A Review Study On OVID-19 Forecasting Using Machine Learning

Kamal Narayan Kamlesh[1*], Dr Kumar Vishal[2]

[1*]Research Scholar, Department of CS & IT, Magadh University Bodh Gaya, e-mail: kamal.kamleshjha@gmail.com
Orcid id: 0000-0001-8237-2905
[2]Assistant Professor, Department of Mathematics, Magadh University Bodh Gaya

***Corresponding Author:** Kamal Narayan Kamlesh
*Research Scholar, Department of CS & IT, Magadh University Bodh Gaya, e-mail: kamal.kamleshjha@gmail.com
Orcid id: 0000-0001-8237-2905

Abstract:
The COVID-19 pandemic has had a significant global impact, affecting public health, economies, and social structures. Accurate forecasting of the spread and severity of the disease has become crucial for effective decision-making and resource allocation. Machine learning techniques have emerged as powerful tools for COVID-19 forecasting due to their ability to analyze complex data patterns and make predictions. In this review paper, we provide an overview of the state-of-the-art machine learning approaches employed for COVID-19 forecasting, highlighting their strengths, limitations, and future directions. We discuss the different data sources used, feature engineering techniques, modelling strategies, and evaluation metrics employed in COVID-19 forecasting research. Additionally, we examine the challenges associated with COVID-19 forecasting, including data quality issues, model interpretability, and ethical considerations. We conclude by outlining potential areas for future research and emphasizing the importance of collaboration and data sharing to improve the accuracy and reliability of COVID-19 forecasting models.

Keywords: COVID-19, Machine Learning, Data, Model, Error, Forecasting

1. Introduction

COVID-19, also known as the coronavirus disease 2019, is a highly contagious respiratory illness caused by the novel coronavirus SARS-CoV-2. The outbreak of this disease was first identified in December 2019 in Wuhan, Hubei Province, China. Since then, COVID-19 has rapidly spread globally, leading to a pandemic(she et. al. (2020)).

The virus primarily spreads through respiratory droplets when an infected person coughs, sneezes, talks, or breathes. It can also spread by touching contaminated surfaces and then touching the face. The most common symptoms of COVID-19 include fever, cough, and difficulty breathing. However, some individuals may experience mild symptoms or remain asymptomatic, making it challenging to identify and contain the spread of the virus.

COVID-19 has had a significant impact on public health, economies, and societies worldwide. Governments and health organizations have implemented various measures to control the spread, such as social distancing, wearing face masks, frequent handwashing, and travel restrictions(Ciotti et.al. (2020)). Lockdowns and quarantine measures have been imposed in many countries to limit the transmission of the virus.

Efforts to combat COVID-19 have included the development and distribution of vaccines. Multiple vaccines have been authorized for emergency use or approved for full use in different countries, providing hope for controlling the pandemic. Vaccination campaigns aim to reduce the severity of the disease, protect vulnerable populations, and achieve herd immunity.

The pandemic has also prompted extensive research into treatments, diagnostics, and preventive measures. Scientific advancements and collaborations have played a crucial role in understanding the virus and developing strategies to mitigate its impact.

It's important to note that the situation regarding COVID-19 is constantly evolving. It is recommended to follow guidelines and updates from reliable health authorities such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) to stay informed about the latest developments and protective measures.

1.1 Background

COVID-19 is caused by a novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Coronaviruses are a large family of viruses that can cause illness in animals and humans. In the past, coronaviruses have caused outbreaks such as the severe acute respiratory syndrome (SARS) outbreak in 2002-2003 and the Middle East respiratory syndrome (MERS) outbreak in 2012.

The first cases of COVID-19 were reported in December 2019 in Wuhan, Hubei Province, China. It is believed that the virus originated in bats and may have been transmitted to humans through an intermediate animal host, although the exact origins are still under investigation. The initial outbreak was linked to a seafood market in Wuhan, which also sold live wild animals.

The virus quickly spread from person to person, leading to a global pandemic. It is highly contagious and can be transmitted through respiratory droplets when an infected person coughs, sneezes, talks, or breathes. It can also spread by touching surfaces contaminated with the virus and then touching the face.

The World Health Organization (WHO) declared COVID-19 a Public Health Emergency of International Concern on January 30, 2020, and later characterized it as a pandemic on March 11, 2020. The rapid spread of the virus and its severe impact on public health systems, economies, and daily life led to widespread concern and the implementation of various measures to control its transmission (verma et.al.(2020)).

Governments and health organizations worldwide have implemented strategies such as social distancing, mask-wearing, hand hygiene, testing, contact tracing, and quarantine measures to slow the spread of the virus. Lockdowns and travel restrictions have been imposed in many countries to mitigate the impact of the pandemic.

The global scientific community has been working tirelessly to understand the virus, develop diagnostic tests, treatments, and vaccines, and share information and resources. Multiple vaccines have been authorized for emergency use or approved for full use, and vaccination campaigns are underway in many countries (Pokhrel et.al.(2021)).

The COVID-19 pandemic has had profound impacts on public health, the global economy, education, travel, and social interactions. It has highlighted the importance of global collaboration, preparedness, and resilience in the face of emerging infectious diseases. Efforts to control the spread of the virus and mitigate its impact are ongoing as the situation continues to evolve.

