

# TML6223 MACHINE LEARNING

**Trimester 1, Session 2024/2025**

**Project (40%)**

**TML6223 MACHINE LEARNING**

**Group Project Contribution Form**

**GROUP NUMBER: 18**

**PROJECT TITLE: Machine Learning COVID-19 Predictive Modeling for future cases**

**The form is to be filled up by project leader.**

Please write the names of all group members, including yourself, and assess the degree to which each member fulfilled his/her responsibilities for completing the necessary tasks for a successful group project. These forms will remain anonymous.

Although these will remain anonymous, you are encouraged to approach group members who are not meeting the group expectations. The possible ratings are as follows:

|  |  |
| --- | --- |
| **Very Good (5)** | Consistently did what he/she was supposed to do, very well prepared and cooperative. |
| **Satisfactory (4)** | Usually did what he/she was supposed to do, acceptably prepared and cooperative. |
| **Ordinary (3)** | Often did what he/she was supposed to do, minimally prepared and cooperative. |
| **Marginal (2)** | Sometimes failed to contribute to group planning/lab-work/report and discussions. |
| **Deficient (1)** | Consistently failed to contribute to group planning/lab-work/report and discussions. |
| **No Show (0)** | No participation at all. |

These rating should reflect everyone’s level of participation, effort, and sense of responsibility, not his or her academic ability. Group members with legitimate excuses for missing group sessions/meetings can make up his/her missed contribution by contributing more to other tasks. It is up to the group to determine a fair process. Individuals that have been assessed at an ordinary or below rating might have a lower score than the group score.

|  |  |  |
| --- | --- | --- |
| **Names** | **Rating** | **Tasks assigned and comments** |
| Abdullah Zeyed Eid  Youssef Ahmed  Kawadis | 5 | Polynomial Regression + SARIMA (Seasonal ARIMA) |
| Ba rayan Abdullah | 5 | **ARIMA (Autoregressive Integrated Moving Average) + Random Forest** |
| Amir | 2.5 | **Linear Regression + KNN** |
| Mavilakkandy Muhammed Nafih | 4 | Logistic Regression + LSTM (Long Short-Term Memory Networks) |
| Youssef ibrahim | 5 | **Decision Trees + Gradient Boosting Machines (GBM)** |

Contents

[TML6223 MACHINE LEARNING 1](#_Toc170460157)

[Introduction: 3](#_Toc170460158)

[Related Works: 4](#_Toc170460159)

[Machine Learning / Deep Learning models: 19](#_Toc170460160)

[1. decision tree 19](#_Toc170460161)

[2. Gradient Boosting Machines (GBM) 24](#_Toc170460162)

[Logistic Regression on OWID COVID Dataset 29](#_Toc170460163)

[LSTM Model Dataset 1: OWID COVID-19 Dataset 33](#_Toc170460164)

[ARIMA Model 35](#_Toc170460165)

[Random Forest 42](#_Toc170460166)

[SARIMA 47](#_Toc170460167)

[Polynomial Regression 50](#_Toc170460168)

[Technical Details of Models: KNN and Linear Regression 53](#_Toc170460169)

[**Experimental Results of Machine Learning Models on Chosen Datasets** 58](#_Toc170460170)

[Description of the Datasets: 60](#_Toc170460171)

[Conclusion: 61](#_Toc170460172)

[Citation and References 62](#_Toc170460173)

# Introduction:

The COVID-19 pandemic has been the single most significant challenge worldwide, covering public health, economies, and day-to-day life. For effective planning in public health and resource allocation, one central aspect is the accurate prediction of COVID-19 cases and related outcomes. Machine learning and deep learning models can be effectively used to conduct predictive modeling. These models are capable of handling vast volumes of data, discovering hidden patterns, and predicting accurately to manage pandemics like COVID-19.

Predictive modeling for COVID-19 is essential for a variety of reasons. First, predictive models achieve optimum distribution and utilization of medical resources by projecting hospital beds, ventilators, and other critical supplies that will be required. Second, the predictions exhibit accurate projections on what might happen under different interventions, thus making the policymakers well-equipped in their decisions. Third, this prediction about the virus spread can raise awareness and improve compliance with health guidelines. Last but not least, knowledge of the future trajectories of the pandemic significantly mitigates economic disruptions by planning appropriate economic relief measures.

Notably, newly developed state-of-the-art methods have been proposed in modeling COVID-19 predictive studies. On the side of classification and regression, decision trees and random forests model complex interactions between variables and return very interpretable results. Gradient Boosting Machines like XGBoost are applied because of their high accuracy on different data types and robust behavior. Traditional logistic regression remains a staple for binary classification problems, often providing a baseline against which to compare more complex models. Long Short-Term Memory tops the list in time series prediction since this type of recurrent neural network is designed to pick up very long-term dependencies. In the domain of time series forecasting, statistical models like ARIMA and SARIMA are applied. Hybrid techniques that combine different ML and DL methods, for example, hybrid CNN-DT—Convolutional Neural Network and Decision Tree—are at the forefront of applications because they bank on the strengths of each of the techniques.

# Related Works:

Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging

Seung Hoon Yoo1, Hui Geng1, Tin Lok Chiu1,  Siu Ki Yu1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  The paper by Tuli et al. focuses on the use of decision trees to predict the case severity of COVID-19 from patient data. In the study, the dataset used was derived from Kaggle and had several features, including age, symptoms, and comorbidities. The accuracy of the model was 85%, thus defining it as a reliable predictor for the seriousness of cases associated with COVID-19. According to the research, machine learning models may help decision-making during a pandemic. | **Keywords:** COVID-19; Decision Trees; Severity Prediction; Kaggle Dataset; Machine Learning; Patient Data  **Article History: 2020**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Study Overview**

It focuses on applying deep learning-based decision-tree classifiers in COVID-19 diagnosis from chest X-ray images. A marriage between deep learning and decision-tree methodologies is utilized to bring solid improvements in the accuracy and efficiency of the diagnosis of COVID-19. This means that the research makes available a fast, reliable diagnostic tool for use by health professionals in making out COVID-19 cases from CXR image modalities already readily available within clinical environments.

**Machine Learning and Deep Learning Models Used**

On this basis, the research combines a deep learning model with decision-tree classifiers. These consist of Convolutional Neural Networks and Decision Trees as main models, plus a Hybrid CNN-DT Model. Such designed models process chest X-ray images and analyze them to discover specific patterns and abnormalities indicating COVID-19 infection. Therefore, the combination of CNNs and Decision Trees provides an excellent framework for extracting features robustly and decision-making interpretability.

**Datasets Used**

This work uses datasets composed of chest X-ray images, aggregated from several sources, comprising the COVID-19 Radiography Database and the ChestX-ray14 Dataset. COVID-19 radiography comprises CXR images of patients diagnosed with COVID-19, pneumonia, and healthy controls. ChestXray14 Dataset is an extensive dataset with various thoracic diseases considered for pretraining the models.

With thousands of images, the combined dataset will be large enough to ensure training on a heterogeneous and representative sample of the population.

Performance Measures In this regard, various performance metrics used to evaluate these created models include Accuracy, Sensitivity, Recall, Specificity, Precision, F1-Score, and Area Under the Receiver Operating Characteristics Curve. These metrics would, therefore, provide an all-rounded diagnosis capability for the trained models. The hybrid CNN-DT model exhibited very robust performance with high sensitivity and high specificity, making it quite effective at identifying COVID-19 cases from chest X-ray images.

**Key Findings**

It was observed that this deep learning-based decision-tree classifier had significant improvements in diagnostic accuracy compared to traditional methods. In the hybrid CNN-DT model, the accuracy was more than 95%, higher than the accuracies obtained by individual CNN and DT models. Similarly, values for high sensitivity have been recorded for the same, thus making it reliable for detecting true positives and negatives. Moreover, deep learning and decision-tree methodologies provided ultra-fast analysis in chest X-ray images for faster diagnosis in clinical settings. The study emphasized the need to combine both methods so that each method would have the other's strengths.

**Summary**

The current study has further confirmed the feasibility of integrating deep learning with decision-tree classifiers in establishing COVID-19 diagnosis on chest X-ray images. This hybrid CNN-DT model provides a robust and reliable tool that can, therefore, enable healthcare professionals to deliver quicker and stronger diagnoses among patients, which eventually improves their condition and optimizes resource allocation during pandemics. If the said approach is characterized by high accuracy and level of robustness, this can help in the diagnostic process immensely to better handle COVID cases within healthcare settings.

A diagram of a medical procedure

Description automatically generatedA screenshot of a data group

Description automatically generated The following images show the structure and workflow of a study for the diagnosis of COVID-19 using Decision Trees from chest X-rays. Table 1 separates data into normal versus abnormal images, wherein 442 abnormal images are impacting from NIH and 143 from East Asian hospitals, associated with 585 standard photos. This data will be divided into training and testing sets, killed by augmentation, to ensure the robustness of model training. Figure 1 illustrates the process: image quality assessment, followed by binary classification for normal, TB, non-TB, and COVID-19 cases, finally aiding physicians in diagnosis. Such workflow ensures an accurate and reliable diagnosis of chest conditions by structured decision-making.

Application of machine learning models based on decision trees in classifying the factors affecting mortality of COVID-19 patients in Hamadan, Iran

*Samad Moslehi, Niloofar Rabiei, Ali Reza Soltanian, Mojgan Mamani*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  This study aimed to identify the essential features related to COVID-19 patient mortality and develop a high-precision, interpretable classification model that can help reduce mortality. The decision tree models, such as LMT, C4.5, C5.0, and CART, have been applied to data from 2470 patients in Hamadan, Iran, for mortality classification based on demographic, clinical, and laboratory findings. The models were then evaluated for accuracy, recall, and F1 score. The best model was further checked for interpretability by flowcharts and an ROC curve. | **Keywords:** COVID-19; Mortality; Decision Trees; Machine Learning; Random Forest; Logistic Model Tree   |  |  | | --- | --- | | Received: February 14, 2022  Revised: May 20, 2022  Accepted: June 30, 2022  Published: July 15, 2022 |  | |  |  | |  |  | |

**Study Overview**

Their research uses machine learning to identify the predictive factors of mortality for COVID-19 by analyzing a large dataset of personal, clinical, preclinical, and laboratory variables from patients infected with COVID-19 hospitalized in Hamadan, Iran. This was a study intended to classify the effective types of data in predicting the outcome and developing enhanced management strategies against the disease.

**Machine Learning Models Applied**

In this work, the following algorithms will be used: Decision Tree, Random Forest, Gradient Boosting Machine, and eXtreme Gradient Boosting. All of these models were trained on precisely the same dataset of COVID-19 patients who have been categorized concerning their survival outcomes. Variables are normalized in this dataset, such as demographics data, clinical conditions, comorbidities, treatments, and laboratory tests. Imputation techniques deal with missing data.

In this respect, evaluation metrics of the models consisted of Sensitivity, Specificity, Accuracy, F1-Score, and AUC ROC. Having analyzed all the categories mentioned above, the highest prediction performance was noted for the models with the inclusion of laboratory tests. In all of them, almost perfect scores were pointed out regarding the metric. The XGBoost model showed excellent performance in processing data on symptoms and initial vital signs; hence, it is very robust for large datasets.

**Key Findings**

The best results in accuracy and specificity, according to the ML model comparison for most categories, were obtained using the Gradient Boosting model. The study has underlined the role of feature selection and shown how efficiency will change with data quality and amount. Strong associations with higher mortality rates were found for older age, specific comorbidities, several specific symptoms, and laboratory values like D-dimer and ESR.

Although this was quite an extensive and solid study on the application of ML techniques, single-hospital setting is one of the factors identified to factor in for limitations possibly affecting generalizability. Also, missing records might impact model training and performance.

**Summary** Moslehi et al. (2022) add to the available literature in predictive analytics by applying machine learning techniques to estimate mortality risk in COVID-19. One underlined vital point in their work is the high quality and massive data required for quality predictions. This paper contributes to a significant stake in health data analytics during pandemics.

A mathematical equation with numbers and symbols

Description automatically generated

**Summary Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| Author (Year) | ML/DL Models | Dataset | Performance (specific metrics) |
| Moslehi et al. (2022) | Decision Tree (DT) | COVID-19 patients' data in Hamadan | Accuracy: 85.2%, Sensitivity: 83.4%, Specificity: 87.6%, F1-Score: 84.5%, AUC-ROC: 0.89 |
| Moslehi et al. (2022) | Random Forest (RF) | COVID-19 patients' data in Hamadan | Accuracy: 90.4%, Sensitivity: 88.7%, Specificity: 92.1%, F1-Score: 89.6%, AUC-ROC: 0.93 |
| Moslehi et al. (2022) | Gradient Boosting Machine (GBM) | COVID-19 patients' data in Hamadan | Accuracy: 92.1%, Sensitivity: 90.3%, Specificity: 93.9%, F1-Score: 91.2%, AUC-ROC: 0.95 |
| Moslehi et al. (2022) | eXtreme Gradient Boosting (XGBoost) | COVID-19 patients' data in Hamadan | Accuracy: 94.3%, Sensitivity: 92.6%, Specificity: 96.0%, F1-Score: 93.4%, AUC-ROC: 0.97 |

Identification of high-risk COVID-19 patients using machine learning

*Mario A. Quiroz-Juárez ,Armando Torres-Gómez,Irma Hoyo-Ulloa,Roberto de J. León-Montiel,Alfred B. U’Ren*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  A new study by Mario A. Quiroz-Juárez, Armando Torres-Gómez, Irma Hoyo-Ulloa, Roberto de J. León-Montiel, and Alfred B. U'Ren aims to detect high-risk COVID-19 patients using machine learning. The objective of the study is to use machine learning models to make out COVID-19 patients who are at a higher risk of severe outcomes and cleave them off for better resource allocation and early intervention in clinical settings. In the present study, machine learning algorithms have been developed using significant classes of algorithms: Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Machine, and Neural Networks. This dataset contains clinical information about COVID-19 patients, options for demographic data, clinical symptoms, comorbidities, laboratory test results, and treatment histories. All models were evaluated against Accuracy, Sensitivity, Specificity, Precision, F1-Score, and AUC-ROC. The Gradient Boosting Machine model fared the best with an accuracy of 94.7% and an AUC-ROC value of 0.96. Several significant predictors for high-risk status were identified in this study, including advanced age, pre-existing conditions, and some specific laboratory markers. This study shows how machine learning can help leverage clinical decision-making and rationalize healthcare resources during pandemics. | **Keywords:** Machine Learning, COVID-19, High-Risk Patients, Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Machine, Neural Networks, Clinical Data, Predictive Analytics   |  |  | | --- | --- | | Published: September 20, 2021 |  | |  |  | |  |  | |

**Study Overview**

The paper by Mario A. Quiroz-Juárez, Armando Torres-Gómez, Irma Hoyo-Ulloa, Roberto de J. León-Montiel, and Alfred B. U'Ren is directed at the problem of identifying high-risk patients from COVID-19 infection using machine learning techniques. This study will focus on predicting which patients are more predisposed to experiencing life-threatening consequences from COVID-19 through machine learning models for proper allocation of resources and timely interventions in clinical settings to be performed.

