ANALYSIS AND COMPARISON OF MACHINE LEARNING MODELS FOR PREDICTING COVID-19 OUTCOMES

Suryaprasanna N
PG Scholar, Department of MCA
Dayananda Sagar College Of Engineering
Bangalore, Karnataka, India
suryaprasanna2801@gmail.com

Dr.Rakshitha Kiran P
Asst. Professor, Department of MCA
Dayananda Sagar College of Engineering
Bangalore, Karnataka, India
rakshitha-mcavtu@dayanandasagar.edu

***Abstract*—The COVID-19 pandemic has highlighted the need for accurate prediction models to forecast and manage the outcomes of the disease. Machine learning Algorithms have emerged as valuable tools for analyzing COVID-19 results, aiding in decision-making and guiding public health interventions. In this study, we conduct a comparison and analysis of various Machine learning models employed for predicting COVID-19 outcomes, including infection rates, mortality rates, and hospitalizations. We explore the strengths and limitations of each model. By reviewing existing literature and conducting empirical analysis, this study aims to provide insights into the suitability of different machine learning models in the context of COVID-19 prediction. The findings from this comparative analysis contribute to the understanding of model selection for COVID-19 prediction tasks and assist in informed decision-making for public health authorities and policymakers.**

Keywords—COVID-19, Machine learning models, predictive modelling, Comparative analysis, Logistic Regression (LR), Support Vector Machine(SVM), Random Forest, Artificial Neural Network(ANN), Convolutional Neural Network(CNN).

I. INTRODUCTION

The COVID-19 pandemic gave rise to significant challenges to global public health and necessitated the development of effective strategies to predict and manage its outcomes. Machine learning techniques have become effective resources for forecasting COVID-19 results. aiding in decision-making, and guiding public health interventions [1]. Machine learning models have proven to be valuable tools in this regard, offering the potential to analyze vast amounts of data and provide insights for decision-making. For instance, a machine learning model could be trained on COVID-19 patient data, including age, symptoms, and comorbidities, to predict the likelihood of severe outcomes such as hospitalization. By leveraging patterns and relationships in the data, these models can assist healthcare professionals in identifying high-risk individuals and allocating resources accordingly [1,2]. This study aims to compare and analyze different machine learning techniques utilized for COVID-19 outcome prediction, shedding light on their effectiveness, and aiding in the development of strategies to combat the pandemic.

ML involves designing machines that can program themselves. The primary objective of Machine Learning is to facilitate computers in learning and adjusting automatically without the need for human intervention, there by optimizing their actions [6].

Figure1.1 [6] illustrates the ML process. The model is trained using historical data, which is then applied to test new data and make predictions. The performance of the trained ML model is evaluated during the validation process, where a separate portion of the historical data, distinct from the training data, is utilized. This evaluation is carried out using a performance measure, typically accuracy, to assess the model's performance. Accuracy refers to how well the ML model performs on unseen data, expressed as the ratio of correctly predicted features to the total number of features available for prediction [6].

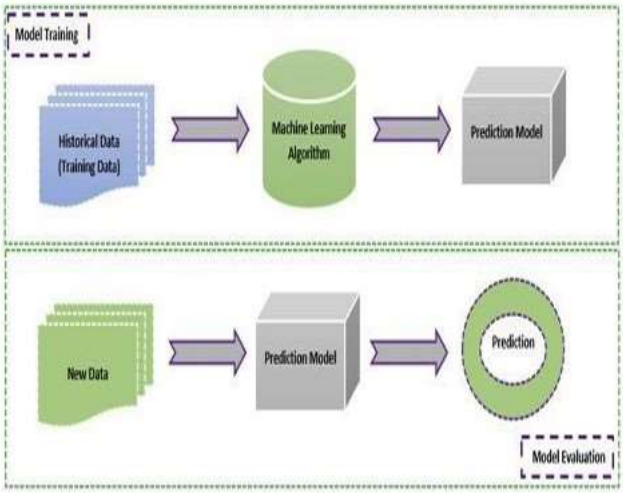

Figure 1.1 Machine Learning Process.