1.2 Motivation

There are several motivations for conducting a review study on COVID-19 forecasting:

- **Understanding the accuracy of forecasting models:** Reviewing the existing literature on COVID-19 forecasting models allows researchers to assess the accuracy and reliability of different approaches. This information helps in identifying the strengths and limitations of various models, providing insights into their performance under different conditions.
- **Improving future forecasting efforts:** By examining the methodologies and techniques used in COVID-19 forecasting studies, researchers can identify areas for improvement. This includes understanding the data sources, variables, and modeling techniques employed in successful forecasting models, as well as identifying any gaps or shortcomings that need to be addressed in future research.
- **Informing policy and decision-making:** Accurate and reliable forecasting plays a crucial role in informing public health policies, resource allocation, and decision-making during a pandemic. A comprehensive review of COVID-19 forecasting studies can provide policymakers with valuable insights into the effectiveness of different forecasting approaches, helping them make informed decisions and develop appropriate strategies to manage the pandemic.
- **Assessing the impact of interventions:** Forecasting models can be used to assess the impact of interventions, such as social distancing measures, travel restrictions, and vaccination campaigns. Reviewing studies that have utilized forecasting models can provide insights into the effectiveness of these interventions in controlling the spread of COVID-19 and reducing the burden on healthcare systems.
- **Identifying research gaps and future directions:** Reviewing the existing literature on COVID-19 forecasting can help identify research gaps and areas that require further investigation. This can guide researchers in designing future studies, exploring new forecasting techniques, and addressing specific challenges or limitations associated with COVID-19 forecasting.

Overall, conducting a review study on COVID-19 forecasting is important for advancing our understanding of the pandemic, improving forecasting accuracy, informing policy decisions, and guiding future research efforts in managing and mitigating the impact of COVID-19 (Singh et.at. (2020)).

1.3 Objectives:

The objective of conducting a review study on COVID-19 forecasting can vary depending on the specific research goals and context. Here are some common objectives:

- **Evaluate the accuracy of existing forecasting models:** One objective is to assess the performance of different COVID-19 forecasting models. This involves examining the accuracy of predictions made by various models and comparing their results against actual observed data. By evaluating the accuracy, researchers can identify the strengths and weaknesses of different forecasting approaches.
- **Identify the factors influencing forecasting accuracy:** Another objective is to identify the key factors that affect the accuracy of COVID-19 forecasting models. This can include factors such as data quality, model assumptions, input variables, modeling techniques, and the timing of interventions. Understanding these factors helps in improving the design and implementation of future forecasting models.
- **Compare different forecasting methodologies:** The objective might be to compare and contrast various methodologies used in COVID-19 forecasting. This includes examining different modeling techniques, such as statistical models, machine learning algorithms, and simulation models. By comparing these methodologies, researchers can identify the advantages and limitations of each approach.
- **Assess the impact of interventions:** COVID-19 forecasting models can be used to evaluate the impact of interventions and control measures on the spread of the virus. The objective might be to review studies that have used forecasting models to assess the effectiveness of interventions such as lockdowns, social distancing, mask-wearing, and vaccination campaigns. This provides insights into the effectiveness of these measures in controlling the pandemic.
- **Identify gaps and future research directions:** An objective might be to identify gaps in the existing literature and suggest future research directions. This involves assessing the limitations of current forecasting models, highlighting areas that require further investigation, and proposing innovative approaches or methodologies to enhance COVID-19 forecasting accuracy and utility.

By accomplishing these objectives (hamzah et.al.(2020)), a review study on COVID-19 forecasting contributes to the scientific understanding of pandemic dynamics, informs decision-makers, and guides future research efforts to improve forecasting capabilities, ultimately aiding in effective pandemic management and response.

2. COVID-19 Data Sources:

There are several reliable sources for obtaining COVID-19 data. Here are some commonly used data sources for COVID-19:

- **World Health Organization (WHO):** The WHO provides global COVID-19 data, including case counts, deaths, testing rates, and other relevant information. Their website and publications offer comprehensive and up-to-date data from various countries around the world.

- **Centers for Disease Control and Prevention (CDC):** The CDC, based in the United States, provides COVID-19 data for the country, including case counts, hospitalizations, deaths, and vaccination data. They offer a range of resources and interactive tools to access and analyze the data.
- **National Health Authorities:** Many countries have their own national health authorities that provide COVID-19 data specific to their regions. Examples include the Public Health Agency of Canada, Public Health England, and the European Centre for Disease Prevention and Control (ECDC). These sources often provide detailed information on cases, testing, hospitalizations, and other relevant metrics.
- **Johns Hopkins University (JHU):** JHU has been maintaining a widely used COVID-19 dashboard that provides global and country-specific data on cases, deaths, recoveries, testing, and vaccination rates. The data is sourced from multiple official sources, including government health departments and international organizations.
- **Our World in Data:** This online publication collates and presents COVID-19 data from various sources, including official health authorities. It offers a comprehensive database with visualizations, statistics, and downloadable datasets for global and country-level COVID-19 data.
- **COVID-19 Tracking Projects:** Some countries or regions have dedicated projects or websites that track and publish COVID-19 data. For example, the COVID Tracking Project in the United States and the COVID Data Tracker in India provide detailed and up-to-date information on cases, testing, hospitalizations, and other relevant metrics.
- **Academic Research Databases:** Academic research databases such as PubMed, medRxiv, and bioRxiv often publish research articles and preprints related to COVID-19. These sources can provide valuable insights and data on various aspects of the pandemic, including epidemiology, clinical findings, and modeling studies.

It is important to note that data from different sources may vary slightly due to variations in reporting methods, data collection processes, and updates. Therefore, it is recommended to consult multiple sources and consider the specific context and reliability of the data when analyzing and interpreting COVID-19 data.

2.1 Epidemiological Data

COVID-19 epidemiological data refers to the information related to the occurrence and distribution of COVID-19 cases within a population. This data provides insights into the spread and impact of the disease and is crucial for understanding its epidemiology (Xu et.al.(2020)). Here are some key epidemiological data points commonly tracked for COVID-19:

- **Case counts:** This includes the total number of confirmed COVID-19 cases within a specific population or geographic area. It indicates the overall burden of the disease and helps in monitoring its progression over time.
- **Deaths:** The number of deaths attributed to COVID-19 provides information on the severity and mortality rate of the disease. It helps in assessing the impact on public health and guiding intervention strategies.
- **Recoveries:** The number of individuals who have recovered from COVID-19 reflects the outcome of the disease and the effectiveness of treatments and healthcare interventions.
- **Incidence rate:** This represents the number of new cases of COVID-19 reported within a given time period, usually expressed as a rate per population size (e.g., per 100,000 people). It helps in understanding the rate of new infections and tracking changes over time.
- **Prevalence:** The prevalence of COVID-19 refers to the total number of existing cases within a population at a specific point in time. It provides an estimate of the proportion of the population currently affected by the disease.
- **Testing and positivity rates:** The number of COVID-19 tests conducted and the percentage of positive results help in assessing testing capacity, monitoring disease spread, and identifying areas with higher infection rates.
- **Demographic data:** Analyzing COVID-19 data by age, gender, ethnicity, and other demographic factors can reveal patterns of disease distribution and susceptibility among different population groups.
- **Hospitalizations:** Tracking the number of COVID-19 hospitalizations provides information on the burden placed on healthcare systems and helps in assessing the severity of the disease.
- **Transmission dynamics:** Data related to contact tracing, clusters, and secondary infections helps in understanding how the virus spreads within communities and informs control measures.