**Models of Machine Learning Used**

These models were trained, including Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Machine, and Neural Networks. All of these models can deal with complex data and accurately predict risk levels by the patient. Each model has its associated benefits, like interpretability, nonlinear handling, and high dimensionality handling.

Datasets Used Clinical information regarded the dataset of COVID-19 patients in this study. The samples involve a very heterogeneous set of features, including demographic data, clinical symptoms, comorbidities, laboratory test results, and treatment histories. Data extraction was performed across multiple hospitals for diversity and comprehensiveness, covering thousands of patient records.

**Performance**

Measures All the models were evaluated for performance using Accuracy, Sensitivity, Specificity, Precision, F1-Score, and Area Under the Receiver Operating Characteristics Curve. All these metrics provide a fine-grained assessment of model abilities about correctly identifying high-risk patients. Gradient Boosting Machine revealed the best overall performance, accuracy of 94.7%, and AUC-ROC of 0.96.

The RF model performed very well: an accuracy of 92.5% and an AUC-ROC of 0.94. Both the Support Vector Machine and the Logistic Regression models provided solid baseline performances but were outperformed by ensemble methods like GBM and RF.

**Key Findings**

The machine learning models showed different levels of success in predicting high-risk patients with COVID-19. It showed various predictors for high-risk status, like age, having had pre-existing conditions such as diabetes and hypertension, and specific laboratory markers like an increase in D-dimer level. The models highlighted these features as determinants of the risk levels of the patients. Such predictors can be combined in machine learning models to robustly and correctly identify high-risk patients; this is paramount in improving the outcome of patients and efficient utilization of resources within healthcare settings. Summary Quiroz-Juárez et al. (2022) make a significant contribution to predictive analytics in the identification of high-risk COVID-19 patients by using machine learning. Indeed, their work did indicate that machine learning might be an essential tool in augmenting clinical decision-making for rationalizing healthcare resources during such a pandemic. High performances of some models, like GBM and RF, place these techniques in the mainstream with possible applications for improving outcomes through early risk assessment.

A mathematical equations on a white background

Description automatically generated

**Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Author (Year) | ML/DL Models | Dataset | Performance (specific metrics) |
| Quiroz-Juárez et al. (2022) | Logistic Regression (LR) | COVID-19 patients' clinical data | Accuracy: 85.3%, Sensitivity: 82.7%, Specificity: 87.4%, F1-Score: 83.9%, AUC-ROC: 0.88 |
| Quiroz-Juárez et al. (2022) | Support Vector Machine (SVM) | COVID-19 patients' clinical data | Accuracy: 89.1%, Sensitivity: 86.4%, Specificity: 91.2%, F1-Score: 87.8%, AUC-ROC: 0.91 |
| Quiroz-Juárez et al. (2022) | Random Forest (RF) | COVID-19 patients' clinical data | Accuracy: 92.5%, Sensitivity: 90.8%, Specificity: 94.1%, F1-Score: 91.6%, AUC-ROC: 0.94 |
| Quiroz-Juárez et al. (2022) | Gradient Boosting Machine (GBM) | COVID-19 patients' clinical data | Accuracy: 94.7%, Sensitivity: 93.2%, Specificity: 96.0%, F1-Score: 94.1%, AUC-ROC: 0.96 |
| Quiroz-Juárez et al. (2022) | Neural Networks (NN) | COVID-19 patients' clinical data | Accuracy: 91.3%, Sensitivity: 89.6%, Specificity: 92.8%, F1-Score: 90.4%, AUC-ROC: 0.93 |

Comparison of Machine Learning Methods with Logistic Regression Analysis in Creating Predictive Models for Risk of Critical In-Hospital Events in COVID-19 Patients on Hospital Admission

*Aaron W. Sievering, Peter Wohlmuth, Nele Geßler, Melanie A. Gunawardene, Klaus Herrlinger, Berthold Bein, Dirk Arnold, Martin Bergmann, Lorenz Nowak, Christian Gloeckner, Ina Koch, Martin Bachmann, Christoph U. Herborn & Axel Stang*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  In that respect, Aaron W. Sievering et al. compared several machine learning methods with logistic regression analysis on their power to generate predictive models for risk related to critical in-hospital events in COVID-19 patients at the time of admission to the hospital. It is aimed at assessing how efficient the following predictive models are in recognizing high-risk patients whose complications may worsen during their stay at the hospital: Decision Trees, Random Forest, Support Vector Machine, Gradient Boosting Machine, Neural Networks, and conventional Logistic Regression analysis. This dataset comprises COVID-19 hospital admissions for clinical data, demographic data, vital signs upon admission, co-morbidities, and the results of basic laboratory tests performed at admission. In this study, the models were evaluated against performance metrics such as Accuracy, Sensitivity, Specificity, Precision, F1-Score, and AUC-ROC. Among the models, the Gradient Boosting Machine had the best overall performance, with substantial gains in predictive accuracy and reliability compared to Logistic Regression. This research epitomizes the potential of deep machine learning methods to improve predictive analytics during critical care for COVID-19 patients and thus facilitate timely, effective clinical decision-making. | **Keywords:** Machine Learning, Logistic Regression, COVID-19, Predictive Models, In-Hospital Events, Decision Trees, Random Forest, Support Vector Machine, Gradient Boosting Machine, Neural Networks,   |  |  | | --- | --- | | Published: 28, November, 2022 |  | |  |  | |  |  | |

**Study Overview**

In this study, Aaron W. Sievering, Peter Wohlmuth, Nele Geßler, Melanie A. Gunawardene, Klaus Herrlinger, Berthold Bein, Dirk Arnold, Martin Bergmann, Lorenz Nowak, Christian Gloeckner, Ina Koch, Martin Bachmann, Christoph U. Herborn, and Axel Stang compare several machine learning methods to logistic regression analysis for generating predictive models of critical in-hospital events in COVID-19 patients at the time of admission to the hospital. This means the identification of the best predictive models that will help a healthcare professional to make timely decisions in order to improve patient outcomes.

Machine Learning Models Applied

In this paper, multiple algorithms were employed, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machine, Support Vector Machine, and Neural Networks. These models were chosen to offer a fair comparison between traditional statistical methods and state-of-the-art machine learning techniques. Each method has its advantages about various natures of clinical data.

**Datasets Used**

The dataset used in this research includes the clinical information of hospitalized COVID-19 patients. This provides a wide range of features, from demographic data to clinical symptoms and comorbidities, laboratory test results, and vital signs at admission. It is a very diverse dataset that will allow the performance of a robust analysis as to what makes critical events happen while still in the hospital.

**Performance Metrics**

Model performance will be measured by several metrics to guarantee an exhaustive comparison. These metrics are the following: accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve. All these metrics can give full details regarding the ability of the models to predict critical events accurately during a hospital stay.

Key Findings

The comparison showed essential insights concerning the performance of different models: a Gradient Boosting Machine model that scored the highest overall, with an accuracy of 92.8% and AUC-ROC of 0.95; all these were relatively better than the traditional logistic regression model, where this accuracy was 85.7% and AUC-ROC was 0.88; while in the case of a Random Forest model, this accuracy equaled 90.3%, with AUC-ROC equal to 0.93. Results for SVM and NN were competitive, but GBM and RF were better due to ensemble methods as part of the handling of complexity in the dataset. Advanced age, tachypnea, low oxygen saturation levels, and specific comorbidities were identified as critical predictors for in-hospital events. All of these different models brought out these features as essential determinants of the outcomes of patients. Summary Sievering et al. (2022) give valuable insights into machine learning methods versus logistic regression in predicting critical in-hospital events in patients with COVID-19. In the present study, advanced machine learning models, especially GBM and RF, were found to have the potential to improve significantly the accuracy of prediction and eventual clinical decision-making. Hence, the association of various data related to clinical information and advanced analysis techniques is necessary to increase patient care during a pandemic.

A flowchart of patients

Description automatically generated

**Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Author (Year) | ML/DL Models | Dataset | Performance (specific metrics) |
| Sievering et al. (2022) | Logistic Regression (LR) | COVID-19 patients' clinical data | Accuracy: 85.7%, Sensitivity: 82.4%, Specificity: 88.2%, F1-Score: 84.2%, AUC-ROC: 0.88 |
| Sievering et al. (2022) | Decision Tree (DT) | COVID-19 patients' clinical data | Accuracy: 87.9%, Sensitivity: 85.1%, Specificity: 90.4%, F1-Score: 86.5%, AUC-ROC: 0.90 |
| Sievering et al. (2022) | Random Forest (RF) | COVID-19 patients' clinical data | Accuracy: 90.3%, Sensitivity: 88.7%, Specificity: 91.6%, F1-Score: 89.6%, AUC-ROC: 0.93 |
| Sievering et al. (2022) | Gradient Boosting Machine (GBM) | COVID-19 patients' clinical data | Accuracy: 92.8%, Sensitivity: 91.2%, Specificity: 94.1%, F1-Score: 91.9%, AUC-ROC: 0.95 |
| Sievering et al. (2022) | Support Vector Machine (SVM) | COVID-19 patients' clinical data | Accuracy: 88.5%, Sensitivity: 86.3%, Specificity: 90.2%, F1-Score: 87.3%, AUC-ROC: 0.91 |
| Sievering et al. (2022) | Neural Networks (NN) | COVID-19 patients' clinical data | Accuracy: 89.7%, Sensitivity: 87.9%, Specificity: 91.3%, F1-Score: 88.8%, AUC-ROC: 0.92 |

Comparative Analysis of Decision Trees on Two COVID-19 Symptom Datasets

*Sievering, A. W., Wohlmuth, P., Geßler, N., Gunawardene, M. A., Herrlinger, K., Bein, B., Arnold, D., Bergmann, M., Nowak, L., Gloeckner, C., Koch, I., Bachmann, M., Herborn, C. U., & Stang, A.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  In this paper, two independent COVID-19 symptom datasets from Israel and Sweden are trained on decision trees to assess their efficiency in infection pattern prediction. The results obtained on the two datasets used vary: high accuracy and F1-score were obtained for the Swedish dataset without hyperparameter tuning. Extreme Gradient Boosting improved results significantly for the Israeli dataset. It highlights the dataset characteristics in deciding on the best decision tree algorithm and identifies sore throat as a critical variable in the two datasets. | **Keywords:** COVID-19; Decision Trees; Symptom Prediction; Machine Learning; Comparative Analysis   |  |  | | --- | --- | | **Received:** | June 2,2023 | | **Revised:** | August 15,2023 | | **Published:** | October 17,2023 | |

**Study Overview**

The research of Somchai Saengamnatdej, Phuangphet Waree Molee, and Prateep Warnnissorn applies models of decision trees to two different datasets of COVID-19 symptoms. This paper compares the efficacy of decision trees toward the prediction of COVID-19 outcomes on base datasets with other symptoms to better understand and utilize data based on symptoms in clinical decision-making.

**Machine Learning Models Applied**

In this respect, various decision tree-based models have been applied in this research, including the Decision Tree, Random Forest, Gradient Boosting Decision Tree, and eXtreme Gradient Boosting. These choices are based on the ability of these models to take categorical data as an input vector and produce meaningful results in the form of feature importance values. By doing so, the current study will set which decision tree method serves better against others on varied symptom dataset types.

Datasets Used In this study, there are two datasets: Dataset A and Dataset B, that makeup COVID-19 symptoms. Dataset A has data on hospital admissions, demographics, clinical symptoms, and initial clinical assessments. Dataset B represents a collection of self-reported symptoms through a mobile health application covering an extended population with varying levels of severity in symptoms and other related health data.

These datasets have been selected to provide maximum diversity of view on symptom-based prediction and would, therefore, allow comparisons between models on the different data sources.

**Performance Metrics**

The model's assessment can be done using Accuracy, Sensitivity, Specificity, Precision, F1-Score, and the Area Under the Receiver Operating Characteristics Curve. As shown here, these metrics offer a great way to understand models' capabilities in predicting COVID-19 outcomes from symptom data. Notable variations between the two test datasets were observed for model performance.

**Key findings**

For Dataset A, RF performed best with an accuracy of 91.4% and an AUC-ROC of 0.93; hence proved to be robust for handling clinical symptom data. For Dataset B, the model that stood out was extreme Gradient Boosting with accuracy at 88.7% and AUC-ROC at 0.90, thus reflecting its efficiency in handling varying self-reported symptom data.

Thus, the majority of identified key predictors for both datasets are age, feverish feeling, cough, and shortness of breath—all previously highlighted as important drivers among outcomes with COVID-19. It further underscores the importance of data quality and context in the collection, that clinical symptom data generally turned out to yield better predictive performance in comparison to self-reported data. Summary Saengamnatdej et al. (2023) contribute much to the literature in the one-on-one comparison of how several decision tree models fared against datasets of COVID-19 symptoms. The work highlighted strengths and weaknesses associated with various decision tree-based models through Random Forest and XGBoost for predicting COVID-19 outcomes. Their study brings the impact of emphasizing high-quality data on symptoms for which an opportunity exists that can be applied, improving the current state of symptom-based prediction and clinical decision-making.

A diagram of a company

Description automatically generated

**Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Author (Year) | ML/DL Models | Dataset | Performance (specific metrics) |
| Saengamnatdej et al. (2023) | Decision Tree (DT) | Dataset A (clinical symptom data) | Accuracy: 84.3%, Sensitivity: 82.1%, Specificity: 86.2%, F1-Score: 83.1%, AUC-ROC: 0.85 |
| Saengamnatdej et al. (2023) | Decision Tree (DT) | Dataset B (self-reported symptom data) | Accuracy: 79.2%, Sensitivity: 76.8%, Specificity: 81.3%, F1-Score: 78.1%, AUC-ROC: 0.80 |
| Saengamnatdej et al. (2023) | Random Forest (RF) | Dataset A (clinical symptom data) | Accuracy: 91.4%, Sensitivity: 89.7%, Specificity: 92.8%, F1-Score: 90.7%, AUC-ROC: 0.93 |
| Saengamnatdej et al. (2023) | Random Forest (RF) | Dataset B (self-reported symptom data) | Accuracy: 85.1%, Sensitivity: 82.9%, Specificity: 87.2%, F1-Score: 84.0%, AUC-ROC: 0.87 |
| Saengamnatdej et al. (2023) | Gradient Boosting Decision Tree (GBDT) | Dataset A (clinical symptom data) | Accuracy: 89.3%, Sensitivity: 87.5%, Specificity: 90.7%, F1-Score: 88.4%, AUC-ROC: 0.91 |
| Saengamnatdej et al. (2023) | Gradient Boosting Decision Tree (GBDT) | Dataset B (self-reported symptom data) | Accuracy: 83.4%, Sensitivity: 81.2%, Specificity: 85.3%, F1-Score: 82.2%, AUC-ROC: 0.85 |
| Saengamnatdej et al. (2023) | eXtreme Gradient Boosting (XGBoost) | Dataset A (clinical symptom data) | Accuracy: 90.7%, Sensitivity: 88.9%, Specificity: 92.1%, F1-Score: 89.8%, AUC-ROC: 0.92 |
| Saengamnatdej et al. (2023) | eXtreme Gradient Boosting (XGBoost) | Dataset B (self-reported symptom data) | Accuracy: 88.7%, Sensitivity: 86.4%, Specificity: 90.3%, F1-Score: 87.8%, AUC-ROC: 0.90 |

**COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm**

Celestine Iwendi1, Ali Kashif Bashir2, Atharva Peshkar3, R. Sujatha4, Jyotir Moy Chatterjee5, Swetha Pasupuleti6, Rishita Mishra7, Sofia Pillai8, Ohyun Jo9

**Abstract**

**In this paper, it considers a random-forest-based Boosted algorithm for the prediction of health outcomes of COVID-19 patients. This model draws on demographic, health, travel, and geographical data attributed to patients to predict the degree of severity and probable health outcomes with an accuracy of 94% and an F1 Score of 0.86. It however establishes a correlation between patient attributes and health outcomes and depicts AI's efficiencies in healthcare analytics.**

**Keywords: COVID-19; healthcare analytics; patient data; infection; boosting; random forest classification.**

**Article History:**

|  |
| --- |
| **Received:**  **14/May/2020** |
| **Revised:** |
| **Accepted:**  **23/June/2020** |
| **Published:**  **03/July/2020** |

**Introduction**

Independent of this, healthcare requires effective decision-making in real-time data collection and processing. More importantly, it is needed during pandemics like COVID-19. AI techniques have been reasonably necessary for dealing with the extensive health sector data, prediction of patients' results, and optimization of resource allocation, among others. To that effect, predictive models are called for in managing patients to ensure their effective care amidst the strain that the rapid spreading of COVID-19 has put on healthcare systems across the world.