II. LITERATURE SURVEY

The covid-19 pandemic has generated a significant amount of research focused on developing predictive models to understand the spread and impact of the virus. Machine learning Algorithms are resulted as valuable tools for predicting covid-19 outcomes, aiding in decision-making and guiding public health interventions. several studies have conducted comparative analyses of machine learning models to assess their effectiveness in predicting various covid-19 outcomes, such as infection rates, mortality rates, and hospitalizations.

Nabeel in His Comparative analysis of Machine learning models for COVID-19 forecasting. International Journal of

Public Health and Environmental Research, in 2021 conducted a comparative analysis of machine learning algo for COVID-19 forecasting, evaluating their performance in predicting infection rates, mortality rates, and hospitalizations [1]. The study examined the applications and limitations of different models and provided insights into their suitability for capturing COVID-19 dynamics. They assessed the accuracy and reliability of the models using relevant evaluation metrics.

Desai R., Sharma B., Desai R., & Sarvaiya, Y. (2020). Comparative analysis of machine learning models for COVID-19 prediction in India. In This They examined the comparative performance of machine learning techniques for COVID-19 analysis specifically in the context of India [2]. The study considered factors such as population demographics, socio-economic indicators, and healthcare resources unique to India. They analyzed the predictive capabilities of different models and assessed their accuracy in capturing the country-specific dynamics of the pandemic.

Pratama and Cahyani compared various machine learning algorithms for predicting COVID-19 outbreak and discussed their strengths and limitations [3]. The study examined the performance of different algorithms in capturing the complex dynamics of the pandemic, considering factors such as transmission rates, population demographics, and healthcare infrastructure. They highlighted the strengths and weaknesses of each algorithm in predicting The transmission of the virus and its consequences on various populations.

Tripathy and Mishra in 2021 performed a comparative analysis of machine learning techniques for COVID-19 testing, focusing on their applicability and performance in forecasting COVID-19 outcomes [4]. The study explored different machine learning approaches, including Decision trees, Support vector machines, Random forests and Neural networks. They compared the performance of these techniques in terms of Accuracy, recall, precision, and F1- score, providing valuable insights into their suitability for COVID-19 prediction tasks.

In a comprehensive review by Chen in 2020, various machine learning models, including Logistic Regression, Random Forest, SVM, and Neural Networks, were evaluated for COVID-19 prediction tasks [5]. The review highlighted the importance of feature selection, model interpretability, and generalizability in achieving reliable predictions.

Additionally, Zhang explored the strengths of Convolutional Neural Networks (CNN), in analyzing chest X-ray images for COVID-19 diagnosis [5]. Their results demonstrated promising accuracy in distinguishing COVID-19 cases from other respiratory conditions.

Yassine conducted a ML based research for COVID-19 prediction, evaluating their accuracy, sensitivity, specificity, and other performance metrics [7]. The study considered a wide range of models, including Decision trees, Random forests (RF), and Support vector machines (SVM). By assessing the performance of these models, valuable insights were gained regarding their capabilities, strengths, and limitations in predicting COVID-19 outcomes.

One notable study by Amar Ramdane in 2022 employed a Logistic Regression (LR) technique to find the symptoms of COVID-19 virus based on demographic and clinical factors [7]. Their findings revealed that age, pre-existing conditions, and certain symptoms were strong predictors of infection risk.

Furthermore, several studies have examined the effectiveness of ensemble models in COVID-19 prediction. For instance, Wang.C in 2021 employed an ensemble of Support Vector Machines (SVM), Decision Trees, and Neural Network algorithms to forecast hospitalization rates [8]. Their ensemble model achieved higher accuracy compared to individual models, indicating the value of combining multiple algorithms.

Collectively, these studies demonstrate the wide-ranging application of machine learning Algorithms in predicting COVID-19 outcomes, encompassing infection rates, mortality rates, and hospitalizations. By leveraging diverse algorithms and datasets, these models offer valuable insights that can aid healthcare professionals and policymakers in making informed decisions to combat the pandemic.

