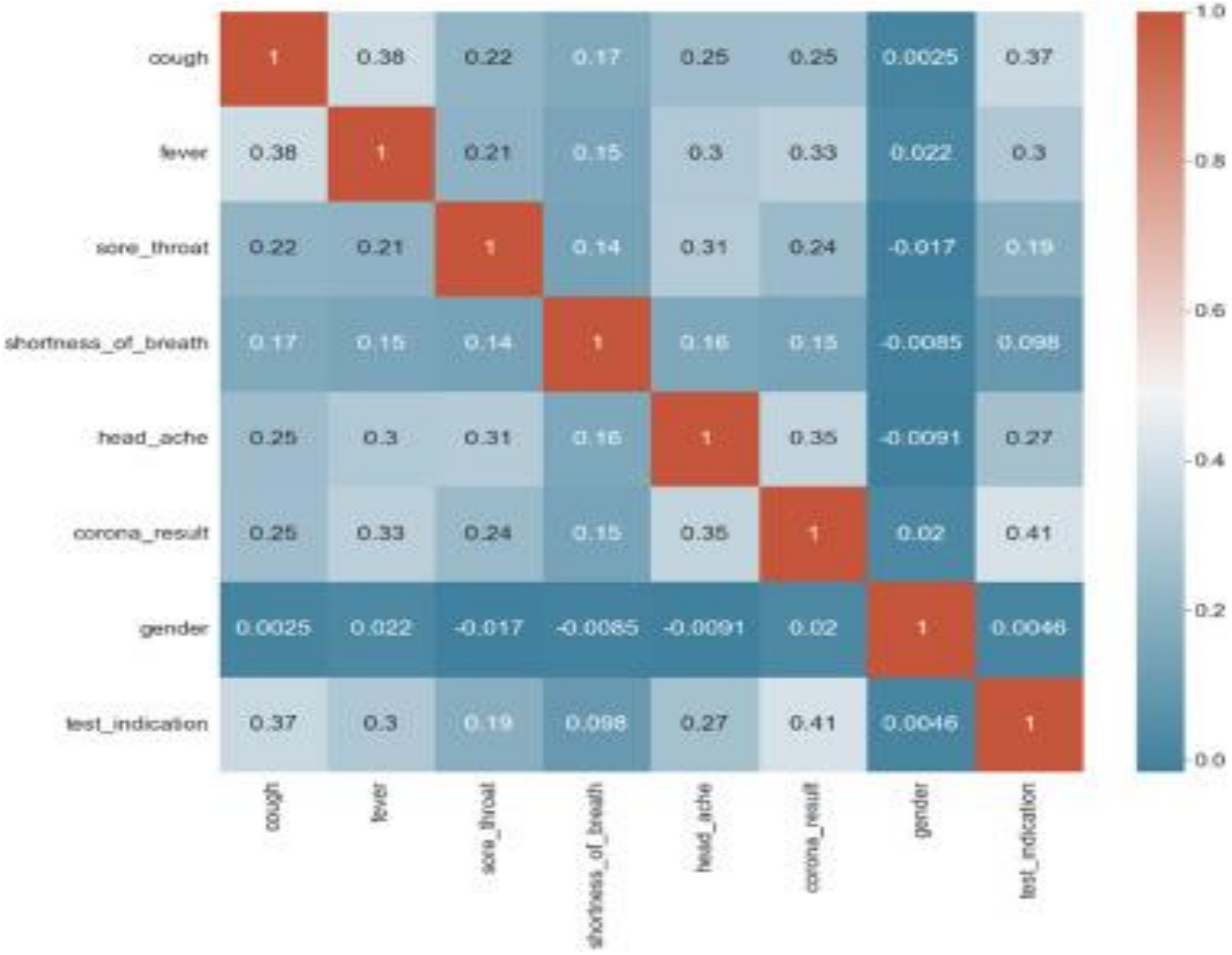
COVID-19 vaccination analysis is essential for adapting vaccination strategies, improving vaccine distribution, and ensuring the world's path to recovery from the pandemic.

BAR-CHART


7
6
5
4
3
2
1
0
log (counts)
Yes
No
Cough
Fever
Sore throat
Shortness of breath
Headache

COVID-19 test results in terms of the features in the dataset.

The dataset is highly imbalanced as there is a higher number of negative cases and a significantly smaller number of positive cases with a ratio of 9.3:1.0. The linear dependence of various features of the used dataset has been measured with the Pearson correlation index


cough 1 0.38 0.22 0.17 0.25 0.25 0.0025 0.37
fever 0.38 1 0.21 0.15 0.3 0.33 0.022 0.3
sore_throat 0.22 0.21 1 0.14 0.31 0.24 -0.017 0.19
shortness_of_breath 0.17 0.15 0.14 1 0.16 0.15 -0.0085 0.098
head_ache 0.25 0.3 0.31 0.16 1 0.35 -0.0091 0.27
corona_result 0.25 0.33 0.24 0.15 0.35 1 0.02 0.41
gender 0.0025 0.022 -0.017 -0.0085 -0.0091 0.02 1 0.0046
test_indication 0.37 0.3 0.19 0.098 0.27 0.41 0.0046 1
cough
fever
sore_throat
shortness_of_breath
head_ache
corona_result
gender
test_indication