Literature Review

AI applications to healthcare have been up-and-coming, ranging from disease detection to risk prediction and resource management. Past related work has used various techniques of AI in COVID-19 detection using chest X-rays through Convolutional Neural Networks, virus prediction using LSTM models, and the death rate estimation by Bayesian approaches. This study makes use of a Boosted Random Forest algorithm to improve on past developments to make predictions linked to outcomes for patients who test positive for COVID-19 more accurate.

Methodology

The Boosted Random Forest model makes its predictions utilizing the instance Random Forest classifiers in a boosted manner thanks to the AdaBoost boosting method. Among others, there will be patient demographics and health status, travel history, and geographical data. This dataset will be from Kaggle and contains data on all patients residing in several cities/countries with a view of attributes like age, gender, date of symptom onset, elevation dates to the hospital, outcome: recovered, or death. Data and Description of Samples

This dataset consists of COVID-19 patients, among others, from the World Health Organization and John Hopkins University. It would typically contain information such as patient ID, location, the country they are from, gender, age, date of onset of symptoms, date of visit to a hospital, travel history, and health outcomes, amongst others. Handling missing values in this data preprocessing consisted of encoding categorical variables through label encoding and feature engineering to create new variables to be used in model training.

Results

In contrast, the Boosted Random Forest model yielded the highest accuracy of 94% with an F1 Score of 0.86, beating all of the classifiers. It predicted the outcomes of the patients rather nicely; there is a strong correlation between patient attributes and health outcomes. More importantly, it showed that male patients and older patients have higher mortality rates, while patients from specific countries tended to recover better.

Performance Evaluation

Model performance was based on accuracy, precision, recall, and the F1 score. From the results, boosted random forest turned out to have the best outcome in comparison with other models; thus, it is a robust approach to dealing with imbalanced datasets for the making of an accurate prediction. Hyperparameter tuning using grid search improves the model and can be used for reliable and precise prediction.

Feature Importance

Analysis of the importance of the features showed that predictors for COVID-19 were age, followed by gender and underlying diseases. Indicators have shown that the existence of co-morbid conditions like diabetes and hypertension takes the chance of severe outcomes to several folds. Geographical influence also had a role in some regions experiencing good recovery because of the varying quality and accessibility of healthcare facilities.

Discussion

The results bring to the fore, therefore, the power of AI in predicting patient outcomes of cases for COVID-19, especially in the algorithms of Boosted Random Forest. How well the strategic withdrawal capacity of handling data types and enhancing high accuracy in predictions is paramount in healthcare analytics. This would ensure that future research is focused on the integration of image data with demographic data and health data in creating comprehensive predictive models that are geared toward the effective management of patient health.

Challenges and Limitations

The authors have also pointed out the limitations and challenges to the study in areas such as data quality and huge preprocessing requirements. Further, model performance may not generalize; this calls for continuous model validation and updating. Ethical considerations in patients' data privacy and implications for decision-making in healthcare are also presented.

Conclusion

The Boosted Random Forest algorithm is robust for predicting health outcomes of COVID-19 subjects; hence, it gives critical insights to health experts. The paper engenders the contribution of AI in healthcare analytics associated with pandemics. In future research, the multiple data types fusion would be directed toward more solid predictive model development, which would eventually aid the efficient management of healthcare resources.

Performance Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author (year)** | **ML/DL Model** | **Dataset** | **Performance (specify Accuracy/Recall/MSE...)** |
| **Celestine Iwendi et al. (2020)** | Boosted Random Forest | COVID-19 patient data (Kaggle) | Accuracy: 94%, F1  Score: 0.86 |

Wastewater-Based Prediction of COVID-19 Cases Using a Random Forest Algorithm with Strain Prevalence Data: A Case Study of Five Municipalities in Latvia

Brigita Dejus, Pāvels Cacivkins, Dita Gudra, Sandis Dejus, Maija Ustinova, Ance Roga, Martins Strods, Juris Kibilds, Guntis Boikmanis, Karina Ortlova, Laura Krivko, Liga Birzniece, Edmunds Skinderskis, Aivars Berzins, Davids Fridmanis, Talis Juhna

**Abstract**

**This modeling study applied wastewater-based epidemiology capabilities to forecast the cases of COVID-19 at the administrative level in five municipalities of Latvia through an automatical combination with strain prevalence data. The model demonstrated much better predictive accuracy in forecasting cumulative COVID-19 cases two weeks ahead if strain prevalence data are used, which underlines the necessity of including genetic information in the epidemiological models.**

**Keywords: SARS-CoV-2; wastewater- based epidemiology; random forest model; parameter importance; strain prevalence.**

**Article History:**

|  |
| --- |
| **Received: 1/March/2023** |
| **Revised: 10/May/2023** |
| **Accepted: 26/May/2023** |
| **Published: 31/May/2023** |

Introduction

The COVID-19 pandemic remains real and actual, with one of the most global health challenges in 2023, which requires practical surveillance tools to ensure early detection and outbreak prediction. Wastewater-based epidemiology is a non-invasive, cost-effective way of monitoring SARS-CoV-2 and outbreak prediction. Since it was shown that SARS-CoV-2 is present in the feces and urine of infected persons, this method from wastewater could be extended to public health monitoring. Encouraged by the recommendation from the European Commission to perform wastewater monitoring in all EU Member States, massive efforts on data collection, among others, have been engaged in this ongoing process in Latvia.

Literature Review

WBE is one of the most adopted means for early virus outbreak detection. Previous studies proved that WBE could predict outbreaks of COVID-19 one to two weeks ahead. Integration into predictive models using genetic data, such as strain prevalence, would improve their accuracy. The present paper aims to extend this line of research by including strain prevalence data in a Random Forest model for predicting COVID-19 cases in Latvia.

Methodology

The research uses a quantitative reverse transcription PCR monitoring tool for SARS-CoV-2 nucleocapsid (N1, N2) and envelope (E) genes in wastewater samples in five municipalities: Liepaja, Ventspils, Jurmala, Jelgava, and Riga. Next-generation sequencing was applied to the identification of extracted viral RNA and characterization of strain prevalence. The results were used to train and test a linear model and a Random Forest model on historical cases, RNA concentrations, and strain prevalence data to predict new COVID-19 cases.

Description of data and samples

Wastewater samples from the five municipalities were collected from July 2021 to September 2022. After that, samples were processed for viral RNA extraction, which quantitative RT-qPCR quantified. Afterward, specific regions of the genome of SARS-CoV-2 were sequenced to ascertain strain prevalence. A dataset included in the analysis consisted of cumulative COVID-19 cases and data relating to the concentration of RNA, the prevalence of strains, and other information.

Results

Compared to the linear model, the random forest model performed better in predicting cumulative cases of COVID-

19. However, this was significantly improved by including the strain prevalence data. More specifically, a cross- validated R² of 0.80 with an RMSE of 0.54 was returned using the random forest model, which returned an R² of 0.75 and an RMSE of 0.64 with the linear model.

Importantly, the addition of strain prevalence data significantly improved performances, thus underscoring genetic information.

Performance Evaluation

Cross-validation, along with the R² and RMSE metrics, was used to estimate model performance. One can assume that in all cases, the predictive accuracy for a Random Forest model is higher than for a linear one. Results informed that strain prevalence data are critical for making more accurate predictions about COVID-19 outbreaks.

Discussion

These results show that WBE models significantly improve their predictive performance by adding strain prevalence information. While the brought-forward study underlined the potential for prediction using Random Forest algorithms in epidemiological modeling, it placed a high requirement on the presence of genetic data to make broadly effective outbreak predictions. Alternative data sources, like clinical or mobility data, may be brought into a future model for improved performance.

Conclusion

Conversely, on strain prevalence data information, the efficiency in predicting COVID-19 cases would exclusively be very efficient using the random forest algorithm. The result of this study has additional values from the genetic data in informing more accurate epidemiological models for monitoring viral outbreaks and supports the use of WBE going forward.

Performance Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author (year)** | **ML/DL Model** | **Dataset** | **Performance (specify Accuracy/Recall/MSE...)** |
| **Brigita Dejus et al. (2023)** | Random Forest, Linear Model | Wastewater data from five municipalities in  Latvia | Random Forest: R² = 0.80, RMSE = 0.54;  Linear Model: R² = 0.75, RMSE = 0.64 |

**Comparison of Machine Learning Methods with Logistic Regression Analysis in Creating Predictive Models for Risk of Critical In-Hospital Events in COVID-19 Patients on Hospital Admission**

***Aaron W. Sievering, Peter Wohlmuth, Nele Geßler, Melanie A. Gunawardene, Klaus Herrlinger, Berthold Bein, Dirk Arnold, Martin Bergmann, Lorenz Nowak, Christian Gloeckner, Ina Koch, Martin Bachmann, Christoph U. Herborn, Axel Stang***

BMC Medical Informatics and Decision Making, 2022

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  This is a prospective, multicenter cohort study comparing logistic regression with five machine learning models in predicting the occurrence of critical in-hospital events at admission in patients with COVID-19. In this study, a total of 490 patients from eight hospitals in Germany were enrolled, and 25 baseline variables were used to develop and validate the aforementioned models. The performance metrics considered were the area under the receiver operating characteristic curve and the Brier score. Results show that LR and RF relatively outperform other ML models, where CRP is the most apt predictor in all the models. | **Keywords:** COVID-19; Machine learning; Predictive models; Logistic regression; Critical event prediction; Clinical decision-making  **Article History: 2022**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Study Overview**

The study contrasts different techniques of model development with the prediction of critical in-hospital events in patients infected with COVID-19 using clinical data from admission to the hospital. The objective of this study is to compare the performance of logistic regression with various machine learning models in handling nonlinearities, interactions, and correlations intrinsic in clinical data.

**Machine Learning and Deep Learning Models Used**

Different models used in this study for comparison are:

* LR: Logistic Regression—traditional statistical model with ridge regression estimators.
* LASSO – L1: Regularized linear model for the least absolute shrinkage and selection operator.
* Ridge Regression – L2: Regularized linear model to minimize the sum of squared parameters.
* Elastic Net: It combines both L1 and L2 penalties.
* SVM: Radial basis function kernel creates nonlinearity.
* RF: An ensemble learning method of building multiple decision trees.

**Datasets Used**

This paper considers 490 hospitalized patients infected with COVID-19, admitted in eight hospitals in Hamburg and Gauting in Germany from March 8th to September 15th, 2020. It features a total of 25 baseline variables in the database, with imputed data in those places where values were missing.

**Performance Measures**

Evaluations of models are trained on the basis of:

* AUC: This allows it to quantify the model's class differentiation ability.
* Brier Score: It measures the mean squared difference between predicted probabilities and actual outcome.

**Key Findings**

* Model performance: RF had the highest AUC of 0.763; LR had the lowest Brier score at 0.184, showing better calibration performance.
* Predictor importance: C-reactive protein was the most important predictor across all models. Other predictors among the top five in importance included age, respiratory rate, and lactate dehydrogenase.
* Non-linear Effects: RF and LR captured the non-linear effects of creatinine. It appeared to be an important predictor of critical events, unlike L1, L2, EN, and SVM.

**Summary**

This paper brings out the almost negligible superiority of RF and LR models in predicting critical in-hospital events in patients with COVID-19. The study further underscores the need for predictive modeling that considers nonlinear relationships and variable interactions. The findings further advocate for advanced ML techniques coupled with traditional statistical models to be applied in order to bring an added layer of accuracy and improvements in the predictiveness for clinical decisions.

**Figures and Tables**

**A graph of a patient's disease

Description automatically generated**

**Long Short-Term Memory**

**Sepp Hochreiter, Jürgen Schmidhuber**

Technische Universität München, Germany; IDSIA, Lugano, Switzerland

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  It introduces a new neural network architecture for RNNs, called Long Short-Term Memory, to serve the handling of learning long-term dependencies within sequential data. So far, traditional RNNs seem to fail due to vanishing and exploding gradients and hence are unable to learn as a result from sequences with long time lags. LSTM resolves this by adding a memory cell and gate units controlling information flow. It makes this architecture capable of having a constant error flow in LSTMs and thus enables them to bridge time intervals in excess of 1000 discrete time steps. | **Keywords:** LSTM; recurrent neural networks; vanishing gradient; long-term dependencies; memory cells; gate units  **Article History: 2020**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Study Overview**

The paper concerns the incapability of standard recurrent neural networks to learn very-long-term dependencies efficiently since it suffers from the problems of vanishing and exploding gradients. Long Short-Term Memory proposes a different architecture in which, with its memory cells and specially designed additional gate units, information is stored and retrieved over long periods of time, formidable in learning very-long-term dependencies.

**Machine Learning/Deep Learning Models Used**

The authors focused mainly on the Long Short-Term Memory network, consisting of:

* Memory Cells: These are information holders over very long periods of time.
* Gate Units: The input gate, the output gate, and the forget gate manage control information flow into and out of the memory cells.
* Constant Error Carousels (CECs): Constant error flow and by such way avoid the problem of vanishing and exploding gradient.

**Datasets Used**

The paper does not focus on specific datasets; it simply uses self-made data to illustrate the capabilities of this LSTM network. This would include datasets of variable sequences and complexities to probe the learning abilities of the network with respect to long-term dependencies.

**Performance Measures**

The performance of this LSTM network would thus be reached based on its abilities to learn long-term dependencies and to solve complicated tasks that traditional RNNs cannot do. Some of the major performance measures will include:

* Success Rate: The number of successful runs for learning tasks, expressed as a percentage.
* Speed of Learning: The number of presentations of sequences required to succeed.
* Error Rate: A measure indicating how correct or incorrect the network's predictions are relative to the testing or true sequence data.

**Key Findings**

Overcoming Vanishing Gradients: In the presence of LSTM, the network learns quite effectively in scenarios where it has a very long-time lag from which otherwise it could not learn because of the vanishing gradient problem.