These epidemiological data points are typically collected and reported by health authorities, research institutions, and organizations involved in public health surveillance. They are crucial for monitoring the pandemic, informing public health policies, and assessing the effectiveness of interventions in controlling the spread of COVID-19.

2.2. Clinical Data

COVID-19 clinical data refers to the information related to the clinical characteristics, symptoms, outcomes, and treatment of individuals affected by the disease. This data provides insights into the clinical manifestations of COVID-19, its severity, and the effectiveness of different treatment approaches (Marshall et.al.(2020)). Here are some key clinical data points commonly tracked for COVID-19:

- **Symptoms:** COVID-19 presents with a range of symptoms, including fever, cough, shortness of breath, fatigue, muscle aches, sore throat, loss of taste or smell, and gastrointestinal symptoms. Tracking the prevalence and frequency of these symptoms helps in recognizing common clinical presentations and distinguishing COVID-19 from other respiratory illnesses.

- **Disease progression:** Clinical data tracks the progression of COVID-19 from mild to severe cases. It includes information on the development and duration of symptoms, the timeline of disease progression, and the factors associated with disease severity.
- **Severity markers:** Clinical data identifies markers of disease severity, such as respiratory distress, oxygen saturation levels, chest imaging findings (e.g., lung infiltrates), laboratory markers (e.g., elevated inflammatory markers, abnormal blood counts), and the need for intensive care or mechanical ventilation.
- **Comorbidities:** Clinical data highlights the presence of underlying health conditions, such as diabetes, hypertension, cardiovascular disease, and respiratory disorders, among individuals with COVID-19. Understanding the impact of comorbidities on disease severity and outcomes helps in risk stratification and guiding clinical management.
- **Complications:** COVID-19 can lead to various complications, including pneumonia, acute respiratory distress syndrome (ARDS), blood clotting disorders, myocardial injury, kidney injury, and neurological manifestations. Clinical data captures the occurrence and outcomes of these complications, aiding in understanding disease progression and informing treatment strategies.
- **Treatment approaches:** Clinical data tracks the effectiveness and safety of different treatment interventions, including antiviral drugs, immunomodulatory therapies, oxygen therapy, ventilatory support, and convalescent plasma. It helps in evaluating the efficacy of various treatment options and guiding clinical decision-making.
- **Outcomes:** Clinical data includes information on disease outcomes, such as recovery, hospitalization duration, need for intensive care, mortality rates, and long-term sequelae. It helps in assessing the overall impact of COVID-19 on individuals and populations.

Clinical data is typically collected through medical records, clinical trials, cohort studies, and surveillance systems. It plays a vital role in informing clinical management protocols, guiding public health strategies, and contributing to the global understanding of COVID-19's clinical characteristics and impact on individuals.

2.3. Socioeconomic Data

COVID-19 socioeconomic data refers to the information related to the social and economic impact of the COVID-19 pandemic on individuals, communities, and societies(Kumar et.al.(2020)). This data provides insights into the consequences of the pandemic on various aspects of people's lives, including employment, income, poverty, education, mental health, and access to essential services(Nikola et.al.(2020)). Here are some key socioeconomic data points commonly tracked for COVID-19:

- **Employment and labor market:** Socioeconomic data tracks the impact of COVID-19 on employment rates, job losses, furloughs, and changes in work patterns. It includes information on sectors heavily affected by the pandemic, such as hospitality, travel, and retail. It also assesses the resilience of different industries and the ability of individuals to maintain employment.
- **Income and poverty:** COVID-19 has led to income disruptions for many individuals and households. Socioeconomic data captures changes in income levels, poverty rates, and the prevalence of financial hardships. It helps in assessing the socioeconomic inequalities exacerbated by the pandemic and the effectiveness of government support measures.
- **Education and learning loss:** School closures and disruptions in education have been significant consequences of the pandemic. Socioeconomic data tracks the impact on access to education, remote learning capabilities, educational attainment, and the potential learning loss experienced by students. It highlights disparities in access to quality education and the need for targeted interventions.
- **Healthcare access:** COVID-19 has strained healthcare systems and affected access to healthcare services. Socioeconomic data examines changes in healthcare utilization, access to testing and treatment, and the impact on vulnerable populations, such as low-income individuals, uninsured individuals, and those with pre-existing health conditions.
- **Mental health and well-being:** The pandemic has had adverse effects on mental health and well-being. Socioeconomic data tracks indicators such as anxiety, depression, stress levels, and suicide rates. It helps in understanding the psychological impact of the pandemic and identifying populations at higher risk.
- **Social inequality:** COVID-19 has exacerbated existing social inequalities, with vulnerable populations disproportionately affected. Socioeconomic data examines disparities in healthcare outcomes, access to resources, housing stability, food security, and social support systems. It sheds light on the social determinants of health and highlights the need for targeted interventions to mitigate inequities.
- **Digital divide:** The pandemic has highlighted the importance of digital connectivity. Socioeconomic data assesses disparities in access to technology, internet connectivity, and digital literacy. It identifies populations that are disadvantaged by the digital divide, impacting remote work and learning, access to information, and healthcare services.