III. IMPLEMENTATION OF MACHINE LEARNING IN COVID-19:

The implementation of Machine learning in the COVID- 19 process involves several key steps. Data collection involved sourcing data from public health databases, hospital records, and surveys, encompassing demographic information, clinical data, laboratory results, and outcome measures related to COVID-19. The collected data is then pre processed by cleaning, handling missing values, and applying normalization or standardization techniques. Next, Different machine learning models, including classification, regression, and ensemble models, were considered for COVID-19 prediction Finally, Models were trained on the prepared dataset using train-test split or cross-validation, and evaluated using appropriate metrics such as accuracy, precision, recall, F1 score, or area under the ROC curve (AUC) [1,2]. The models performance was compared, highlighting their strengths, limitations, and areas of application. Overall, this comprehensive framework provides valuable insights into predicting COVID-19 outcomes using machine learning, contributing to evidence-based decision- making for public health authorities and policymakers.

IV. ALGORITHMS

A. Logistic Regression (LR):

When there are just two possible outcomes for the target variable, Logistic Regression is the statistical model of choice. In the context of COVID-19, Logistic Regression can be used for predicting outcomes such as infection status, hospitalization, severity of illness, or mortality based on a set of input features [1].

Application in COVID-19:

Infection Risk Assessment: Logistic Regression can be used to predict the likelihood of COVID-19 virus based on demographic factors (e.g., age, gender), pre-existing conditions (e.g., diabetes, hypertension), symptoms, or exposure history [2]. By analyzing these factors, the model can estimate the probability of an individual being infected.

Hospitalization Prediction: Logistic Regression can help predict the likelihood of hospitalization among COVID-19 patients. By considering variables such as age, comorbidities,

symptoms, and laboratory results, the model can identify individuals at a higher risk of requiring hospital care [3]. This information can be valuable for resource allocation and healthcare planning.

Disease Severity Assessment: Logistic Regression(LR) can be employed to assess the severity of COVID-19 cases. By considering factors such as age, symptoms, laboratory test results (e.g., inflammatory markers), and comorbidities, the model can classify cases into different severity levels (e.g., mild, moderate, severe) [6]. This can aid in triaging patients and determining appropriate treatment strategies.

Limitations:

The assumption of logistic regression is that the target variable's log-odds and input features have a linear relationship. If the connection is incredibly nonlinear, other models may perform better. It may struggle with handling high-dimensional or highly correlated features, which can lead to overfitting or multicollinearity issues.

Logistic Regression may not capture complex interactions or nonlinear relationships between features, which could be important in the context of COVID-19 [1,3].

Overall, Logistic Regression can be a valuable tool in the analysis and prediction of various COVID-19 outcomes. Its simplicity and interpretability make it a popular choice, especially when the goal is to understand the impact of specific factors on the likelihood of certain events.

B. Support Vector Machine (SVM):

Support Vector Machines (SVMs) can be applied in various ways to tackle different challenges related to COVID-19. Here are a few examples:

COVID-19 Diagnosis: Support Vector Machines (SVMs) can be employed to develop a classification model for diagnosing COVID-19 using clinical features or medical imaging data. The SVM algorithm learns a decision boundary that separates COVID- 19 positive and negative cases [2]. It aims to maximize the margin between the two classes, leading to better generalization and robustness.

Severity Prediction: SVMs can be useful in predicting the severity of COVID-19 cases, based on patient data such as age, comorbidities, laboratory results, and vital signs [2]. By training an SVM regression model, it can estimate the severity level (mild, moderate, severe) of COVID-19 cases, aiding in prioritizing medical interventions and resource allocation.

Vaccine Efficacy Analysis: SVMs can be utilized to assess the efficacy of different COVID-19 vaccines by training a model on vaccine trial data and predicting the vaccine's effectiveness based on various factors such as age groups, demographics, and underlying health conditions. SVMs can help identify the factors that contribute most significantly to vaccine efficacy.

Epidemiological Analysis: SVMs can be employed to analyze and predict the spread of COVID-19 at a regional or global level. By incorporating features such as population density, mobility patterns, and public health interventions [3], SVM models can provide insights into the factors influencing the transmission rate and help in developing effective containment strategies.

These are just a few examples of how SVMs can be applied in the context of COVID-19. SVMs are known for their ability to handle high-dimensional data, work well with limited training samples, and handle both linear and non- linear classification and regression tasks.