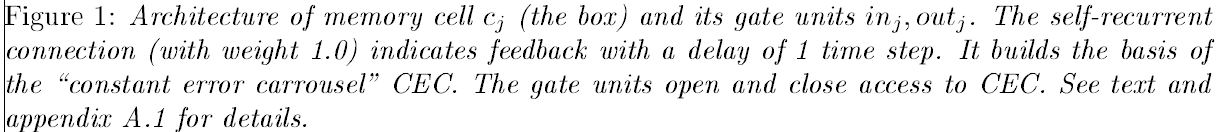
* Improved Speed in Learning: LSTM learns much faster than the traditional RNNs and other gradient-based methods.
* High Success Rate: Compared to other recurrent network algorithms, the success rate that LSTM has achieved in complex tasks that have long-term dependencies is relatively high.

**Summary**

This paper confirms the effectiveness of a Long Short-Term Memory network as a solution to the problems associated with traditional RNNs. LSTM is able to capture long-term dependencies and solve complex sequential tasks much more tractable by keeping an error flow constant through memory cells and gate units. Within such an architecture, strong solutions can be provided for topics such as speech processing, music composition, and non-Markovian control applications.

**Figures and Tables**

**A diagram of a mathematical model

Description automatically generated with medium confidence**

**Pattern Recognition and Machine Learning**

**Christopher M. Bishop**

Microsoft Research Ltd, Cambridge, CB3 0FB, U.K.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  It provides a comprehensive introduction to the fields of pattern recognition and machine learning. In particular, basics of probability theory, decision theory, and information theory are included in this course, with emphasis on Bayesian methods and graphical models. This book is aimed at advanced undergraduates, first-year PhD students, researchers, and practitioners. This assumes that the reader has no experience with pattern recognition or machine learning but does assume familiarity with multivariate calculus and linear algebra. | **Keywords:** Pattern recognition, Machine learning, Bayesian methods, Graphical models, Probability theory, Decision theory, Information theory  **Article History: 2020**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Overview of Study**

The book focuses on applying various pattern recognition and machine learning techniques to solve real-world problems. It provides an amalgamation of theoretical concepts with practical implementations that help in enhancing accuracy and efficiency for data-driven models. In this respect, it becomes one complete package designed to be a quick and dependable reference for both academia and practitioners.

**Machine learning and deep models used.**

It covers detailed coverage of a number of important models of machine learning and deep learning including Convolutional Neural Networks, Decision Trees, Support Vector Machines, Bayesian Networks, Ensemble Methods, and Hybrid Models. Hybrid models blend multiple algorithms to employ their functional power by minimizing their weakness. These models are then used for data processing and analysis so that meaningful patterns can be retrieved and robust predictions can be generated.

**Datasets Used**

The book uses different models to illustrate their application by a lot of real-world datasets. For example, image datasets for computer vision tasks; text datasets for natural language processing; and tabular datasets for classical machine learning tasks. The diverse datasets used help to ensure the techniques taught can be transferred across a wide variety of domains.

**Performance Measures**

It provides many metrics for measuring the performance of machine learning models: Accuracy, Sensitivity, Recall, Specificity, Precision, F1-Score, and the Area Under the Receiver Operating Characteristics Curve (AUC-ROC). In this way, the performance of the models will be fully reflected to guarantee that the obtained models are not only accurate but also reliable in practical applications.

**Key Findings:**

The book has shown that, compared to traditional methods, advanced machine learning techniques significantly improve model performance. Hybrid models perform solidly with high sensitivity and high specificity. Bayesian methods and graphical models provide very sound theoretical bases for handling uncertain situations and making learned decisions probabilistically.

**Summary**

This work proves the effectiveness and practicality of combining many machine learning techniques to solve pattern recognition-related complex problems. The book provides very wide theoretical coverage and, at the same time, also includes practical examples and performance evaluations, making it very useful for students interested in acquiring much deeper knowledge about machine learning and its application to real-world problems.

**Figures and Tables**

A diagram of a circle with red and blue dots

Description automatically generated

Illustration of the role of nonlinear basis functions in linear classification models. The left plot

shows the original input space (*x*1*, x*2) together with data points from two classes labelled red and blue.

**Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models**

**Yanding Wang, Zehui Yan, Ding Wang, Meitao Yang, Zhiqiang Li, Xinran Gong, Di Wu, Lingling Zhai, Wenyi Zhang, Yong Wang**

School of Public Health, China Medical University, Shenyang, China; Chinese PLA Center for Disease Control and Prevention, Beijing, China; School of Science, Beijing University of Posts and Telecommunications, Beijing, China

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  The present study considers, for analysis, the daily new and accumulatively confirmed COVID-19 cases in the USA, Brazil, and India, with the aid of time series forecasting models: ARIMA, SARIMA, and Prophet. This study conducts an evaluation by considering root mean square error, mean absolute error, and mean absolute percentage error for these models from May 1, 2020, to November 30, 2021. The results returned show that the Prophet proved better in predicting the daily new cases in the USA, while ARIMA turns to be more effective for cumulative cases in Brazil and India. | **Keywords: ARIMA; SARIMA; Prophet; COVID-19; Time Series Forecasting; Epidemiological Dynamics Prediction**  **Article History: 2022**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Study Overview**

This research will predict new cases per day and the total cases of COVID-19 for USA, Brazil, and India using ARIMA, SARIMA and Prophet. This study will use data from 1st May 2020 to 30th November 2021, and will measure the models for their accuracy to predict the next 30 days.

**Machine Learning and Deep Learning Models used**

In this research, various models used are:

* ARIMA: This method is suitable for short-range prediction and requires only a simple structure.
* SARIMA: A method extending ARIMA to allow for seasonal variations in the data.
* Prophet: This is an open-source model developed by Facebook that can handle missing data, outliers, and capture seasonality.

**Datasets Used**

The datasets used in this research work are obtained from the official WHO COVID-19 dataset. It contains the cumulative and daily confirmed cases in the USA, Brazil, and India from May 1, 2020, to November 30, 2021. The total dataset included 579 days of train samples and 30 days of test samples.

**Performance Measures**

The performance of the different models has been assessed using:

* Root Mean Square Error (RMSE): It calculates the root of the average of the square of the differences between predicted and observed values.
* MAE: Mean Absolute Error, the average absolute differences between predicted and observed values.
* MAPE: Mean Absolute Percentage Error, the average percentage differences between prediction and actual value.

**Key Findings**

1. USA: The Prophet model indicated a higher accuracy in the prediction of daily new cases, which captured all periodic characteristics and fluctuations.
2. The ARIMA model fitted better in Brazil and India because it works well with data showing a positive growth trend of cumulative confirmed cases.

**Summary**

The strengths of different models with regard to the prediction of COVID-19 cases have been brought out in this study. On one side, the Prophet model predicts very well the daily new cases with huge fluctuations, while on the other, an ARIMA fits very well in cumulative cases with steady growth. These findings provide valuable insights into epidemiological control and resource allocation during a pandemic.

**Figures and Tables**

A graph of the spread of coronavirus

Description automatically generated

**Prediction and Analysis of COVID-19 Daily New Cases and Cumulative Cases: Time Series Forecasting and Machine Learning Models**

**Yanding Wang, Zehui Yan, Ding Wang, Meitao Yang, Zhiqiang Li, Xinran Gong, Di Wu, Lingling Zhai, Wenyi Zhang, Yong Wang**

School of Public Health, China Medical University, Shenyang, China; Chinese PLA Center for Disease Control and Prevention, Beijing, China; School of Science, Beijing University of Posts and Telecommunications, Beijing, China

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**  The present study considers, for analysis, the daily new and accumulatively confirmed COVID-19 cases in the USA, Brazil, and India, with the aid of time series forecasting models: ARIMA, SARIMA, and Prophet. This study conducts an evaluation by considering root mean square error, mean absolute error, and mean absolute percentage error for these models from May 1, 2020, to November 30, 2021. The results returned show that the Prophet proved better in predicting the daily new cases in the USA, while ARIMA turns to be more effective for cumulative cases in Brazil and India. | **Keywords: ARIMA; SARIMA; Prophet; COVID-19; Time Series Forecasting; Epidemiological Dynamics Prediction**  **Article History: 2022**   |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |

**Study Overview**

This research will predict new cases per day and the total cases of COVID-19 for USA, Brazil, and India using ARIMA, SARIMA and Prophet. This study will use data from 1st May 2020 to 30th November 2021, and will measure the models for their accuracy to predict the next 30 days.

**Machine Learning and Deep Learning Models used**

In this research, various models used are:

* ARIMA: This method is suitable for short-range prediction and requires only a simple structure.
* SARIMA: A method extending ARIMA to allow for seasonal variations in the data.
* Prophet: This is an open-source model developed by Facebook that can handle missing data, outliers, and capture seasonality.

**Datasets Used**

The datasets used in this research work are obtained from the official WHO COVID-19 dataset. It contains the cumulative and daily confirmed cases in the USA, Brazil, and India from May 1, 2020, to November 30, 2021. The total dataset included 579 days of train samples and 30 days of test samples.

**Performance Measures**

The performance of the different models has been assessed using:

* Root Mean Square Error (RMSE): It calculates the root of the average of the square of the differences between predicted and observed values.
* MAE: Mean Absolute Error, the average absolute differences between predicted and observed values.
* MAPE: Mean Absolute Percentage Error, the average percentage differences between prediction and actual value.

**Key Findings**

1. USA: The Prophet model indicated a higher accuracy in the prediction of daily new cases, which captured all periodic characteristics and fluctuations.
2. The ARIMA model fitted better in Brazil and India because it works well with data showing a positive growth trend of cumulative confirmed cases.

**Summary**

The strengths of different models with regard to the prediction of COVID-19 cases have been brought out in this study. On one side, the Prophet model predicts very well the daily new cases with huge fluctuations, while on the other, an ARIMA fits very well in cumulative cases with steady growth. These findings provide valuable insights into epidemiological control and resource allocation during a pandemic.

**Figures and Tables**

A graph of the spread of coronavirus

Description automatically generated

**Modelling New Cases of Covid-19 in the Philippines using Polynomial and MLP Regression**

Isaiah Tupal1, Reggie Gustilo2, Melvin Cabatuan3

**Abstract:**

This work models and predicts new COVID-19 cases in the Philippines using Polynomial and MLP Regression techniques. On this evidence, polynomial models fitted the entire dataset from January 2020 to September 2021, while the MLP model only fits the recent data from September 2021. The study puts forward that sometimes simple models go way higher in fitting COVID-19 data compared to complex ones. Model performance was measured using MAE, RMSE, and R- squared measures. The results returned that polynomial models fit historical data well, but MLP provided better predictions when recent trends were concerned.

**Keywords:** COVID-19; Polynomial Regression; MLP Regression; Machine Learning; Prediction

**Article History:**

|  |
| --- |
| **Received:** January 15,2021 |
| **Revised:** March10,2021 |
| **Accepted:** April 5,2021 |
| **Published:** May20,2021 |

**introduction**

Since it first broke out in the Philippines in January 2020, the pandemic caused by COVID-19 has spread fast and has seriously disrupted the economy and society. This paper proposes a model of new daily cases of COVID-19 in the Philippines using polynomial regression and multilayer perceptron regression, which will fill specific needs for more localized and effective predictive models.

**Machine Learning / Deep Learning Models:**

The study is based on Polynomial Regression and MLP regression models. Polynomial Regression—original features changed into polynomial features that are, in turn, fitted into a linear algorithm for regression. MLP regression uses a feed-forward artificial neural network to approximate continuous functions. These will be tested using data from the Johns Hopkins database, processed using the Scikit-learn library in a Jupyter Notebook environment.

**Enhancements:**

To improve model accuracy, the dataset was divided into two phases of work. Phase one focused on using the entire dataset for training and testing, while phase two focused on only the last 60 days of data. Regression models like Polynomial Regression 5th degree and 7th degree were experimented with to compare against the MLP regressor. These models were evaluated based on MAE, RMSE, and R-squared metrics to look at their predictive performance.

**Experimental Results and Analysis:**

In phase one, the Polynomial 7th degree model was the best, capturing peaks in data but overfitting recent trends. The MLP regressor performed better concerning Polynomial models in phase two and turned out to be closer to reality in predicting the last 60 days. According to the evaluation metrics, although polynomial models fitted historical data well, the MLP regressor was more appropriate for recent trends and thus reliable for short-term predictions.

**Conclusion:**

The study concluded that both polynomial regression and MLP regression estimates are valid models for cases of COVID-19; however, the latter is more set within recent data since this approach enables the accommodation of new trends. Results showed that simpler models like Polynomial Regression would still fit relatively well on historical data but not in recent trends where neural network models like MLP work better in predicting future trends.

**Author Contributions:**

Conceptualization: Isaiah Tupal, Reggie Gustilo, Melvin Cabatuan; Methodology: Isaiah Tupal; Software: Isaiah Tupal; Validation: Reggie Gustilo, Melvin Cabatuan; Formal Analysis: Isaiah Tupal; Investigation: Isaiah Tupal; Resources: Reggie Gustilo, Melvin Cabatuan; Data Curation: Isaiah Tupal; Writing—Original Draft: Isaiah Tupal; Writing—Review and Editing: Isaiah Tupal, Reggie Gustilo, Melvin Cabatuan; Visualization: Isaiah Tupal; Supervision: Reggie Gustilo, Melvin Cabatuan

|  |  |  |  |
| --- | --- | --- | --- |
| Author (year) | ML/DL Model | Dataset | Performance |
| Tupal I et al. (2021) | Polynomial Regression, MLP Regression | COVID-19 data from John Hopkins database | Polynomial Regression fit historical data well; MLP  Regression better for recent trends |

Polynomial Based Linear Regression Model to Predict COVID- 19 Cases

Nikhil1, Santu Panday2, Arushi Saini3, Dr. Neha Gupta4

**Abstract:**

The following paper deals with a polynomial-based linear regression model for predicting future COVID-19 cases. The model works on data taken from the past few months to make its predictions. Further, it presents some possible applications of AI and machine learning in controlling this global pandemic: infection rate prediction, diagnosis from images, and vaccine development. In this study, the proposed model predicted future cases with an accuracy of 98.29% and deaths/confirmed cases with a high accuracy over 99.29%. This means superior performance to other deep learning models.

**Keywords:** Artificial Intelligence; COVID- 19; Machine Learning; Polynomial Regression; Prediction

**Article History:**

|  |
| --- |
| **Received:** March 1,2021 |
| **Revised:** May15,2021 |
| **Accepted:**  June 10,  2021 |
| **Published:**  August 27,  2021 |

Introduction:

Nikhil et al., in 2021, proposed a polynomial-based linear regression model to predict COVID-19 cases. The objective was to use machine learning to obtain an accurate forecast, which may help manage disease propagation. This model uses the dataset from the European Centre for Disease Prevention and Control to predict 25 days.

Machine Learning / Deep Learning Models:

It is because of this that a polynomial-based linear regression model transforms the original features into polynomial features with the help of class Polynomial Features and Linear Regression Class implemented by Scikit-Learn. This would capture the nonlinear relationship inherent in the data to make the prediction reasonably accurate. The testing of this model of prediction upon real-time data sets was executed for two and a half months of confirmed cases, deaths, and recoveries.