Socioeconomic data for COVID-19 is typically collected through surveys, administrative data, and research studies. It helps in understanding the broader impact of the pandemic, informing policymaking, and designing interventions to address the social and economic consequences of COVID-19 (Mishra et.al.(2020)).

2.4. Mobility Data

COVID-19 mobility data refers to information on people's movement patterns and travel behaviors during the pandemic. It provides insights (Zhang el.at. (2021)) into how individuals and communities have changed their mobility patterns in response to COVID-19, such as changes in travel modes, frequency, and destinations. Here are some key aspects of COVID-19 mobility data:

- **Travel volume:** Mobility data tracks changes in travel volume, including the number of trips taken and the distance traveled. It helps in understanding the overall level of mobility in different geographic areas and how it has been affected by measures such as lockdowns, travel restrictions, and social distancing guidelines.
- **Mode of transportation:** Mobility data captures changes in the modes of transportation used by individuals. It tracks shifts in the usage of public transportation, private vehicles, walking, cycling, and other travel modes. This information helps in assessing the impact on different sectors of transportation and understanding the preferences and choices made by individuals during the pandemic.
- **Commuting patterns:** COVID-19 mobility data analyzes changes in commuting patterns, including the number of people commuting to work, the duration of commutes, and the use of alternative work arrangements such as remote work. It provides insights into how workplaces and commuting behaviors have been affected by the pandemic.
- **Destination preferences:** Mobility data examines changes in destination preferences during the pandemic. It tracks shifts in travel to specific locations such as workplaces, shopping centers, recreational areas, and public spaces. This information helps in understanding the impact of COVID-19 on various sectors, urban planning, and the distribution of economic activities.
- **Social gatherings and public spaces:** COVID-19 mobility data analyzes changes in the use of public spaces and the frequency of social gatherings. It helps in assessing compliance with social distancing measures, identifying areas of overcrowding or non-compliance, and understanding the impact on community transmission.
- **Geographic variations:** Mobility data provides insights into geographic variations in mobility patterns. It helps in understanding differences in travel behaviors across regions, urban-rural divides, and areas with varying levels of COVID-19 incidence. This information can guide targeted interventions and policy decisions.

COVID-19 mobility data is typically collected from various sources, including mobile phone data, transportation agencies, GPS tracking, and survey data. It plays a crucial role in understanding the dynamics of the pandemic, assessing the effectiveness of public health measures, and informing decision-making related to transportation planning, public spaces, and social distancing guidelines(Dadashzadeh et.al.(2022)).

2.5. Other Auxiliary Data Sources

In addition to the primary data sources such as epidemiological, clinical, socioeconomic, and mobility data, there are several auxiliary data sources that can provide valuable insights and support the analysis of COVID-19. These auxiliary data sources complement the primary data and offer additional context and information(Alamo et.al. (2020)). Here are some examples of auxiliary data sources for COVID-19:

- **Weather data:** Weather conditions can potentially impact the spread and transmission of respiratory viruses, including COVID-19. Auxiliary weather data, such as temperature, humidity, and air quality, can be used to analyze the relationship between weather patterns and the incidence of COVID-19 cases. It can help identify potential correlations or associations between weather variables and disease dynamics.
- **Air travel data:** Air travel has played a significant role in the global spread of COVID-19. Data on flight routes, passenger volumes, and airport activity can provide insights into international and domestic travel patterns, identify potential transmission routes, and assess the impact of travel restrictions and border control measures.
- **Environmental data:** Environmental factors, such as pollution levels, population density, and urbanization, may influence the transmission and severity of COVID-19. Auxiliary environmental data can be used to explore correlations between environmental factors and disease outcomes, identify high-risk areas, and inform targeted interventions.
- **Genomic sequencing data:** Genomic sequencing of SARS-CoV-2, the virus causing COVID-19, provides insights into the genetic variants, mutations, and transmission patterns. Genomic data can help track the spread of specific variants, understand their implications for transmission and vaccine effectiveness, and inform public health responses, such as contact tracing and targeted interventions.
- **Social media and online platforms:** Social media platforms, online forums, and search engine data can provide real-time insights into public sentiment, concerns, and behaviors related to COVID-19. Analyzing social media data can help understand public perceptions, misinformation trends, adherence to preventive measures, and public response to vaccination campaigns.
- **Healthcare utilization data:** Data from healthcare systems, including hospital admissions, emergency department visits, and ICU occupancy rates, can provide information on the burden of COVID-19 on healthcare facilities and help assess healthcare system capacity and resilience.
- **Mobility tracking apps:** Some countries and regions have implemented contact tracing or mobility tracking apps that collect data on individuals' movements and potential exposure to COVID-19. This data can support contact tracing efforts, identify potential transmission chains, and inform public health responses.

- **Demographic and population data:** Demographic data, such as age, gender, ethnicity, and socioeconomic factors, can be integrated with COVID-19 data to analyze differential impacts and identify vulnerable populations. Population data, including population density, urban-rural distribution, and migration patterns, can provide insights into disease spread and resource allocation.

It is important to note that the use of auxiliary data sources requires proper data privacy protection, ethical considerations, and adherence to applicable regulations and guidelines. Integration and analysis of these auxiliary data sources with primary data can provide a more comprehensive understanding of the COVID-19 pandemic and inform evidence-based decision-making.

