Covid-19 Analysis and Prediction with machine learning algorithms

Arshia Ilaty
2^{nd} semester student of UNR in the Master of Computer Science and Engineering

Abstract— **In the healthcare industry, the role of artificial intelligence and data science is becoming increasingly important due to the rise of automation. Healthcare organizations are working on developing automated systems that can better manage patient care. Machine learning (ML) is a technology that can help diagnose and treat diseases more accurately, potentially leading to earlier diagnoses and fewer patients. In this chapter, the authors discuss two ML solutions for predicting COVID-19 infections and forecasting future cases. The autoregressive integrated moving average (ARIMA) time series and two classifiers, random forest and extra tree classifier (ETC), are compared, with ETC achieving the highest accuracy of 93.62**

These forecasting techniques could be used to take corrective measures against infectious diseases like COVID-19. The paper aims to introduce the basics of ML and demonstrate how it can be used to predict and forecast COVID-19, potentially informing future healthcare automation tasks.

I. Introduction and Motivation

The COVID-19 pandemic has had a significant impact on global health and the world economy. Governments and healthcare organizations have struggled to manage the pandemic and mitigate its effects. Accurate prediction and forecasting of COVID-19 cases can help healthcare professionals and policymakers make informed decisions and allocate resources effectively. Machine learning (ML) is a technology that can help address this challenge. ML algorithms can analyze large volumes of data and identify patterns and trends that may predict future COVID-19 cases.

The motivation for this project and paper is to explore the use of ML for COVID-19 prediction and forecasting. The primary goal of the project is to develop an ML-based system that can accurately predict the number of COVID-19 cases in a given region and forecast future cases. The project will involve collecting and analyzing data on COVID-19 cases, hospitalizations, and deaths from various sources, including government agencies and healthcare organizations.

The motivation for this project stems from the urgent need for accurate COVID-19 prediction and forecasting. The pandemic has had a significant impact on global health and the world economy. Accurate prediction and forecasting of COVID-19 cases can help healthcare professionals and policymakers make informed decisions and allocate resources effectively. ML-based systems can analyze large volumes of data and identify patterns and trends that may predict future COVID-19 cases with higher accuracy than traditional statistical models.

The potential benefits of developing an ML-based COVID-19 prediction and forecasting system are significant. Firstly, it can help healthcare professionals and policymakers make informed decisions and allocate resources effectively. Accurate prediction and forecasting can help hospitals and healthcare organizations prepare for an influx of patients and ensure that they have the necessary resources, such as ventilators and personal protective equipment. Secondly, it can help governments and public health officials develop effective public health interventions, such as vaccination campaigns and social distancing measures. Lastly, it can help individuals and businesses make informed decisions about their activities and investments, reducing uncertainty and minimizing the economic impact of the pandemic.

Overall, the use of ML for COVID-19 prediction and forecasting is a promising area of research that can provide significant benefits to healthcare professionals, policymakers, governments, and the public. The primary goal of this project and paper is to develop an ML-based system that can accurately predict the number of COVID-19 cases in a given region and forecast future cases. By doing so, we hope to contribute to the development of more effective COVID-19 prediction and forecasting systems that can help manage the pandemic and mitigate its effects.

II. Prediction and Forecasting Techniques

A. Model selection and development

Several ML techniques are utilized for predicting and forecasting future events. Support vector machine,

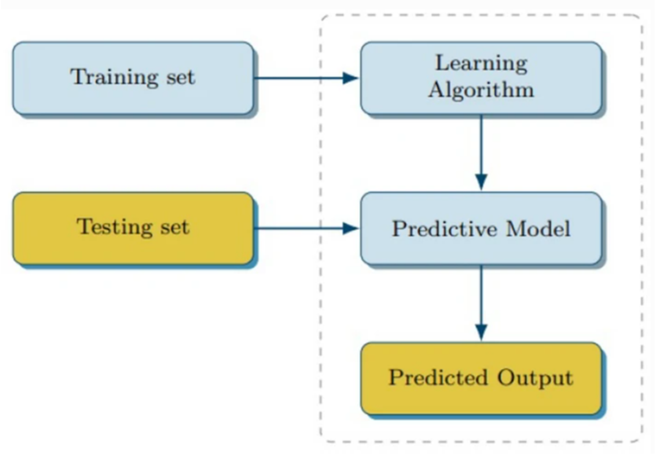

Fig. 1. Training and Testing split.png.