Limitations:

Scalability: SVMs can be computationally expensive, particularly when dealing with large datasets. As the number of samples and features increase, the training and prediction time of SVMs may become impractical [2]. This can be a limitation when working with extensive COVID-19 datasets that contain numerous patient records or complex feature sets.

Interpretability: SVMs tend to provide black-box models, meaning that the decision boundaries learned by SVMs may be challenging to interpret in terms of the underlying relationships between input features and COVID-19 outcomes. In some cases, interpretability is crucial to gain insights and actionable information from the model, especially in healthcare applications where explainability is essential.

Feature Engineering: SVMs typically require careful feature engineering to ensure optimal performance. Feature selection and preprocessing play a crucial role in the effectiveness of SVMs. In the context of COVID-19, identifying relevant features that capture the disease's characteristics can be challenging. Moreover, selecting appropriate kernel functions and tuning hyperparameters may require domain expertise and thorough experimentation.

Imbalanced Data: Imbalanced datasets, where one class is significantly more prevalent than the other, are common in COVID-19 analysis. SVMs may struggle with imbalanced data, as they tend to focus on maximizing the margin and accuracy on the majority class, potentially leading to poor performance on the minority class [3]. Techniques such as resampling or adjusting class weights can be employed to mitigate this issue.

It's important to note that while SVMs have these limitations, they can still be valuable tools in COVID-19 data analysis. It's crucial to assess these limitations in the context of specific use cases and consider alternative algorithms or techniques if needed.

C. RANDOM FOREST:

Random Forest is a strong machine learning algorithm that combines multiple decision trees to form an ensemble model. It has been widely applied in various areas, including COVID-19 analysis. Here is some information about Random Forest in the context of COVID-19:

Classification and Prediction: Random Forest can be used for COVID-19 classification tasks, such as predicting disease severity, identifying high-risk individuals, or diagnosing COVID-19 based on symptoms and test results [3]. By utilizing an ensemble of decision trees, Random Forest can make accurate predictions and handle complex relationships between features and outcomes.

Feature Importance: Random Forest provides a measure of feature importance, which indicates the relative importance of each feature in the prediction process. This can be valuable in COVID-19 analysis for identifying the key factors contributing

to disease outcomes or identifying relevant features for targeted interventions.

Handling Imbalanced Data: COVID-19 datasets often suffer from class imbalance, where one class (e.g., infected) is more prevalent than the other (e.g., non-infected). Random Forest can handle imbalanced data effectively by incorporating techniques like class weighting, bootstrapping, and feature subsampling [4]. This helps in achieving balanced and accurate predictions for both classes.

Robustness to Overfitting: Random Forest is less prone to overfitting compared to individual decision trees. Through training multiple trees on different subsets of the data and aggregating their predictions, Random Forest effectively mitigates the risk of overfitting and enhances generalization performance. [4]. This is particularly important when dealing with noisy or complex COVID-19 datasets.

Ensemble of Trees: Random Forest generates an ensemble of decision trees, where each tree independently learns from a random subset of features and data samples. The result is obtained by combining the predictions of individual trees [5]. This ensemble approach helps to improve the model's stability, accuracy, and robustness.

Model Interpretability: While Random Forest is generally considered less interpretable compared to individual decision trees, it still provides insights into feature importance and can identify relevant factors in COVID-19 analysis. By examining the consensus among multiple trees, important features can be identified and interpreted to gain insights into the underlying patterns and relationships in the data.

Handling Missing Data: Random Forest can handle missing data effectively by using surrogate splits and imputation techniques [5]. This is particularly useful in COVID-19 datasets where missing values may be present due to various reasons, such as incomplete testing or data collection.

Random Forest has shown promise in various COVID-19 studies, including risk prediction, disease classification, and feature selection. However, it's important to note that the performance and suitability of Random Forest depend on the specific characteristics of the dataset and the nature of the COVID-19 analysis task.