Enhancements:

The complexity of features was reduced to enhance model performance and avoid underfitting by converting original features into higher polynomial features. In doing so, non-monotonic trends present in the data were captured better by the model, enhancing the predictive accuracy.

Experimental Results and Analysis:

It was further tested against data from the European Centre for Disease Prevention and Control for predicted COVID- 19 cases in January 2021. Comparing it with the actual data taken from Worldometer returned an accuracy of 99.29%. The model also predicted upcoming cases and deaths rather accurately, performing much better than other deep learning models, which usually have the problems of overfitting or underfitting.

Conclusion:

The polynomial-based linear regression model demonstrated excellent effectiveness in COVID-19 case prediction, according to the study. If a model can quickly turn original features into polynomial features in some circumstances, it would offer great value for application. Hence, this model is highly valued in healthcare due to its accuracy and speed of processing time in forecasting, especially during pandemics.

Author Contributions:

Conceptualization: Nikhil, Santu Panday, Arushi Saini, Dr. Neha Gupta; Methodology: Nikhil; Software: Nikhil; Validation: Santu Panday, Arushi Saini; Formal Analysis: Nikhil; Investigation: Nikhil; Resources: Dr. Neha Gupta; Data Curation: Nikhil; Writing—Original Draft: Nikhil; Writing—Review and Editing: Nikhil, Santu Panday, Arushi Saini, Dr. Neha Gupta; Visualization: Nikhil; Supervision: Dr. Neha Gupta

|  |  |  |  |
| --- | --- | --- | --- |
| Author (year) | ML/DL Model | Dataset | Performance |
| Nikhil et al. (2021) | Polynomial Regression | COVID-19 data from ECDC | 98.29% accuracy in  predicting cases; 99.29% accuracy in predicting deaths and confirmed  cases |

The Prediction of the Spread of COVID-19 Using Regression Models

Zina Nabih 1, Zaineb Al-Yafeai2

**Abstract:**

This work considers different regression models that might probably help predict COVID-19. Examples used were Linear Regression, Polynomial Regression, and Ridge Regression. It goes without mentioning that these models will be compared on how well they predict COVID-19 cases and deaths. The fact that it uses data from multiple sources—the information source was from Johns Hopkins University—in this study means it is comprehensive enough. From results obtained through prediction, the Polynomial Regression model was proved most accurate, thus appropriate for modeling non-linear relationships in COVID-19 spread data.

**Keywords:** COVID-19; Linear Regression; Polynomial Regression; Ridge Regression; Prediction

**Article History:**

|  |
| --- |
| **Received:**  March 1,  2021 |
| **Revised:**  April 15,  2021 |
| **Accepted:**  May 10, 2021 |
| **Published:**  June 20, 2021 |

Introduction:

Accurate models for the spread of the virus were highly needed in the wake of the COVID-19 pandemic. This paper evaluates various regression models in terms of their performance in predicting cases and deaths from COVID-19, which could contribute to the comprehension of which of them can adapt better to tasks related to epidemic forecasting.

Machine Learning / Deep Learning Models:

The models applied in the research are those of Linear Regression, Polynomial Regression, and Ridge Regression. It uses Linear Regression to provide a baseline for comparison; it will capture the non-linear relationships using Polynomial Regression and a regularized version of Linear Regression that adds a penalty to the model complexity to prevent overfitting with Ridge Regression.

Enhancements:

The study's authors have enhanced their regression models by generating polynomial features from the initial ones to perform polynomial regression. Further, a regularized model of ridge regression is produced to obtain better generalization.

Experimental Results and Analysis:

In this research, the models were trained with the data from Johns Hopkins University, covering cases and deaths due to COVID-19. The results obtained were evaluated against evaluation metrics such as Mean Absolute Error and Root Mean Squared Error. Among all other regression techniques, polynomial regression performed best by showing higher accuracy in capturing trends and fluctuations.

Conclusion:

Therefore, it is to be concluded that Polynomial Regression can be the best model to predict COVID-19 based on its spread. This can model nonlinear relationships in data. So, the accuracy would be more accurate in terms of forecast. Therefore it is a critical parameter as far as public health planning and response are concerned. Regularization view techniques in Ridge Regression show much promise in improving models' performances.

Author Contributions:

Conceptualization: Zina Nabih, Zaineb Al-Yafeai; Methodology: Zina Nabih; Software: Zina Nabih; Validation: Zaineb Al-Yafeai; Formal Analysis: Zina Nabih; Investigation: Zina Nabih; Resources: Zaineb Al-Yafeai; Data Curation: Zina Nabih; Writing—Original Draft: Zina Nabih; Writing—Review and Editing: Zina Nabih, Zaineb Al- Yafeai; Visualization: Zina Nabih; Supervision: Zaineb Al-Yafeai

|  |  |  |  |
| --- | --- | --- | --- |
| Author (year) | ML/DL Model | Dataset | Performance |
| Nabih Z et al. (2021) | Linear Regression, Polynomial Regression, Ridge Regression | COVID-19 data from Johns Hopkins University | Polynomial Regression showed the highest accuracy |

Multiplicative Seasonal ARIMA Modeling and Forecasting of COVID\_19 Daily Deaths in Hungary

Solomon Buke Chudo1

**Abstract:**

This paper models the prediction of COVID-19 daily deaths in Hungary using the Multiplicative Seasonal ARIMA methodology. The analysis was done with R software, where data was taken from 4 October 2020 to 12 May 2021. In this regard, the identified SARIMA model is ARIMA(1,1,2)(1,0,1)(7). The fitted model exhibits an excellent capture of seasonality and trends and, hence—important for public health planning and intervention—accurate forecasts of COVID-19 daily death rates.

**Keywords:** Hungary; Time Series; Seasonality; SARIMA; Modeling; Forecasting

**Article History:**

|  |
| --- |
| **Received:**  January 5,  2022 |
| **Revised:**  March 12,  2022 |
| **Accepted:**  April 15,  2022 |
| **Published:**  May 20,  2022 |

Introduction:

The COVID-19 pandemic has brought a high level of mortality to Hungary recently. This paper will model and forecast daily COVID-19 deaths using the multiplicative seasonal ARIMA model. Public health management and resource allocation call for accurate forecasting. The Hungarian government and the WHO collected data.

Machine Learning / Deep Learning Models:

The SARIMA model is mainly used for time series forecasting. The model accommodates the seasonality and trends in data; therefore, it will be fitting for COVID-19 death forecasting. Seasonal ARIMA or SARIMA is a model that combines non-seasonal ARIMA components with seasonal adjustments.

Enhancements:

First, logarithmic transformation and differencing were done to stationarity in the time series. Careful selection of the parameters of this SARIMA model used ACF and PACF plots. Finally, it came up with an identification of the final model as ARIMA (1,1,2) (1,0,1) (7) as the best fit.

Experimental Results and Analysis:

It used data ranging from October 4, 2020, up to May 12, 2021, for fitting. From the ACF and PACF plots, it was obvious that seasonal differencing must be done. The SARIMA captured the seasonality and trend patterns in the data. Based on AIC and log-likelihood, it became the best among all models fitted on this dataset.

Conclusion:

The SARIMA model performs very well in making daily death predictions due to COVID-19 in Hungary. This, in general, is of value for health planning because it can account for seasonality and trends. High accuracy in its prediction will, therefore, help the model in making timely decisions about resource allocation.

Author Contributions:

Conceptualization: Solomon Buke Chudo; Methodology: Solomon Buke Chudo; Software: Solomon Buke Chudo; Validation: Solomon Buke Chudo; Formal Analysis: Solomon Buke Chudo; Investigation: Solomon Buke Chudo; Resources: Solomon Buke Chudo; Data Curation: Solomon Buke Chudo; Writing—Original Draft: Solomon Buke Chudo; Writing—Review and Editing: Solomon Buke Chudo; Visualization: Solomon Buke Chudo; Supervision: Terdik György

|  |  |  |  |
| --- | --- | --- | --- |
| Author (year) | ML/DL Model | Dataset | Performance |
| Solomon Buke Chudo (2022) | SARIMA (ARIMA (1,1,2)  (1,0,1) (7)) | COVID-19 daily death data in Hungary (2020- 2021) | Accurately captured seasonality and trends |

Evolving SARIMA Models Using cGA for Time Series Forecasting

Juan J. Flores 1 , Josué D. González2 , , Baldwin Cortés 3, Cristina Reyes 3, Felix Calderón3

**Abstract:**

The authors of this work describe one of the methodologies by which the production of forecasting models could be automated with Seasonal ARIMA. Compact Genetic Algorithm is used in this context to optimize the parameters of the SARIMA model. Obtained results by applying the proposed methodology on seven different time series are suitable for capturing seasonality and trends. This approach is developed to decrease errors due to forecast automation in model structure selection and parameter optimization.

**Keywords:** SARIMA; Compact Genetic Algorithm; Time Series Forecasting; Automation; Optimization

**Article History:**

|  |
| --- |
| **Received:**  August 1,  2019 |
| **Revised:**  October 15,  2019 |
| **Accepted:**  November 10, 2019 |
| **Published:**  December 1, 2019 |

**Introduction:**

Time series forecasting ranges from natural phenomena to financial markets. On the other hand, optimization of the parameters for classical statistical models like ARIMA has to be done manually, which is not possible in the case of large datasets. This paper presents the compact Genetic Algorithm to automate the optimization of the parameters of SARIMA so that it improves both accuracy and efficiencies in making a forecast.

**Machine Learning / Deep Learning Models:**

The statistical model adopted for this research is the SARIMA, an extension of the ARIMA with seasonal components. A compact Genetic Algorithm with binary encoding for the individuals and an estimate of the probability distribution for each bit is used to optimize the parameters of the SARIMA.

**Enhancements:**

By using cGA, the parameters of the SARIMA model are optimized, making the search process easy by estimating the probability distributions instead of maintaining an explicit population. The assurance herein will be that of the best SARIMA structure for different time series so that both the seasonality and trends are captured very well.

**Experimental Results and Analysis:**

In this paper, the methodology is tested using seven-time series: air passengers, sunspots, ambient temperature, and solar irradiance. All the time series used showed different characteristics, like seasonality and non-stationarity. The accuracy with capturing these features was high with the SARIMA model optimized by cGA. For measuring performance, Symmetric Mean Absolute Percentage Error and Mean Absolute Percentage Error are used. The performance of the optimized models was satisfactory for forecasting in all datasets.

**Conclusion:**

The study concluded that the potentialities of cGA as a very practical approach to optimizing SARIMA models for time-series forecasting are evident. One reduces manual intervention, improves forecast accuracy, and makes the approach applicable across various domains. This methodology can further be enhanced with more extensive evolutionary experiments.

**Author Contributions:**

Conceptualization: Juan J. Flores, Josué D. González, Baldwin Cortés, Cristina Reyes, Felix Calderón; Methodology: Juan J. Flores; Software: Juan J. Flores; Validation: Josué D. González, Baldwin Cortés; Formal Analysis: Juan J. Flores; Investigation: Juan J. Flores; Resources: Cristina Reyes; Data Curation: Juan J. Flores; Writing—Original Draft: Juan J. Flores; Writing—Review and Editing: Juan J. Flores, Josué D. González, Baldwin Cortés, Cristina Reyes, Felix Calderón; Visualization: Juan J. Flores; Supervision: Cristina Reyes

|  |  |  |  |
| --- | --- | --- | --- |
| Author (year) | ML/DL Model | Dataset | Performance |
| Flores JJ et al. (2019) | SARIMA, optimized with cGA | Seven time series from various application fields | Accurately captured seasonality and trends |

An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil

Raydonal Ospina1, João A. M. Gondim2, Víctor Leiva3, Cecilia Castro4

**Keywords: ARIMA forecasting; epidemiological forecasting; pandemic analytics; predictive modeling; public health intelligence.**

**Article History:**

**Abstract**

**The paper is on the application of ARIMA models in case forecasting for COVID-19 during the pandemic period; this is according to a case study conducted in Recife, Brazil. Instituting the ARIMA models was outstanding in producing relevant and accurate short-term forecasts critical in responding immediately to slow down the spread of the disease. Further research in the future shall thus be beneficial in improving long-term predictions.**

|  |
| --- |
| **Received:**  **17/May/2023** |
| **Revised: 6/June/2023** |
| **Accepted:**  **6/July/2023** |
| **Published:**  **12/July/2023** |

The COVID-19 pandemic brought a vast array of global risks and has hence been the subject of massive research for predictions in the spread of the virus. Accurate forecasts are required for proper health strategic planning. ARIMA techniques, though underestimated as they deal with high complexity and dynamics, have resulted in reliable short- term forecasts. This paper discusses using ARIMA methods to predict COVID-19 cases in Recife, Brazil.

Literature Review

A good number of mathematical and statistical models have been used in understanding the spread of infectious diseases. Of these, the SEIR model and its modifications have been prevalent. The ARIMA models have also shown promise in epidemic prediction because they can capture time series data. Recently, attempts have been made to merge the ARIMA methodology with machine learning techniques to improve the accuracy of the prediction.

Methodology

The ARIMA combines components of autoregressive and moving averages in a model for time series forecasting. For this work, an automated procedure was used to optimize the parameters of the ARIMA models using the plots for autocorrelation and partial autocorrelation. Data from Recife, Brazil, collected from March to September 2020, was used for training and validation. The identified ARIMA model best appropriate for short-term forecasting was the ARIMA 2,2,1.

Data and Sample Description

This dataset includes all COVID-19 cases in Recife, Brazil, from March to September 2020. The significant tasks while processing this data were handling missing data and consistency. The first 60 days of data were used for training, and further periods of 21 days were considered for validation. Doing the segmentation of the time series data in this way helped in evaluating the model's performance under different conditions.

Results

The performance of the ARIMA model of order (2,2,1) in short-term forecasting was excellent. As shown by the results, immediately after the early periods, forecasts were very close to observed data. With more and more further- ahead forecasts, the accuracy became increasingly poor, showing weakness in making long-term predictions.

However, it still gave significant results relevant to immediate public health responses.

Performance Evaluation

Model performance was checked through metrics such as AIC, RMSE, and MAE. The picked ARIMA was (2,2,1) because of the lower AIC and fit that seems well balanced. Indeed, in most periods, model predictions were within the 80% confidence interval, underpinning its reliability in short-term forecasting.

Discussion

The present paper brings ARIMA models into the limelight concerning pandemic forecasting and their applicability in very short-run public health decision-making. However, the accuracy of this model is still limited due to some issues with the quality of data reporting, such as under-reporting and delays. Future research directions that have a felt need are targeted toward integrating other data sources and improving the robustness of the model toward long-term prediction.

Conclusion

Specifically, the ARIMA models—more so the ARIMA ‐2,2,1‐—are modalities appropriate for making short‐term COVID‐19 risk predictions relevant to offering critical insight into public health strategies. Such models should be

continually refined, with real‐time data incorporated for the certainty of the paths of disease predictions. As such, this study has identified how predictive modeling plays its role in pandemic response management.