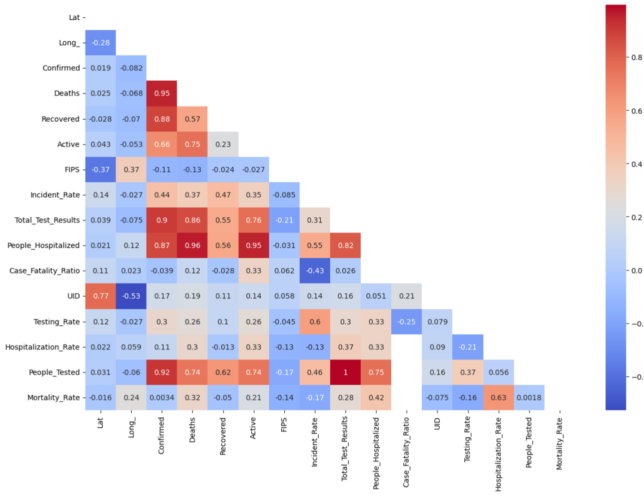

Fig. 2. Correlation matrix values for Covid-19-US Dataset.

linear regression, logistic regression, naive Bayes, decision trees (random forest and ETC), K-nearest neighbor, and neural networks (multilayer perceptron) are some of the ML techniques used for prediction.

Similarly, various ML techniques used for forecasting future events include the naive approach, moving average, simple exponential smoothing, Holt's linear trend model, Holt-Winters model, Seasonal Autoregressive Integrated Moving Average Exxogenous Model (SARIMAX), and Autoregressive Integrated Moving Average Model (ARIMA).

Each ML technique has its own distinctive features and is utilized differently based on the accuracy results. The model with the highest accuracy during the model evaluation process is selected for prediction or forecasting. In the same manner, ETC was chosen for the symptom-based prediction of COVID-19, and the ARIMA forecasting model was employed for predicting the number of confirmed cases of COVID-19 in the US, as they produced the most accurate results among all classifier and forecasting methods evaluated during the model performance evaluation.

Overall, Various machine learning models such as SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayes Classifier, Multilinear Regression, Logistic Regression, Random Forest Classifier, and XGBoost Classifier are developed using two datasets. The evaluation of the models is done by calculating the R-Squared (coefficient of determination) regression score and accuracy. The dataset is divided into training and testing sets with a ratio of 70:30.

B. correlation matrix for feature selection and engineering

A correlation matrix is a table utilized in the feature selection process to determine correlation coefficients between variables or features. Each cell in the matrix signifies the correlation between two variables. It is employed to summarize a vast dataset and to identify the features that have the highest correlation.

The correlation coefficient's value near 1 signifies that features participating in correlation are highly correlated to each other; on the other hand, the correlation coefficient's value near 0 signifies that features are less correlated to each other.

C. Deep learning application models

Artificial Intelligence (AI) based algorithms have the ability to learn from historical data to provide predictions for future outcomes. Two key subsets of AI are machine learning (ML) and deep learning (DL), which rely on computer algorithms to improve their own performance over time. DL is particularly useful for image classification, speech recognition, bioinformatics, and other areas, but its use has been limited by computational complexity and processing power. However, with advancements in big data, larger and deeper networks are now possible, enabling computers to learn, analyze, and react to complex situations more quickly than humans.