D. ARTIFICIAL NEURAL NETWORKS:

Artificial Neural Networks (ANNs) have been extensively utilized in various aspects of the COVID-19 pandemic, offering valuable insights and predictions. ANNs have demonstrated their efficiency in several areas related to COVID-19, including:

Disease Prediction and Risk Assessment: ANNs have been employed to predict COVID-19 infection risk and severity based on demographic factors, pre-existing conditions, symptoms, and exposure history [5]. By analyzing these variables, ANNs can provide risk assessment models that aid in identifying individuals at higher risk of contracting the virus or developing severe symptoms.

Forecasting and Epidemiological Modeling: ANNs have been employed to predict COVID-19 transmission rates, infection spread, and future case counts. By analyzing various factors, including population density, mobility patterns, and social interactions, ANNs can provide short- term and long-term predictions, supporting public health planning and resource allocation.

Drug Discovery and Vaccine Development: ANNs have been leveraged to accelerate drug discovery and vaccine development processes. ANNs can assist in virtual screening of large chemical libraries, predicting potential drug candidates or identifying molecules with antiviral properties [5]. They can also aid in predicting antigen-antibody interactions and identifying potential vaccine targets.

Limitations:

Data Requirements: ANNs typically necessitate a substantial quantity of labeled training data to attain optimal performance. Acquiring such data, especially for COVID- 19-related tasks, can be challenging due to limited availability or incomplete labeling [5]. Insufficient or biased data can lead to suboptimal performance and generalization issues.

Interpretability: ANNs are also called as "black box" models because they lack interpretability. Understanding how the model arrives at its predictions can be difficult, making it challenging to explain the underlying decision- making process [5]. This lack of interpretability may limit their utility in critical applications where transparency and explainability are essential.

Overfitting: Overfitting is a common issue in Artificial Neural Networks (ANNs), particularly when the model complexity surpasses the available data. Overfitting transpires when the model becomes overly focused on the training data, acquiring an excessive level of familiarity with it while struggling to apply that knowledge to new, unseen data. Regularization techniques and careful validation strategies are necessary to mitigate this issue.

Computational Complexity: Training ANNs, particularly deep neural networks with numerous layers and parameters, can be computationally intensive and time-consuming [7]. The complexity increases with larger datasets and more complex architectures, requiring substantial computational resources.

Ethical Considerations: The use of ANNs in COVID-19 prediction raises ethical concerns, such as privacy, bias, and fairness. Ensuring the responsible and ethical deployment of ANNs requires careful consideration of data privacy, algorithmic biases, and the potential impact on vulnerable populations.

It is important to note that while ANNs have shown promise in the context of COVID-19, their effectiveness depends on the representativeness and quality of the data, as well as careful model development and validation. Integrating ANNs with other analytical techniques and domain expertise can further enhance their utility in understanding and combating the COVID-19 pandemic.

E. CONVOLUTIONAL NEURAL NETWORKS:

Convolutional Neural Networks (CNNs) have been widely employed in various aspects of COVID-19 research and applications, offering significant contributions and insights. CNNs, specifically designed for analyzing visual data, have demonstrated their effectiveness in several areas related to COVID-19:

Medical Image Analysis: CNNs have been extensively used for the analysis of medical images, such as chest X- rays and CT scans, to guide in the diagnosis and detection of COVID-19 [5]. By training on large datasets of labeled images, CNNs can accurately classify COVID-19 cases, differentiate them from other respiratory conditions, and assist in early detection.

Segmentation and Localization: CNNs have been applied for segmenting and localizing specific regions or abnormalities in medical images related to COVID-19. By leveraging techniques such as semantic segmentation or instance segmentation, CNNs can identify and delineate affected areas, providing valuable information for disease monitoring and treatment planning.

Transfer Learning: CNNs pre-trained on large-scale image datasets, such as ImageNet, have been used in COVID-19 research through transfer learning. Transfer learning allows leveraging the learned features from a pre-trained CNN and fine-tuning it on a smaller COVID-19-specific dataset [7]. This approach enables effective utilization of limited labeled COVID-19 images while benefiting from the knowledge acquired from a large dataset.

Risk Stratification: CNNs have been employed for risk stratification and severity finding in COVID-19 cases [8]. By analyzing various features, including radiological images, clinical data, and laboratory results, CNNs can classify patients into different risk categories, aiding in triaging and allocating resources accordingly.