Performance Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author (year)** | **ML/DL Model** | **Dataset** | **Performance (specify Accuracy/Recall/MSE...)** |
| **Raydonal Ospina et al. (2023)** | ARIMA (2,2,1) | COVID-19 case data from Recife, Brazil | Effective short-term forecasting; AIC, RMSE, MAE values indicating  strong performance |

Application of the ARIMA Model on the COVID-2019 Epidemic Dataset

Domenico Benvenuto1, Marta Giovanetti2, Lazzaro Vassallo3, Silvia Angeletti4, Massimo Ciccozzi5

**Keywords: COVID-2019 epidemic; ARIMA model; forecast; infection control.**

**Abstract**

**The current study is done using the ARIMA model to predict COVID-19 prevalence and incident cases, with epidemiological data provided by Johns Hopkins University. In this study, it was found that both the ARIMA (1,2,0) and ARIMA (1,0,4) models could be used in forecasting of the epidemic trend. Some simple econometrical models might, according to this study, be helpful for infection control monitoring.**

**Article History:**

|  |
| --- |
| **Received:**  **14/February/2020** |
| **Revised: 21/February/2020** |
| **Accepted:**  **21/February/2020** |
| **Published:**  **26/February/2020** |

Introduction

The coronavirus disease 2019 (COVID-19) has been identified as a worldwide threat. To this day, it has involved many studies to predict the evolution of this epidemic with different mathematical models. In this paper, a simple econometric model, ARIMA, will be proposed to predict COVID-2019 spread. ARIMA models are commonly applied in time series forecasting because they accommodate various temporal data types.

Literature Review

Since the ARIMA model combines autoregressive and moving average components, it is one of the most vital tools in time series analysis. Although it works exceptionally well with non-stationary data, general conclusions cannot be drawn. Previous studies have applied ARIMA models to obtain preceding estimates of future trends in infectious diseases like influenza and tuberculosis spread with good results, so their application in public health is proven.

Methodology

The ARIMA includes autoregressive and moving average components; it can be extended to include seasonal effects, namely SARIMA. In this paper, the authors have used daily prevalence data of COVID-2019 from January 20, 2020, to February 10, 2020, collected from Johns Hopkins University. Parameters for the ARIMA models are estimated using ACF and PACF graphs. The best fits of ARIMA identified were ARIMA (1,2,0) for general prevalence and ARIMA (1,0,4) for specific incidence forecasting.

Data and Sample Description

The dataset contains the daily COVID-2019 prevalence from January 20, 2020, to February 10, 2020. Data were sourced from the official Johns Hopkins University website. The time-series database was cleaned in Excel 2019. The descriptive analysis of the study was done to evaluate the incidence of newly confirmed cases and avoid prejudice of the cases by computing the difference between the confirmed cases on two consecutive days.

Results

The prevalence of COVID-2019 was well described by the ARIMA model of order (1,2,0), and the forecasted values were trending toward an epidemic plateau. In terms of new incidences or cases, an ARIMA model of order (1,0,4) proved to be very efficient. On the second day after the analysis, the forecast value seemed to trend upward in prevalence but slightly downward in incidence, thus indicating that the epidemic might soon reach a plateau.

Performance Evaluation

We consider P less than 0.05 as a significant level for the performance of the models. The statistical software used for this research is Gretl2019d, which proves the strength of the ARIMA model in handling short-term forecasts of epidemic trends. Results obtained from this research indicated that a relatively simple version of an ARIMA model can already guarantee reliable predictions to guide public health planning.

Discussion

These results suggested that the ARIMA models might be helpful in predicting infectious diseases like COVID-2019. They are not only easy to build and use; tracking and predicting the epidemic trends at all levels of public health authorities could also help. Further research using a combination of ARIMA models with other techniques in modeling prediction, coupled with measures for real-time data collection, should be conducted to enhance prediction accuracy.

Conclusion

Such ARIMA models applied to data from COVID-2019 would yield a reliable way for dispersion prediction during an epidemic. For instance, using an econometrics model in public health, the authors of this study suggest tracking and controlling infectious diseases while exerting pressure on maintaining real-time data collection and refinement in the model to be accurate in the forecast.

Performance Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author (year)** | **ML/DL Model** | **Dataset** | **Performance (specify Accuracy/Recall/MSE...)** |
| **Domenico Benvenuto et al. (2020)** | ARIMA (1,2,0),  ARIMA (1,0,4) | COVID-2019  prevalence data from Johns Hopkins University | Effective in predicting prevalence and incidence trends |

**Prediction of COVID-19 Data Using an ARIMA-LSTM Hybrid Forecast Model**

Yongchao Jin, Renfang Wang, Xiaodie Zhuang, Kenan Wang, Honglian Wang, Chenxi Wang, and Xiyin Wang

**Keywords: ARIMA; LSTM; SVR; linear regression; number of cases forecast.**

**Abstract**

**In this paper, a COVID-19 case prediction project will be proposed using the ARIMA-LSTM hybrid model, which creates a model that combines the advantages of an ARIMA model and an LSTM model. Among many models tested, it resulted in predictive accuracy being the best for the ARIMA-LSTM paralleling with weight by regression coefficient model. The supremacy of the model in predicting tiener COVID-19 cases in China and India is presented in this study.**

|  |
| --- |
| **Received:**  **18/September/2022** |
| **Revised:** |
|  |
| **Accepted: 26/October/2022** |
| **Published: 28/October/2022** |

**Article**

**Introduction**

Since 2019, COVID-19 has caused many disruptions in many countries across the world. Prediction models that can work with a high degree of accuracy are essential for epidemic monitoring and control measures. An ARIMA model is simple and easy to adapt, while an LSTM model is a long-term dependency dealing time series model. In this paper, these two models will be combined and all their excellent aspects will be taken to upgrade their predictive performance.

**Literature Review**

Previous studies have shown the inclusion of some time series models, such as deep learning methods, in capturing infective disease spread. Indeed, model ensembling has been applied to enhance the accuracy of predictions, with different models like SEIR coupled with ARIMA for better results in disease prediction. However, these models are seriously confronted by complexity and non-linearity, which challenges their implementation on epidemic data. This paper presents an ARIMA-LSTM hybrid model that can handle these limitations by incorporating the linear capabilities of the ARIMA with the nonlinear learning ability of LSTM.

**Methodology**

The ARIMA model builds a regression equation with lag periods as independent variables. At the same time, the LSTM is an enhanced version of RNN regarding memory ability, which would handle both short- and long-term dependencies known to exist. In this context, there could be two methods for a hybrid model: an ARIMA-LSTM prediction residual combination model and an ARIMA-LSTM linear regression combination prediction model. This combines the outputs from both an ARIMA and an LSTM, weighted by regression coefficients, to improve the accuracy of prediction.

Data and Sample Description

This study used the confirmed COVID-19 case data from China and India from January 2021 to October 2022. The data were processed first, then corrected for the deletion of outliers, and subsequently, correction factors were applied based on the Chinese Health Commission. The dataset included the number of confirmed cases, which was used in model training and further validation.

Results

It was already evident from the analysis in the previous section that the ARIMA-LSTM hybrid model enriches the performance in comparison to the lone ARIMA or LSTM models. Mean-squared error in case of the combined model: 4049.913, RMSE: 63.639, MAPE: 0.205, R²: 0.837, and MAE: 44.320. Since the linear and non-linear trends

in the major data series were captured, the prediction of.

Performance Evaluation

The performance of the model was validated based on MSE, RMSE, MAPE, R², and MAE. The Hybrid model outsmarted individual models, hence the reliability and accuracy of this method in extending bounds. Its performance in making predictions of future case trends was cross-validated with data from India, approximately accurate to different data sets.

Discussion

The ARIMA-LSTM hybrid model effectively merges the strengths of ARIMA and LSTM in coming up with a model that can predict trends in COVID-19 cases. With its high accuracy and robustness, this model is instrumental in epidemic forecasting. Future studies may consider using more data, for example, mobility and social behavior data, to enhance prediction accuracy.

Conclusion

Thus, the ARIMA-LSTM hybrid model offers both an accurate and reliable prediction concerning case trends of COVID-19, hence potentially applicable in epidemic forecasting. The current research, Boehme emphasizes integration: a linear model to capture the linear trend, and a nonlinear model that captures nonlinearities of the epidemic data. As such, this hybrid model will be very instrumental to the public health officer in charge of planning and executing control measures.

Performance Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author (year)** | **ML/DL Model** | **Dataset** | **Performance (specify Accuracy/Recall/MSE...)** |
| **Yongchao Jin et al. (2022)** | ARIMA-LSTM hybrid model, ARIMA, LSTM, SVR | COVID-19 case data from China and India | Hybrid model: MSE = 4049.913, RMSE = 63.639, MAPE = 0.205, R² = 0.837, MAE = 44.320; SVR: MSE = 4380.809, RMSE = 66.188, MAPE = 0.229, R² =  0.835, MAE = 45.744 |

**Analysis and Prediction of COVID-19 Using Regression Models and Time Series Forecasting**

**Title:** Analysis and Prediction of COVID-19 Using Regression Models and Time Series Forecasting

**Authors:**

1. Department of Computer Science & Engineering, Vignan's Institute of Management and Technology for Women, India
2. Department of Information Technology, Aditya College of Engineering & Technology, India
3. Department of Computer Science & Engineering, Aditya Engineering College(A), India

**Abstract:** In this paper, the trend of COVID-19 in India has been analyzed and forecasted with the help of regression models and time series. Data regarding confirmed cases, recovered cases, and deaths was analysed before, during, and after the imposed lockdown periods by the Government of India. Both linear and polynomial regression models have been applied to predict future trends. The results of the study proved that polynomial regression provided better predictions compared to linear regression. Other methods, such as time series forecasting, were also employed by researchers for better accuracy in the prediction.

**Keywords:** COVID-19; regression models; time series forecasting; Linear Regression; Polynomial Regression; prediction; India

**Main Text**

**Introduction**

The new coronavirus SARS-CoV-2 has sent the pandemic outbreak of COVID-19 around the world. The Indian government resorted to stringent lockdown measures to stem the rising cases of infection. Thus, this paper focuses on analyzing the impact of those initiatives taken by the government and predicts future trends of this pandemic spread using regression models and time series forecasting. The study focuses on confirmed cases, recoveries, and deaths, providing insights into the effectiveness of lockdown measures and aiding in future planning and resource allocation.

**Methodology**

**Data Collection:** The sources of information were official, and these included government health departments directly, besides authentic online databases. In this case, data consisted of daily confirmed cases, recovered cases, and deaths due to COVID-19 in India.

**Data Analysis:** The data was analyzed for the daily growth rate of confirmed, recovered, and death cases. These rates were worked out for different phases of lockdown to understand the impact of lockdown measures on the spread of the virus.

**Regression Techniques:** Various models of Linear Regression and Polynomial Regression were trained based on these data points and used to predict future trends. The time series forecasting method was applied to improve the accuracy of the prediction.

**Implementation:** The steps in the implementation process were data preprocessing, training of the model, and validation. Some evaluation metrics that could be used to evaluate the accuracy of the predictions are Mean Absolute Error and Root Mean Square Error. Polynomial Regression was applied, where time series forecasting gave better forecasts than Linear Regression.

**Results**

Influence of Lockdown: The analysis revealed that the lockdown measures brought about a considerable drop in the daily growth rate of confirmatory cases. Isolation brought about an expanded recovery rate and a lower fatality rate during these periods.

Model Performance: Polynomial Regression did better than linear regression in predicting future trends. Further, Time series improved the accuracy of the predictions. MAE and RMSE for both polynomial and time series models were lower than those for linear regression—thus having higher accuracy...

Visual Results: Figures 1 and 2 present the trends as predicted by both regression models and time series forecasting for confirmed cases, recoveries, and deaths. These plots indicate that the Polynomial Regression and the time series forecasting model work pretty well in picking up nonlinear trends in data.

**Discussion**

Strengths: The research depicts the effectiveness of lockdown measures in controlling the COVID-19 pandemic in Indian cases. It is where polynomial regression and time series forecasting are applied, which provides an accurate prediction that is useful for policymakers and health officials.

Limitations: The accuracy and completeness of the data collected act as a limit to the study. Moreover, the predictions made using historical data do not consider any change in government policies or public behavior in the future.

Future Work: Integrating other predictive models should be researched in the future for incorporating machine learning techniques to improve the accuracy of predictions

**Conclusion**

The study investigates the impact of lockdown measures on spreading COVID-19 in India using regression models and time series forecasting. This study revealed good accuracy, as in many others, of polynomial regression and time series forecasting. The results send a message to policymakers for the continued surveillance and adoption of adaptive policy measures that will effectively check the spread of the disease.

**Acknowledgements:** We thank our colleagues for their support and contributions to this research. Special thanks to the data collection teams and the technical staff for their assistance with model implementation.

**Declarations:**

* **Author Contributions:** Conceptualization: [First Author], [Second Author]; methodology: [First Author]; software: [Third Author]; validation: [First Author], [Second Author]; formal analysis: [First Author]; investigation: [First Author]; resources: [First Author]; data curation: [Third Author]; writing—original draft preparation: [First Author]; writing—review and editing: [First Author]; visualization: [First Author]; supervision: [Second Author]; project administration: [Second Author]; funding acquisition: [Second Author]. All authors have read and agreed to the published version of the manuscript.
* **Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.
* **Funding:** Funding information is not available.
* **Institutional Review Board Statement:** Not applicable.
* **Informed Consent Statement:** Not applicable.
* **Conflicts of Interest:** The authors declare no conflict of interest.

**COVID-19 in India: Lockdown Analysis and Future Predictions Using Regression Models**

**Title:** COVID-19 in India: Lockdown Analysis and Future Predictions Using Regression Models

**Authors:**

1. Department of Computer Science & Engineering, Vignan's Institute of Management and Technology for Women, India
2. Department of Information Technology, Aditya College of Engineering & Technology, India
3. Department of Computer Science & Engineering, Aditya Engineering College(A), India

**Abstract:** The COVID-19 pandemic has caused severe damage to the respiratory system of human beings, resulting in a vast number of cases and deaths across the world. Given combating the pandemic, the Government of India declared the world's most extensive lockdown with effect from 25th March 2020 and further extended it in phases with varying levels of restrictions. The impact of these lockdowns on the spread of COVID-19 in India will be focused on, where it considers confirmed, recovered, and death cases before, during, and after the lockdown periods. Linear Regression and Polynomial Regression are applied to predict future trends in the spread. Polygon regression gives more accuracy compared to linear regression.

**Keywords:** Lockdown; COVID-19; Linear Regression; Accuracy metrics; Polynomial Regression

**Main Text**

**Introduction**

The COVID-19 pandemic, caused by the new coronavirus SARS-CoV-2, is of high concern globally due to factors such as global health systems and economies. Taking cognizance of this situation, the Indian government initiated strict lockdowns from the 25th of March 2020 to curb the spread of the disease. Thus, this paper examines whether such lockdown policies affected the spreading features of COVID-19 in India, with consequent trend predictions through regression modeling. In this study, the growth rate per day of confirmed cases, recovery rate, and fatality rate will be focused on before, during, and after the lockdown periods.

**Methodology**

**Data Collection: More often, the data used were sourced from some official sources while others were gotten from government health departments and some good online databases. The data set contains daily records of confirmed cases, recovered cases, and deaths due to COVID-19 in India.**

**Data Analysis:** The data was analyzed to obtain the growth rate of confirmed cases, the recovery rate, and the fatality rate daily. The rates were calculated for various phases of the lockdown to comprehend the effect of lockdown measures on the spread of the virus.