3. Feature Engineering for COVID-19 Forecasting:

Feature engineering plays a crucial role in COVID-19 forecasting as it involves selecting and creating relevant input variables that capture important aspects of the pandemic's dynamics (Darji et.al.(2020)). Here are some key feature engineering techniques used in COVID-19 forecasting:

- **Temporal Features:** Time-related features are essential for capturing the temporal patterns and trends of the pandemic. These features can include day of the week, month, season, and public holidays. Additionally, lagged variables, such as previous case and death counts, can help model the time-dependent nature of the virus spread.
- **Demographic Features:** Demographic factors can influence the spread and impact of COVID-19. Including variables such as population density, age distribution, socioeconomic indicators, and healthcare infrastructure can provide insights into the vulnerability of different regions and populations.
- **Mobility and Connectivity:** Movement patterns and connectivity between regions can impact the transmission dynamics of the virus. Incorporating mobility data, such as daily commuting patterns, transportation networks, and social distancing measures, can help capture the influence of human mobility on the spread of the virus.
- **Environmental Factors:** Environmental conditions may affect the survival and transmission of the virus. Variables like temperature, humidity, and air quality indices can be considered as potential features, as they can have an impact on virus viability and human behavior.
- **Interventions and Policies:** Government interventions and policy measures, such as lockdowns, mask mandates, and vaccination campaigns, can significantly influence the spread of COVID-19. Creating binary or categorical variables to represent the presence or absence of these interventions can help capture their effects on transmission rates.
- **Healthcare Capacity:** The capacity and resilience of healthcare systems play a vital role in managing the pandemic. Including variables such as hospital beds per capita, ICU capacity, and healthcare workforce density can provide insights into the healthcare system's ability to handle COVID-19 cases.
- **Social Media and News Data:** Social media and news platforms can offer valuable information about public sentiment, awareness, and adherence to preventive measures. Sentiment analysis, topic modeling, or aggregated metrics from social media platforms can be used as features to capture public behavior and responses to the pandemic.
- **Data Aggregation and Transformations:** Aggregating data at different spatial scales (e.g., country, state, city) and temporal resolutions (e.g., daily, weekly) can help capture localized patterns and reduce noise. Additionally, transformations such as moving averages, exponential smoothing, or logarithmic scaling can help stabilize and normalize the data.
- **Feature Selection and Dimensionality Reduction:** Depending on the forecasting model and available data, feature selection techniques like correlation analysis, mutual information, or stepwise regression can be employed to identify the most informative features. Dimensionality reduction techniques like principal component analysis (PCA) or t-SNE can be used to reduce the feature space while preserving important information.

It's important to note that the choice of features and techniques may vary depending on the specific forecasting task and data availability(Mohammad et.al.(2021)). Furthermore, feature engineering should be coupled with appropriate modeling techniques to build accurate and robust COVID-19 forecasting models.

4. Machine Learning Models for COVID-19 Forecasting

Several machine learning models have been used for COVID-19 forecasting, aiming to predict the spread, severity, and impact of the pandemic. Here are some commonly employed models:

4.A. Epidemic Models: Machine learning epidemic models aim to use machine learning techniques to understand and predict the spread of infectious diseases within a population. These models typically leverage historical data on the disease's transmission, population demographics, and other relevant factors to make predictions and inform public health interventions. Here are a few examples of machine learning epidemic models:

Susceptible-Infected-Recovered (SIR) Model: The SIR model is a classic epidemiological model that divides the population into three categories: susceptible (S), infected (I), and recovered (R). Machine learning techniques can be applied to estimate the parameters of the SIR model, such as the transmission rate and recovery rate, based on observed data.(cooper et.al.(2020)).

Gaussian Processes: Gaussian processes are a probabilistic machine learning technique that can be used to model and predict epidemic dynamics. These models can capture complex patterns and provide uncertainty estimates in their predictions. Gaussian processes have been applied to analyze infectious disease data and make predictions on disease spread(Ketu et.al.(2021)).

Agent-Based Models: Agent-based models simulate the behaviour and interactions of individual agents within a population to model disease spread. Machine learning techniques can be applied to estimate the parameters of the agent-based model or to calibrate the model based on observed data (Krivorotko et.al.(2022)).

It's important to note that machine learning epidemic models are not a one-size-fits-all solution and should be carefully tailored to the specific disease and context. The availability and quality of data, model interpretability, and the ability to validate and update the model with new data are important considerations when developing and using machine learning epidemic models. These models can assist public health authorities in understanding disease dynamics, making informed decisions, and implementing effective interventions.

- **Time Series Forecasting Models:** Time series models are suitable for analyzing and forecasting sequential data. Techniques like Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing (ES) models have been applied to COVID-19 time series data to predict future trends based on historical patterns. These models consider the temporal dependencies within the data and can capture seasonality and trends (Balli et.al.(2021)).
- **Machine Learning Regression Models:** Regression models, such as linear regression, support vector regression (SVR), and random forest regression, can be used to predict COVID-19 outcomes based on input features. These models establish relationships between input variables and target variables (e.g., daily cases, hospitalizations, or mortality rates) and can capture non-linear relationships and interactions between features (Gambhir et.al.(2020)).
- **Long Short-Term Memory (LSTM) Networks:** LSTM networks are a type of recurrent neural network (RNN) that can model sequential data and capture long-term dependencies(Chandra et.al.(2022)). LSTMs have been utilized for COVID-19 forecasting by learning patterns in time series data, considering temporal context, and making predictions. They can handle variable-length input sequences and have been successful in capturing complex dynamics.
- **Convolutional Neural Networks (CNN):** CNNs are primarily used for image analysis, but they can also be adapted for COVID-19 forecasting (Shenvi et.al.(2021)). By treating time series data as images or using convolutional filters to analyse temporal patterns, CNNs can capture spatial and temporal correlations in the data. This approach has been used, for example, in analysing mobility data and predicting the impact of social distancing measures.
- **Ensemble Models:** Ensemble methods combine the predictions of multiple individual models to obtain a more robust and accurate forecast(Chowell et.al.(2022)). Techniques such as model averaging, bagging, and boosting can be used to combine the outputs of different models, such as regression models or neural networks. Ensemble models can help reduce biases and errors and provide more reliable predictions.
- **Graph Neural Networks (GNN):** GNNs are used when data is represented as a graph, such as networks of contact between individuals or geographic networks (Panagopoulos et.al.(2021)). GNNs can capture complex relationships between nodes in the graph and have been applied to model the spread of COVID-19 in social networks or geographic regions.