This study focuses on the development and evaluation of clinical predictive models to determine COVID-19 infection using laboratory findings. Six different types of models were trained and evaluated: Artificial Neural Network (ANN), Convolutional Neural

Networks (CNN), Long-Short Term Memory (LSTM), Recurrent Neural Networks (RNN), CNNLSTM, and CNNRNN. ANN is a type of information processing system inspired by the biological nervous system, composed of neurons, activation functions, input and output layers, and hidden layers. CNN is a variant of neural networks, particularly useful for image classification, featuring convolutional layers, pooling layers, fully-connected layers, and a classification layer. RNN is a type of feedforward neural network with an internal memory, where the output of each input depends on the past computation. LSTM is a modified version of RNN that is able to remember past data more effectively, resolving the vanishing gradient problem of RNN.

Alongside these four DL models, we also developed the LSTM model.

D. Evaluation

This study compared various machine learning models including logistic regression, naive Bayes, SVM, ANN, and decision tree, to identify the most accurate model for predicting COVID-19 infection cases. The decision tree model was found to be the most accurate in terms of overall accuracy with 94.99

III. Related works

Based on the analysis of selected papers, supervised learning is the most dominant type of machine learning applied in production lines, accounting for 92.9

This analysis highlights the dominance of supervised learning and the prevalence of classification tasks in production-line applications of machine learning. Logistic regression, ANN, and CNN were the most frequently applied algorithms, with logistic regression being the most popular. These findings suggest that machine learning applications in production lines have the potential to improve efficiency and accuracy, and further research is needed to explore the full potential of machine learning in this domain.

IV. Open Issues and Continuing Research

As of now, in 2023, the COVID-19 pandemic has affected over 229 countries and caused more than 750 million confirmed cases worldwide. Numerous studies have been conducted on predicting global outbreaks, and this research aims to review the most significant forecasting models for COVID-19 while providing a concise analysis of the published literature. This study focuses on identifying critical subject areas using keyword analysis and determining several criteria that can guide future researchers. Additionally, this paper highlights the most useful models employed by researchers in predicting this pandemic. Furthermore, this study could assist researchers in identifying gaps in the research area and developing new machine-learning models for forecasting COVID-19 cases.

V. Conclusion

Machine learning has shown great potential in the field of healthcare, especially in the diagnosis and prediction of COVID-19 infections. From the literature review, it is evident that supervised learning is the most dominant type of machine learning used in production lines, with logistic regression being the most commonly used algorithm. The results of the study indicate that decision tree models perform best in terms of accuracy when compared to other models developed with logistic regression, naive Bayes, SVM and ANN. However, SVM and Naïve Bayes models perform better in terms of sensitivity and specificity, respectively. Furthermore, the development and use of machine learning algorithms can play a crucial role in the prediction, diagnosis, and containment of COVID-19, especially in developing countries where healthcare systems are limited.

Overall, the use of machine learning in healthcare has the potential to revolutionize the way we diagnose and treat diseases, especially in the context of pandemics such as COVID-19. The study presented here highlights the importance of machine learning algorithms as effective tools for the prediction and diagnosis of COVID-19 infections, and it is hoped that this will inspire further research in this area to help combat future pandemics.

References

[1] Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19. 2021:381–97. doi: 10.1016/B978-0-12-824536-1.00027-7. Epub 2021 May 21. PMCID: PMC8138040.

[2] Meraihi, Y., Gabis, A.B., Mirjalili, S. et al. Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. SN COMPUT. SCI. 3, 286 (2022). https://doi.org/10.1007/s42979-022-01184-z

[3] Osama Shahid, Mohammad Nasajpour, Seyedamin Pouriyeh, Reza M. Parizi, Meng Han, Maria Valero, Fangyu Li, Mohammed Aledhari, Quan Z. Sheng, Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance, Journal of Biomedical Informatics, Volume 117, 2021, 103751, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2021.103751.

[4] Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z. et al. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inform Decis Mak 22, 2 (2022). https://doi.org/10.1186/s12911-021-01742-0