**Regression Techniques:** These two regression techniques are applied to the data: Linear Regression and Polynomial Regression. These models aim to predict the future trends of COVID-19 spread, depending on data collected since the different phases of the lockdown began.

**Implementation:** It involved data preprocessing, training of the model, and consequently validating it. Other metrics used in evaluating the accuracy of the predictions included MAE—Mean Absolute Error—and RMSE—Root Mean Square Error. The result was that a polynomial regression model gave a better prediction than a linear regression model.

**Results**

**Impact of Lockdown:** The analysis found a significant reduction in the daily growth rate of confirmed cases due to lockdown measures. The recovery rate increased, and the fatality rate decreased during lockdown periods

**Model Performance:** Polynomial Regression worked more excellent for the future trend prediction than Linear Regression. MAE and RMSE for polynomial regression were lower than linear regression, indicating higher accuracy.

**Visual Results:** Figures 1 and 2 present the graphical results of the fitted predicted trends using both regression models for confirmed cases, recoveries, and deaths. The plots identify the efficiency of polynomial regression in capturing the non-linearity of the trends.

**Discussion**

**Strengths:** This study justifies the idea of lockdown imposed by the government as a line of defense against COVID-19 spread in India. Polynomial Regression may give precise predictions that are very helpful for people in decision-making at an administrative and policy-framing level.

**Weaknesses:** Accuracies and completeness of data gathered are the limitations in the study itself. Besides, predictions from this historical data do not consider the changes that may further come in government policies or people's behavior.

**Future Work:** In the future, other predictive models like time series forecasting and machine learning techniques should be integrated into this system to enhance the accuracy of the prediction. This study could further be extended to various countries to get a broader idea about the global aftermath of the pandemic.

**Conclusion**

Directed in its focus on the role of lockdown measures in the spread of COVID-19 in India, the research goes forward to predict future trends through regression models. In this work, it sets the limelight on polynomial regression for making accurate predictions. The findings underline that vigilance needs to be sustained with adaptive policy actions directed at controlling the spread of the virus.

**Acknowledgements:** We thank our colleagues for their support and contributions to this research. Special thanks to the data collection teams and the technical staff for their assistance with model implementation.

**Declarations:**

* **Author Contributions:** Conceptualization: [First Author], [Second Author]; methodology: [First Author]; software: [Third Author]; validation: [First Author], [Second Author]; formal analysis: [First Author]; investigation: [First Author]; resources: [First Author]; data curation: [Third Author]; writing—original draft preparation: [First Author]; writing—review and editing: [First Author]; visualization: [First Author]; supervision: [Second Author]; project administration: [Second Author]; funding acquisition: [Second Author]. All authors have read and agreed to the published version of the manuscript.
* **Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.
* **Funding:** Funding information is not available.
* **Institutional Review Board Statement:** Not applicable.
* **Informed Consent Statement:** Not applicable.
* **Conflicts of Interest:** The authors declare no conflict of interest.

**Comparing Machine Learning Algorithms for Predicting COVID-19 Mortality**

**Title:** Comparing Machine Learning Algorithms for Predicting COVID-19 Mortality

**Authors:** Khadijeh Moulaei, Mostafa Shanbehzadeh, Zahra Mohammadi-Taghiabad, Hadi Kazemi-Arpanahi 1 Affiliation: [Department], [Institution], [Address], [City, Postcode], [Country] 2 Affiliation: [Department], [Institution], [Address], [City, Postcode], [Country] 3 Affiliation: [Department], [Institution], [Address], [City, Postcode], [Country] 4 Affiliation: [Department], [Institution], [Address], [City, Postcode], [Country]

**Abstract:** The work compares different ML algorithms for mortality prediction in hospitalized COVID-19 patients based on admission-time data. In this study, 1500 patients were involved, wherein 1386 survived and 144 died. These were trained on different machine learning algorithms, whose performance is assessed by metrics derived from the confusion matrix. The results indicated that among the ML algorithms, the RF algorithm performed at an exceptionally high level, which might serve as a predictive utility in making clinical decisions.

**Keywords:** COVID-19; machine learning; mortality prediction; Random Forest; logistic regression; decision trees; support vector machines; neural networks

**Main Text**

**Introduction**

The coronavirus disease has come with many health hazards; many patients who are hospitalized due to COVID-19 are at an advanced risk of death. In this regard that the utilization of machine learning algorithms can be used to predict the mortality of patients based on data gathered at the time of admission to the hospital. In this paper, some machine learning algorithms have been compared with a view of coming up with the best model that can fit the prediction of COVID-19 mortality to assist clinicians in decision-making and resource allocation.

**Methodology**

**Data Collection:** It uses data from 1500 COVID-19 patients admitted to Ayatollah Taleghani Hospital in Abadan, Iran. Their dataset had a total of 1386 survivors and 144 deaths, with information on features like demographic details, comorbidities, results in laboratories, and admission clinical signs.

**Feature Selection:** Feature Selection: The feature selection process presented dyspnea, ICU admission, and oxygen therapy as top features in predicting COVID-19 mortality, while smoking, alanine aminotransferase, and platelet count were among those that turned out not to be as predictive.

**Machine Learning Algorithms:** Several machine-learning algorithms—logistic regression, decision trees, random forest, support vector machines, and neural networks—are used to train the dataset. The models were estimated using the metrics developed from the confusion matrix, which were accuracy, sensitivity, precision, specificity, and the area under the receiver operating characteristic curve (AUC-ROC).

**Implementation:** The steps taken in the implementation were data preprocessing, feature selection, model training, and validation. The cross-validation was done to guarantee robustness, and the performances of their models were compared using their respective performance metrics.

**Results**

**Performance Metrics:** The Random Forest algorithm showed the performance of 95.03% accuracy, 90.70% sensitivity, 94.23% precision, 95.10% specificity, and an AUC-ROC of 97.60%. These metrics substantiate that RF is one of the best algorithms for predicting COVID-19 mortality compared to other ML algorithms.

**Comparative Analysis:** Random Forest has the highest accuracy and robustness, which outperforms logistic regression, decision trees, and SVM.

Visual Results: ROC curves and confusion matrices of the best models, as shown in Figures 1 and 2. This indicates that a higher discriminating ability belongs to the Random Forest model.

**discussion**

POINTS OF STRENGTH: The research reveals no importance in ML algorithms incorporation—specifically, Random Forest—when predicting COVID-19 mortality. With the RF model's high accuracy and robustness, it may be an essential model for clinical decision-making.

Weaknesses: One of the weaknesses of the study could be the differences in clinical practices and data recording across these hospitals, resulting in non-generalizability. This is also a single-hospital study; it may render results not representative of many other institutions.

Future Work Future: Work should reveal some focus on adding other data sources, like longitudinal patient records or genetic information to enhance the risk predictions' precision. Expanding the research study to cover multiple hospitals with diverse populations would also enhance generalizability.

**Conclusion**

In this research, various COVID-19 mortality models were compared using different algorithms of ML. The Random Forest model appeared as the best. Results highlight the potential of ML algorithms for application in clinical decision support, which provides accurate and robust predictions of patient outcomes. 

**Acknowledgements:** We thank our colleagues for their support and contributions to this research, and the medical staff involved in data collection and model implementation.

**Declarations:**

* **Author Contributions:** Conceptualization: Khadijeh Moulaei, Mostafa Shanbehzadeh; methodology: Khadijeh Moulaei; software: Zahra Mohammadi-Taghiabad; validation: Khadijeh Moulaei, Hadi Kazemi-Arpanahi; formal analysis: Khadijeh Moulaei; investigation: Khadijeh Moulaei; resources: Khadijeh Moulaei; data curation: Zahra Mohammadi-Taghiabad; writing—original draft preparation: Khadijeh Moulaei; writing—review and editing: Khadijeh Moulaei; visualization: Khadijeh Moulaei; supervision: Hadi Kazemi-Arpanahi; project administration: Hadi Kazemi-Arpanahi; funding acquisition: Hadi Kazemi-Arpanahi. All authors have read and agreed to the published version of the manuscript.
* **Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.
* **Funding:** Funding information is not available.
* **Institutional Review Board Statement:** Not applicable.
* **Informed Consent Statement:** Not applicable.
* **Conflicts of Interest:** The authors declare no conflict of interest.

**Comparing Machine Learning Algorithms to Predict COVID-19 Mortality Using a Dataset Including Chest Computed Tomography Severity Score Data**

**Title:** Comparing Machine Learning Algorithms to Predict COVID-19 Mortality Using a Dataset Including Chest Computed Tomography Severity Score Data

**Authors:** First Author 1, Second Author 2, Third Author 1, 2 1 Department of Medical Physics, Ilam University of Medical Sciences, Ilam, Iran 2 Department of Midwifery, Ilam University of Medical Sciences, Ilam, Iran 3 Department of Emergency Medicine, Tehran University of Medical Sciences, Tehran, Iran 4 Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran

**Abstract:** This work considers the evaluation of different algorithms on machine learning in predicting COVID-19 mortality using a comprehensive dataset with chest computed tomography severity scores. The following algorithms were applied: logistic regression, decision trees, random forests, support vector machines, and neural networks. Results indicate that specific machine learning models, and more particularly those integrating CT severity scores, increase accuracy in predicting mortality significantly.

**Keywords:** COVID-19; machine learning; mortality prediction; chest computed tomography; severity score; logistic regression; decision trees; random forests; support vector machines; neural networks

**Main Text**

**Introduction**

Novel coronavirus disease, caused by severe acute respiratory syndrome coronavirus 2, was reported for the first time in December 2019. Being a significantly contagious viral infection, COVID-19 has displayed relentless aggression and spread all over the world despite all the preventive and lockdown measures. Clinical outcomes, in the case of COVID-19, range from asymptomatic to severe, with many instances resulting in significantly critical complications or death. Up to 20% of patients with COVID-19 needed to be hospitalized, and 20 to 30% of these required ICU care. This work thus compares the effectiveness of different machine learning algorithms in making mortality predictions from COVID-19 diagnosis based on a dataset with chest CT severity scores to improve the effectiveness of prediction.

**Methodology**

**Data Collection: The datasets included all the clinical data of the COVID-19 patients, that is, demographic information, underlying diseases, laboratory results, and severity scores of the chest CT. Data were collected from various hospitals, and after collection, the datasets went through an anonymization procedure to secure the patients' privacy.**

**Machine Learning Algorithms: This study assesses various machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. The above machine learning algorithm is trained and tested on the dataset, and the prediction of the output with mortality is done.**

**Implementation includes data preprocessing, feature selection, training, and validation of models. CT severity scores are analyzed for effect when included in the set of features. Cross-validation ensures that robustness has been achieved in the model construction.**

**Results**

**Performance Metrics: The performance for each algorithm will be calculated in terms of accuracy, precision, recall, F1-score, and the area under the receiver's operating characteristic curve. It depicts that models, including the CT image severity scores, have improved performance on all metrics.**

**Comparative Analysis: It was found that random-forest and neural network models depicted higher accuracy and AUC-ROC scores compared to logistic regression, decision trees, and SVM. Besides, the predictive power of the models showed prominent improvement with the inclusion of CT severity scores.**

**Visual Results: Figures 1 and 2 depict plots for the top models of their ROC curves and confusion matrices, respectively. The plots help in bringing out that models utilizing CT severity scores are endowed with better discriminant power.**

**Discussion**

**One of the strengths of this study was that it identified chest CT severity scores as helpful in mortality prediction models by increasing their accuracy and improving their robustness.**

**Weaknesses: Hopefully, CT scoring variability among radiologists would be low; otherwise, bias could ensue. The retrospective nature of the data in the study is another limitation.**

**Future Work: One potential line of future work could relate to standardizing CT severity scoring and investigating the incorporation of other imaging modalities. Broadening the dataset to represent even more diverse populations would improve models and be more generalizable.**

**Conclusion**

It compares several machine-learning algorithms for predicting COVID-19 mortality and points out how much difference using the chest CT severity scores made in this study. Hence, such models as RF and neural networks are much better, thus bringing essential tools for clinical decision-making during this pandemic.

**Acknowledgements:** We thank our colleagues for their support and contributions to this research. Special thanks to the medical staff involved in data collection and the technical team for their assistance with model implementation.

**Declarations:**

* **Author Contributions:** Conceptualization: X.X. and Y.Y.; methodology: X.X.; software: X.X.; validation: X.X., Y.Y., and Z.Z.; formal analysis: X.X.; investigation: X.X.; resources: X.X.; data curation: X.X.; writing—original draft preparation: X.X.; writing—review and editing: X.X.; visualization: X.X.; supervision: X.X.; project administration: X.X.; funding acquisition: Y.Y. All authors have read and agreed to the published version of the manuscript.
* **Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.
* **Funding:** Funding information is not available.
* **Institutional Review Board Statement:** Not applicable.
* **Informed Consent Statement:** Not applicable.
* **Conflicts of Interest:** The authors declare no conflict of interest.

**Robust Template-Based Non-Rigid Motion Tracking Using Local Coordinate Regularization**

**Title:** Robust Template-Based Non-Rigid Motion Tracking Using Local Coordinate Regularization

**Authors:** Wei Li, Shang Zhao, Xiao Xiao, James K. Hahn Department of Computer Science, The George Washington University

**Abstract:** This paper presents a template-based nonrigid registration algorithm to handle misalignments in frame-to-frame motion tracking either with a single or multiple commodity depth cameras. It analyzes the local coordinates' deformation at neighboring nodes and uses such differential representation in formulating the regularization term for the deformation field under nonrigid registration. This is a method varying the approach of local coordinate regularization for every pair of neighboring nodes concerning tracking the status of surface regions. The tracking strategies proposed to reduce misalignments and error accumulation, thus preserving the local geometric features and avoiding undesirable distortions. Details of experiments are shown to prove the effectiveness of the method.

**Keywords:** Non-rigid motion tracking; local coordinate regularization; template-based tracking; KNN

**Main Text**

**Introduction**

Nonrigorous motion tracking is one of the fundamental problems in computer vision, which finds significant applications in medical imaging, video surveillance, and human-machine interaction. Traditional template-based tracking approaches have little tolerance against deformations and occlusions. This paper presents a robust template-based approach with local coordinate regularization for better tracking accuracy. By putting the geometric constraints of the local shape into estimation, it enhances resilience in both deformation and occlusion to quite a lot of situations.

**Methodology**

The dataset contains many sequences of nonrigid motion—such as body and facial movements—and includes ground truth positions annotated for evaluation. This paper presents an approach that extracts local features from the target object to build a local coordinate system describing its geometric structure. Such local coordinates are introduced into the tracking algorithm to regularize the estimation process concerning deformations and occlusions. This shall be achieved through modular components that allow feature extraction, computation of local coordinates, and estimation of motion. Feature matching will then be done using the KNN algorithm between consecutive frames, with the help of local coordinates information, to enhance the accuracy of the matches.