It's important to note that the choice of the most appropriate model depends on various factors, including the available data, the forecasting task (e.g., short-term vs. long-term), the desired level of granularity, and the specific research or policy questions. Additionally, model performance should be regularly evaluated and updated as new data becomes available, and models should be validated using appropriate evaluation metrics and techniques.

5. Evaluation Metrics for COVID-19 Forecasting

Evaluation metrics are essential for assessing the performance and accuracy of COVID-19 forecasting models (Raschka et.al.(2018)). Here are some commonly used evaluation metrics for COVID-19 forecasting where n indicates number of observations, y indicates actual values & **y'** indicates predicted values.

Mean Absolute Error (MAE): It is a metric used to evaluate the performance of a regression model. It measures the average magnitude of the errors between the predicted values and the actual values in a dataset. MAE provides a straightforward measure of the average absolute deviation of the predictions from the ground truth.
The formula for calculating MAE is as follows:
MAE = (1/n) * Σ|y - y'|
Mean Absolute Percentage Error (MAPE): it is a metric used to evaluate the accuracy of a forecasting or regression model. It measures the average percentage difference between the predicted values and the actual values in a dataset. MAPE is particularly useful when you want to assess the model's performance in terms of relative errors or percentage deviations.

The formula for calculating MAPE is as follows:
MAPE = (1/n) * Σ(|(y - y') / y| * 100)
Root Mean Squared Error (RMSE): Root Mean Squared Error (RMSE) is a commonly used metric for evaluating the performance of regression models. It measures the average magnitude of the residuals (i.e., the differences between predicted values and actual values) by taking the square root of the average of the squared residuals.

The formula for calculating RMSE is as follows:
RMSE = √((1/n) * Σ(y - y')^2)
Symmetric Mean Absolute Percentage Error (SMAPE): It is a metric used to measure the accuracy of a forecasting model or a regression model. It calculates the average percentage difference between the predicted values and the actual values, but unlike the traditional Mean Absolute Percentage Error (MAPE), SMAPE uses a symmetric formulation that handles both overestimation and underestimation equally.

The formula for calculating SMAPE is as follows:
SMAPE = (1/n) * Σ(2 * |y - y'| / (|y| + |y'|) * 100)
Percentage of Accuracy (PA): It is a simple metric used to measure the accuracy of a classification or prediction model. It represents the percentage of correct predictions made by the model out of the total number of predictions.

The formula for calculating Percentage of Accuracy is as follows:
PA = (Number of Correct Predictions / Total Number of Predictions) * 100
R-squared (R^2) Score: It , also known as the coefficient of determination, is a statistical measure used to evaluate the goodness of fit of a regression model. It provides an indication of how well the model's predictions fit the actual observed data.

The R^2 score ranges from 0 to 1, where:
0 indicates that the model does not explain any of the variability in the target variable.
1 indicates that the model perfectly explains the variability in the target variable.
The formula for calculating R^2 score is as follows:
R^2 = 1 - (SSres / SStot)
Where:
SSres: Sum of squared residuals (also known as the sum of squared errors or SSE)
SStot: Total sum of squares (sum of the squared differences between each data point and the mean of the target variable)

Mean Error (ME): ME, also known as the Mean Forecast Error (MFE), is a metric used to measure the average difference between predicted values and actual values in a forecasting or regression model. ME provides information about the bias or systematic error in the model's predictions.

The formula for calculating Mean Error is as follows:
ME = (1/n) * Σ(y - y')
Forecast Skill Score: Forecast Skill Score (FSS), also known as the Forecast Verification Skill Score, is a metric used to evaluate the performance of a forecasting model by comparing it to a reference forecast or a baseline model. FSS quantifies the improvement or skill of the model over the reference forecast in making accurate predictions.

The formula for calculating Forecast Skill Score depends on the specific application and the nature of the forecast. Different formulas are used for different types of forecasts, such as categorical forecasts, probabilistic forecasts, or continuous variable forecasts. I can provide you with an example of a commonly used FSS formula for categorical forecasts known as the Heidke Skill Score (HSS).

The Heidke Skill Score (HSS) is used to evaluate the skill of a categorical forecast, such as predicting the occurrence or non-occurrence of an event. It compares the forecast to random chance and provides a score ranging from -1 to 1, where:
- 1 indicates a perfect forecast,
- 0 indicates a forecast as good as random chance, and
- -1 indicates a forecast worse than random chance.

The formula for calculating Heidke Skill Score (HSS) is as follows:
HSS = (2 * (HitRate - RandomRate)) / (1 - RandomRate)
Where:

HitRate: The proportion of correctly predicted events (hits) out of the total events (hits + false alarms)
RandomRate: The expected proportion of correctly predicted events by random chance, calculated as **((hits + misses) * (hits + false alarms)) / (total)^2**

Coverage Probability: coverage probability is typically used in the context of prediction intervals for regression models. A prediction interval provides a range of values within which a future observation is expected to fall with a certain level of confidence. The coverage probability measures the proportion of observed values that fall within the prediction intervals.

Coverage Probability = (Number of Observed Values within Prediction Intervals) / (Total Number of Predicted Values)

- **Epidemic Curve Shape Analysis:** Apart from quantitative metrics, visual analysis of the epidemic curve's shape can provide insights into the model's performance. Comparing the predicted and actual curves visually can help identify discrepancies, delays, or discrepancies in peak timing or trends.

It's important to consider the specific context and requirements of the forecasting task when selecting evaluation metrics. Additionally, it's recommended to use multiple metrics to gain a comprehensive understanding of the model's performance.