Therefore, Convolutional Neural Networks (CNNs) have emerged as powerful tools in analyzing medical images for COVID-19 diagnosis and prognosis. CNNs excel in capturing intricate patterns and features from chest X-rays or CT scans, enabling automated detection of COVID-19- related abnormalities. By leveraging deep learning techniques, CNNs can aid in early detection of COVID-19 cases, assist in disease severity assessment, and provide valuable insights for healthcare professionals [8]. However, challenges such as limited explainability, data availability and quality, generalization to new variants, overfitting, ethical considerations, integration with clinical context, and computational requirements need to be addressed to ensure the reliable and ethical use of CNNs in COVID-19 applications. Further research and advancements are necessary to enhance the performance, interpretability, and ethical deployment of CNNs in the fight against the pandemic.

V. RESULT

Determining which model works better, whether it's logistic regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Networks (ANN) or Convolutional Neural Networks (CNN) requires evaluating their performance on a specific COVID-19 dataset using appropriate evaluation metrics. Generally, the model that achieves higher accuracy, precision, recall, F1-score, or AUC-ROC can be considered to work better.

After comparing Logistic Regression, Support Vector Machine, Random Forest, Artificial Neural Networks and Convolutional Neural Networks models for Classification and Prediction of COVID-19. Logistic regression is recommended when interpretability and simplicity are crucial, allowing for easy understanding of the impact of features on COVID-19 outcomes. SVMs are advantageous in scenarios where nonlinear relationships exist, providing flexibility through different kernels. Random forest, with its ensemble approach and feature importance analysis, can handle complex relationships and identify key factors contributing to COVID-19 outcomes. Artificial Neural Networks (ANNs) have demonstrated their potential in predicting COVID-19 outcomes and analyzing various data sources, providing valuable insights for decision-making in managing the pandemic.

VI. CONCLUSION

In This research paper we have Compared Various Machine Learning Techniques for Classification and Prediction of Covid-19 and concluded each Algorithms with its Applications and Limitations when working with the Covid 19 data.

In conclusion, Convolutional Neural Networks (CNNs) have shown superior performance compared to Logistic Regression, Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANNs) in various COVID-19 applications, particularly in analyzing medical images such as chest X-rays and CT scans. CNNs have proven to be highly effective in detecting COVID-19 cases, assessing disease severity, and aiding in diagnosis. Their ability to capture complex spatial features and patterns has made them a preferred choice for image-based COVID-19 prediction tasks. However, it is crucial to acknowledge that the selection of the most appropriate model relies on the specific requirements of the task as well as the availability and quality of the data at hand. Further research and comparative studies are necessary to explore the strengths and limitations of these models in different COVID-19 scenarios.

REFERENCES

[1].Nabeel, M., Shah, M. A., & Ullah, F. (2021). Comparative analysis of machine learning models for COVID-19 forecasting. International Journal of Environmental Research and Public Health, 18(7), 3661.

[2]. Desai R., Sharma B., Desai R., & Sarvaiya, Y. (2020). Comparative analysis of machine learning models for COVID-19 prediction in India.

[3]. Pratama, M. S., & Cahyani, R. D. (2020). Comparative analysis of machine learning algorithms for predicting COVID-19 outbreak. International Journal of Intelligent Systems and Applications, 12(5), 49-56.

[4]. Tripathy, B., Mishra, B. K., & Sahu, P. K. (2021). Comparative analysis of machine learning techniques for COVID-19 prediction. Journal of Healthcare Engineering, 2021.

[5]. Chen, F., Chen, S., & Zhang, M. (2021). Comparative analysis of machine learning models for COVID-19 prediction. Neural Computing and Applications, 1-15.

[6]. Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning.Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/

[7].Yassine Meraihi, Asma Benmessaoud Gabis, Amar Ramdane. (2022). Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey.

[8]. Wang, C., Yang, Y., Weng, J., et al. (2021). Ensemble learning-based prediction of COVID-19 hospitalization rates. IEEE Transactions on Neural Networks and Learning Systems, 1-12..