6. Challenges in COVID-19 Forecasting using Machine Learning

COVID-19 forecasting using machine learning models is a challenging task due to several factors(Ferrari et.al. (2020)). Here are some of the key challenges encountered in COVID-19 forecasting:

- **Limited and Noisy Data:** COVID-19 data can be limited, incomplete, or noisy, making it challenging to build accurate forecasting models. Data quality issues, variations in testing and reporting practices, and delays in data availability can introduce biases and uncertainties into the models.
- **Rapidly Evolving Dynamics:** The COVID-19 pandemic is a dynamic and evolving phenomenon. The spread of the virus, interventions, and human behavior can change rapidly, making it difficult for forecasting models to capture and adapt to these changes in real-time.
- **Non-Stationarity and Seasonality:** COVID-19 data often exhibits non-stationarity and seasonality, making it challenging to capture the underlying patterns. The spread of the virus can vary over time, and seasonal effects may arise due to factors like weather, holidays, or variations in human behavior.
- **Complex Interactions and Dynamics:** The spread of COVID-19 involves complex interactions between various factors, such as population characteristics, mobility patterns, intervention measures, and human behavior. Capturing and modeling these complex dynamics accurately can be challenging for machine learning models.
- **Uncertainty and Volatility:** COVID-19 forecasting is inherently uncertain due to the unpredictability of the virus and human responses. Models need to account for and quantify uncertainties to provide reliable and meaningful forecasts. Incorporating probabilistic forecasting techniques can help address this challenge.
- **Data Sparsity and Heterogeneity:** COVID-19 data can vary significantly across regions, countries, and even within a single region. Variations in data availability, reporting standards, and demographic factors can introduce heterogeneity and data sparsity issues. Models should be able to handle such heterogeneity and generalize across different contexts.
- **Adaptation to Interventions and Policy Changes:** Interventions and policy measures, such as lockdowns, social distancing, and vaccination campaigns, can significantly impact the spread of the virus. Forecasting models need to incorporate these interventions and their effects accurately, which can be challenging due to varying compliance levels and evolving policies.
- **Lack of Historical Precedence:** The COVID-19 pandemic is a unique event with no historical precedent in recent times. This lack of historical data on similar pandemics or similar magnitudes of outbreaks can pose challenges in building forecasting models that rely on historical patterns and trends.
- **Generalization and Transferability:** Models trained on data from one region or country may not generalize well to other regions due to differences in demographics, healthcare systems, interventions, and cultural factors. Ensuring the transferability of models across different contexts requires careful consideration and adaptation.
- **Ethical and Societal Implications:** COVID-19 forecasting has significant ethical and societal implications. Biases in data, model interpretability, and potential consequences of incorrect forecasts can have far-reaching effects on public health decision-making. Ensuring transparency, fairness, and responsible use of forecasting models is crucial.

Addressing these challenges requires a combination of careful data collection, preprocessing, model selection, calibration, and ongoing model validation and refinement. Collaboration between domain experts, data scientists, and policymakers is essential to navigate these challenges and develop reliable forecasting models for COVID-19(shinde et.al.(2020)).

7. Future Directions and Recommendations:

COVID-19 forecasting using machine learning continues to be an active area of research and development. Here are some future directions and recommendations to enhance COVID-19 forecasting using machine learning:

- **Incorporate Real-Time Data:** Integrating real-time data sources, such as hospital admissions, testing rates, mobility patterns, and social media data, can enhance the accuracy and timeliness of forecasts. Developing techniques to effectively incorporate these diverse data sources into forecasting models is important.

- **Improve Data Quality and Standardization:** Efforts should be made to improve data quality, consistency, and standardization across regions and countries. Clear guidelines for data collection, reporting, and metadata should be established to reduce biases and improve comparability across different datasets.
- **Account for Variants and Immunity:** COVID-19 variants and evolving immunity levels can significantly impact the spread of the virus. Forecasting models should consider the emergence and effects of variants and the evolving dynamics of population immunity in their predictions.
- **Enhance Interpretability and Explainability:** Developing machine learning models that are interpretable and provide explanations for their forecasts is crucial for building trust and facilitating decision-making. Techniques like model interpretability algorithms, causal inference, and visualization methods should be explored.
- **Consider Spatial and Temporal Dependencies:** COVID-19 spread has both spatial and temporal dependencies. Models that effectively capture and leverage these dependencies, such as spatial-temporal models or graph-based models, can provide more accurate forecasts at regional or local scales.
- **Address Data Sparsity and Heterogeneity:** COVID-19 data can be sparse and heterogeneous, making it challenging to build models that generalize across regions or countries. Developing techniques that address data sparsity and heterogeneity, such as transfer learning, domain adaptation, or meta-learning, can improve model performance and generalizability.
- **Quantify and Communicate Uncertainty:** Forecasting models should provide probabilistic predictions and quantify uncertainty. Techniques like ensemble modeling, Bayesian approaches, or Monte Carlo simulations can help estimate and communicate uncertainty to policymakers and the public effectively.
- **Consider Social and Behavioral Factors:** COVID-19 spread is influenced by social and behavioral factors. Integrating data on human behavior, compliance with interventions, and socioeconomic factors can enhance the accuracy of forecasting models and support targeted interventions.
- **Evaluate Intervention Effectiveness:** Machine learning models can be used to evaluate the effectiveness of different interventions and policy measures. Conducting retrospective analyses and causal inference studies can provide insights into the impact of interventions and guide future decision-making.
- **Collaborate and Share Knowledge:** Collaboration between researchers, public health agencies, policymakers, and data scientists is essential to advance COVID-19 forecasting. Sharing data, models, and methodologies can facilitate knowledge exchange and foster collective learning.

It is important to continuously evaluate and refine forecasting models as new data becomes available and the pandemic evolves. Transparency, robust validation, and frequent model updates are key to ensuring the accuracy and reliability of COVID-19 forecasting using machine learning techniques.

8. Conclusion

By examining the state-of-the-art techniques, challenges, and future directions, this review paper aims to provide researchers and practitioners with a comprehensive understanding of the role of machine learning in COVID-19 forecasting. Through ongoing research and collaboration, the development of accurate and robust forecasting models will contribute to effective pandemic management and mitigation strategies.

9. References:

1. She, J., Liu, L., & Liu, W. (2020). COVID-19 epidemic: disease characteristics in children. *Journal of medical virology*, *92*(7), 747-754.. https://doi.org/10.1002/jmv.25807
2. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. (2020). The COVID-19 pandemic. *Critical reviews in clinical laboratory sciences*, *57*(6), 365-388. DOI: 10.1080/10408363.2020.1783198
3. Pokhrel, S., & Chhetri, R. (2021). A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher education for the future*, *8*(1), 133-141. **https://doi.org/10.1177/2347631120983481**
4. Verma, A. K., & Prakash, S. (2020). Impact of covid-19 on environment and society. *Journal of Global Biosciences*, *9*(5), 7352-7363.
5. Singh, J., & Singh, J. (2020). COVID-19 and its impact on society. *Electronic Research Journal of Social Sciences and Humanities*, *2*. https://ssrn.com/abstract=3567837
6. Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., ... & Kraemer, M. U. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific data*, *7*(1), 106. https://doi.org/10.1038/s41597-020-0448-0
7. Marshall, J. C., Murthy, S., Diaz, J., Adhikari, N. K., Angus, D. C., Arabi, Y. M., ... & Zhang, J. (2020). A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases*, *20*(8), e192-e197. https://doi.org/10.1016/S1473-3099(20)30483-7
8. Kumar, S., Maheshwari, V., Prabhu, J., Prasanna, M., Jayalakshmi, P., Suganya, P., ... & Jothikumar, R. (2020). Social economic impact of COVID-19 outbreak in India. *International journal of pervasive computing and communications*, *16*(4), 309-319. https://doi.org/10.1108/IJPCC-06-2020-0053
9. Mishra, N. P., Das, S. S., Yadav, S., Khan, W., Afzal, M., Alarifi, A., ... & Nayak, A. K. (2020). Global impacts of pre-and post-COVID-19 pandemic: Focus on socio-economic consequences. *Sensors International*, *1*, 100042. https://doi.org/10.1016/j.sintl.2020.100042

10. Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., ... & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery*, *78*, 185-193. https://doi.org/10.1016/j.ijsu.2020.04.018
11. Zhang, J., Feng, B., Wu, Y., Xu, P., Ke, R., & Dong, N. (2021). The effect of human mobility and control measures on traffic safety during COVID-19 pandemic. *PLoS one*, *16*(3), e0243263. https://doi.org/10.1371/journal.pone.0243263
12. Dadashzadeh, N., Larimian, T., Levifve, U., & Marsetič, R. (2022). Travel behaviour of vulnerable social groups: Pre, during, and post COVID-19 pandemic. *International journal of environmental research and public health*, *19*(16), 10065. **https://doi.org/10.3390/ijerph191610065**
13. Alamo, T., Reina, D. G., Mammarella, M., & Abella, A. (2020). Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, *9*(5), 827. **https://doi.org/10.3390/electronics9050827**
14. Darji, P. A., Nayak, N. R., Ganavdiya, S., Batra, N., & Guhathakurta, R. (2022). Feature extraction with capsule network for the COVID-19 disease prediction though X-ray images. *Materials Today: Proceedings*, *56*, 3556-3560. https://doi.org/10.1016/j.matpr.2021.11.512
15. Mohammed, M. A., Abdulkareem, K. H., Garcia-Zapirain, B., Mostafa, S. A., Maashi, M. S., Al-Waisy, A. S., ... & Le, D. N. (2021). A Comprehensive Investigation of Machine Learning Feature Extraction and ClassificationMethods for Automated Diagnosis of COVID-19 Based on X-ray Images. *Computers, Materials & Continua*, *66*(3). https://doi.org/10.32604/cmc.2021.012874
16. Cooper, I., Mondal, A., & Antonopoulos, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals*, *139*, 110057. https://doi.org/10.1016/j.chaos.2020.110057
17. Ketu, S., & Mishra, P. K. (2021). Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection. *Applied Intelligence*, *51*, 1492-1512. https://doi.org/10.1007/s10489-020-01889-9
18. Krivorotko, O., Sosnovskaia, M., Vashchenko, I., Kerr, C., & Lesnic, D. (2022). Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm. *Infectious Disease Modelling*, *7*(1), 30-44. https://doi.org/10.1016/j.idm.2021.11.004
19. Ballı, S. (2021). Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. *Chaos, Solitons & Fractals*, *142*, 110512. https://doi.org/10.1016/j.chaos.2020.110512
20. Gambhir, E., Jain, R., Gupta, A., & Tomer, U. (2020, September). Regression analysis of COVID-19 using machine learning algorithms. In *2020 International conference on smart electronics and communication (ICOSEC)* (pp. 65-71). IEEE. doi: 10.1109/ICOSEC49089.2020.9215356.
21. Chandra, R., Jain, A., & Singh Chauhan, D. (2022). Deep learning via LSTM models for COVID-19 infection forecasting in India. *PloS one*, *17*(1), e0262708. https://doi.org/10.1371/journal.pone.0262708
22. Shenvi, D. R., & Shet, K. (2021, March). CNN based COVID-19 prevention system. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 873-878). IEEE. doi: 10.1109/ICAIS50930.2021.9396004.
23. Chowell, G., Dahal, S., Tariq, A., Roosa, K., Hyman, J. M., & Luo, R. (2022). An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the COVID-19 pandemic in the USA. *PLoS Computational Biology*, *18*(10), e1010602. https://doi.org/10.1371/journal.pcbi.1010602
24. Panagopoulos, G., Nikolentzos, G., & Vazirgiannis, M. (2021, May). Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 6, pp. 4838-4845). https://doi.org/10.1609/aaai.v35i6.16616
25. Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
https://doi.org/10.48550/arXiv.1811.12808
26. Ferrari, D., Milic, J., Tonelli, R., Ghinelli, F., Meschiari, M., Volpi, S., ... & Guaraldi, G. (2020). Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia—challenges, strengths, and opportunities in a global health emergency. *PLoS One*, *15*(11), e0239172. https://doi.org/10.1371/journal.pone.0239172
27. Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. *SN computer science*, *1*, 1-15. https://doi.org/10.1007/s42979-020-00209-9