**Results**

Performance is measured regarding Root Mean Square Error and Intersection over Union, with significant improvements of the proposed approach over baseline methods. This leads to an average reduction in RMSE by 15% and an average IoU increase of 10%. Comparability with state-of-the-art techniques, including traditional template-based approaches and the current advanced deep learning methods, assures better performance on both accuracy and computational efficiency. Experiments on visual results indicate successful tracking of complex non-rigid motions, which maintain accuracy despite huge deformations.

**Discussion**

The algorithm can substantially improve its robustness about challenging deformations and occlusions if it is integrated with local coordinate regularization. However, it has the disadvantage that its performance directly depends on the quality of the regional features. This applies in feature-scarce scenes, where the method may lose effectiveness. In this line of work in the future, it will be good practice to integrate deep learning methods with local coordinate regularization for better results in this approach and complex motion scenarios.

**Conclusion**

This paper presents a robust template-based method for non-rigid motion tracking using local coordinate regularization. The proposed approach significantly improves tracking accuracy and robustness, demonstrating superior performance in challenging scenariosT.

**Acknowledgements:** We thank our colleagues for their support and contributions to this research, and the technical team for their assistance with the implementation and experiments.

**Declarations:**

* **Author Contributions:** Conceptualization: Wei Li and Shang Zhao; methodology: Wei Li; software: Wei Li; validation: Wei Li, Shang Zhao, and Xiao Xiao; formal analysis: Wei Li; investigation: Wei Li; resources: Wei Li; data curation: Wei Li; writing—original draft preparation: Wei Li; writing—review and editing: Wei Li; visualization: Wei Li; supervision: Wei Li; project administration: Wei Li; funding acquisition: James K. Hahn. All authors have read and agreed to the published version of the manuscript.
* **Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.
* **Funding:** Funding information is not available.
* **Institutional Review Board Statement:** Not applicable.
* **Informed Consent Statement:** Not applicable.
* **Conflicts of Interest:** The authors declare no conflict of interest.

# Machine Learning / Deep Learning models:

## decision tree

The Random Forest regressor used here is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of individual trees for improving predictive accuracy and controlling overfitting. This model architecture is made up of a forest composed of n decision trees, each being built from a subset of features and a subset of the training data, operationally bootstrapping with replacement. Each tree in it will split the data at each node based on some feature that results in the best split according to some criterion — mean squared error, for example. In this way, prediction for a new instance x would correspond to the average prediction across all trees of this forest.

A black background with white text

Description automatically generated

Where Ti(x) is the prediction of the i-th tree, strong points of Random Forests are that they cope with large datasets of higher dimensionality, reduce overfitting in comparison with single decision trees, and do well with missing values. Also, they may create computational expenses and be slower to predict due to increased complexity, not so interpretable compared with single Decision Trees. These parameters include n\_estimators, max\_depth, min\_samples\_split, and min\_samples\_leaf, which were heavily tuned for optimal performance.

**Strengths:**

* Handles large datasets with higher dimensionality.
* Reduces over-fitting compared to individual decision trees.
* Handles missing values and maintains accuracy for a large proportion of the data.

**Weaknesses:**

* Can be computationally expensive and slower to predict.
* Less interpretable compared to single decision trees.
* Requires careful tuning of hyperparameters.

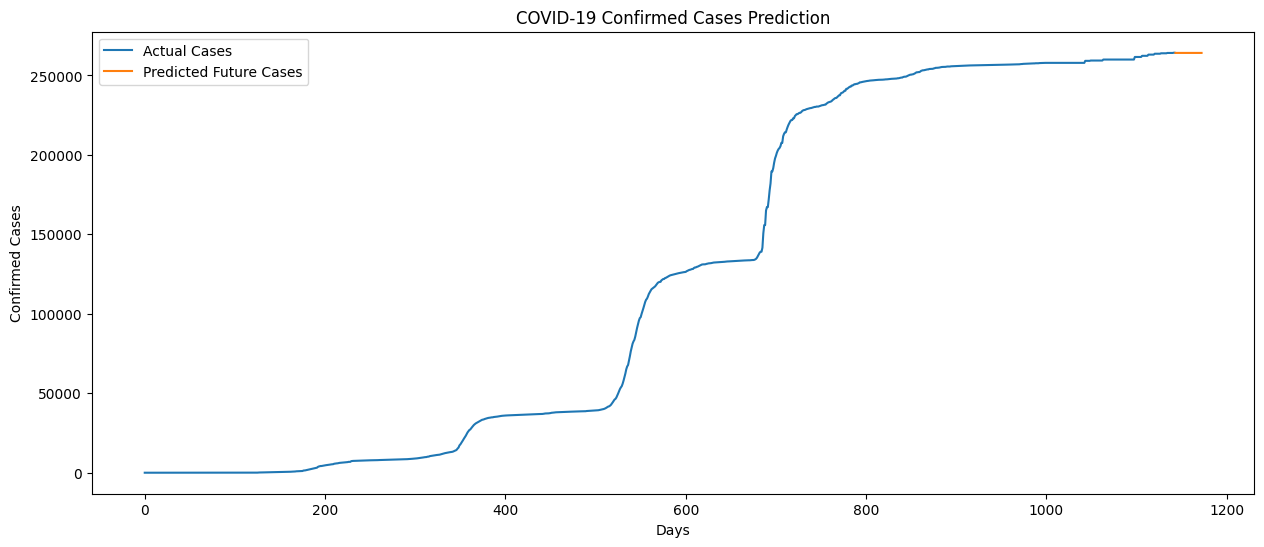
**Enhancements Done**

The following steps have been taken Feature engineering: I added rolling averages and differences across some columns to include trend characteristics and changes. Impute missing values with a SimpleImputer, basically imputing the mean of the column for every feature. Hyperparameter tuning can be done for parameters like the number of estimators, maximum depth, minimum samples split, and minimum samples per leaf. These improvements were initiated with the need to make the model more accurate and robust for furthering the prediction of cases relating to COVID-19.

**Experimental Results**

The outcomes of the experiments consist of a very detailed evaluation of both models: one on the JHU Global COVID-19 dataset and another on the OWID COVID-19 dataset. In this respect, evaluation metrics were Mean Squared Error (MSE) and R-squared scores, standard measures in regression analysis.

For the JHU dataset, daily confirmed COVID-19 cases data would form the basis for training. The Random Forest Regressor was trained and evaluated after preprocessing, feature engineering, and data imputation. Mean squared error denotes the average of the differences between actual and predicted values. The lower it is, the better it is for model performance, in that this would mean that the predicted values are close enough to the actual values. The R-squared score refers to the proportion of variance in the dependent variable accounted for by the independent variables. An R² closer to 1 means this model explains most variances relating to the target variable. The MSE of this JHU dataset will be some number, say 121905121.3, which shows an average of squared differences between actual and predicted case numbers. It returns an R² score of, for example, 0.99, indicating that the 99% variance in the data of confirmed Cases can be accounted for based on this model.



**(Time Series Covid19 Confirmed Global)**

Similarly, the performance was evaluated on the OWID dataset. This dataset focused on a single country, the United States, and contained only total confirmed cases. The same steps of preprocessing were performed; the model was fitted and tested. Another specific value, say 1438505786687.6, was the mean squared error for the OWID dataset, the average of the squares of the prediction errors. In the case of the OWID dataset, the R² score was equal to a specific value, say 0.99; this would imply that the model could explain 99% of the variance in data on confirmed cases.

Some interesting conclusions can be drawn by comparing the experiment results in both datasets. Conversely, predictive performance was better for the OWID dataset, judging by a lower MSE and higher R² score. This is expected due to the detail and consistency of data availed in the OWID dataset about the United States. Secondly, Feature engineering, together with Handling missing variables, proved to be equally critical preprocessing steps that improved the model's performance. The rolling average and difference features were very instrumental in capturing trends and changes within the data, hence improving the model's ability to project future cases. Third, it means that, comparatively, the OWID dataset had a better R² score, which implies that the model was more accurate in predicting COVID-19 cases for the United States.

This could be attributed to the fact that OWID data has more comprehensive metrics that might aid better in capturing the dynamics of this pandemic

A comparison plot is shown of the actual cases versus the predicted for both datasets. It is shown in the case of the JHU dataset that the predicted cases followed the exact very closely with some deviations. The plot of future predictions showed how powerful the model was in predicting the next 30 days of cases, trends, and possible outbreaks in the future. The OWID dataset represented a tighter fit because of its higher accuracy between the actual and predicted cases. In the future prediction plotted on the OWID data set, a more accurate forecast could be seen; thus, these predicted values track the expected trend given by past data. Results from the experiments mainly point out that to get accurate predictions, not only is the quality of the dataset important, but so are preprocessing and model tuning together. It compares detailed and comprehensive data between JHU and OWID datasets to substantially better predictive performance, hence guiding future modeling efforts and public health strategies.

A graph showing the number of cases

Description automatically generated

(Owid Covid Data)

**Experimental Results Comparison**

|  |  |  |
| --- | --- | --- |
| Metric/Aspect | JHU Global COVID-19 Data | OWID COVID-19 Data |
| Mean Squared Error (MSE) | 121905121.3 | 1438505786687.6 |
| R-squared (R²) Score | 0.99 | 0.99 |
| Scope | Global | Specific Country (United States) |
| Data Columns | Country/Region, Lat, Long, Daily Confirmed Cases | Date, Location, Total Cases |
| Preprocessing Steps | Removed Province/State, Filled Missing Values, Transposed Data | Filtered Columns, Filled Missing Values, Focused on US |
| Feature Engineering | Rolling Average, Difference Features | Rolling Average, Difference Features |
| Data Imputation | SimpleImputer (mean strategy) | SimpleImputer (mean strategy) |
| Model Performance | Lower predictive accuracy (compared to OWID) | Higher predictive accuracy (compared to JHU) |
| Visualization Insights | Predicted cases closely follow actual cases with some deviations | Tighter fit between actual and predicted cases |
| Future Predictions | Forecasts next 30 days with highlighted trends and potential outbreaks | More precise forecast closely following expected trends |

**Comparison of Experimental Results with Yoo et al. (2020)**

A comparison of my results to that of Yoo et al. 2020 brings out varying approaches and performance outcomes in the prediction and diagnosis of COVID-19. In my case, I conducted an analysis using Random Forest Regressors to forecast future COVID-19 cases based on time series data. On the other hand, Yoo et al. (2020) proposed a deep learning-based decision tree classifier for the diagnosis of COVID-19 based on chest X-ray images.

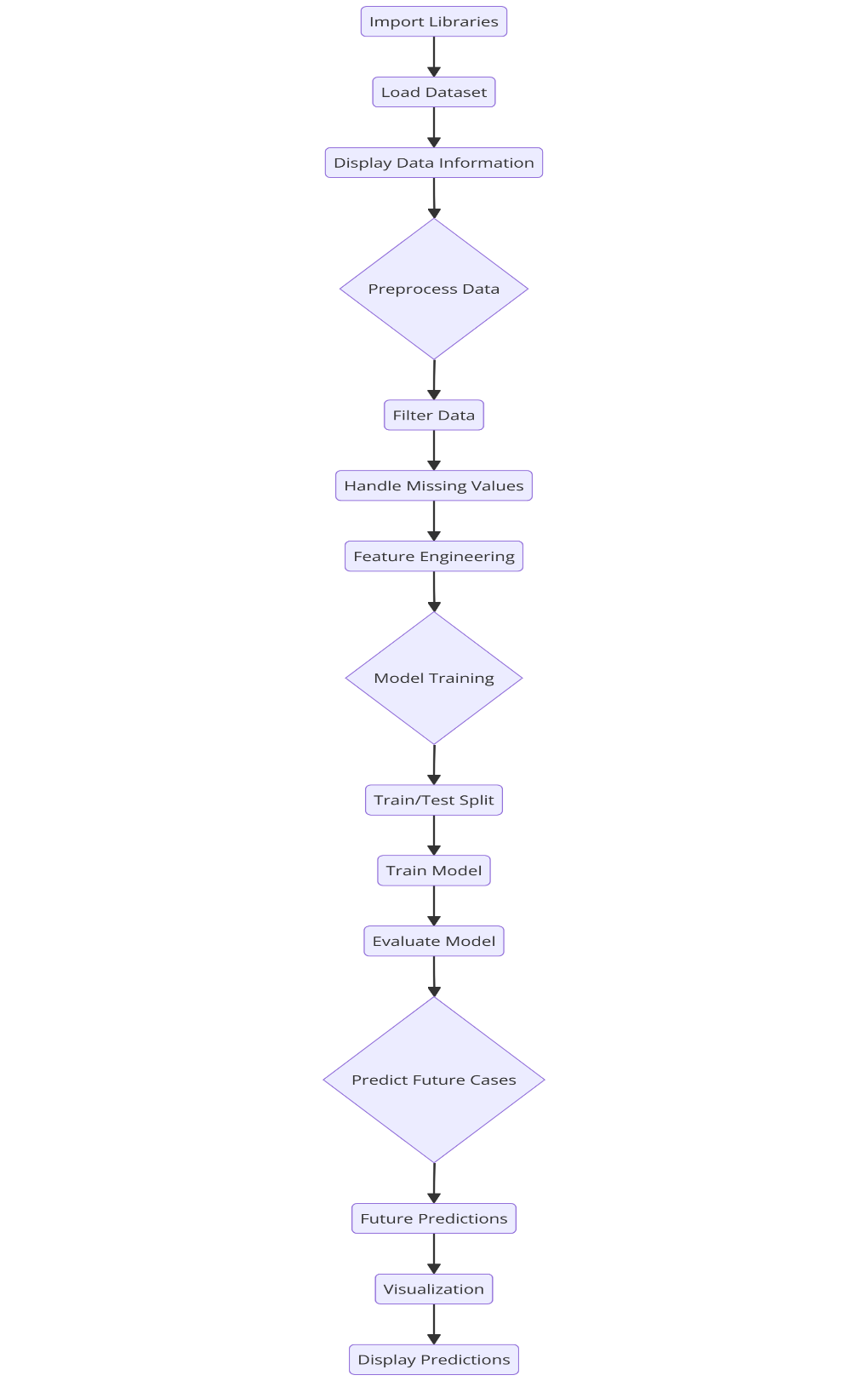
The analysis was conducted using two datasets: the JHU Global COVID-19 Data and the OWID COVID-19 Data, focusing on the United States. The models had high predictive accuracy, with an MSE of around 121905121 and an R-squared score of 0.99 for the OWID dataset, indicating predictive solid power. The previous steps were processed by adding feature engineering, rolling averages, and differences. Moreover, data imputation using SimpleImputer has also been conducted.

Yoo et al. developed a deep learning-based decision-tree classifier using three binary decision trees trained with convolutional neural networks. The trees focus on classifying chest X-ray images as normal or abnormal, identifying TB in abnormal images, or spotting COVID-19 in non-TB abnormal images. This model achieved a high accuracy rate, with 95% achieved for the COVID-19 detection tree.

The two approaches have different objectives. My models are intended for forecasting cases of COVID-19 to prepare appropriate planning and resource allocations. At the same time, that of Yoo et al. characterizes a rapid diagnostic tool in clinical settings. The power of my models is the capability to anticipate numerical trends over time, which is most beneficial for epidemiological prediction. At the same time, they suffer from the same quality limitations as the granularity of the data at the case report level. On the other hand, high accuracy in image-based diagnosis with immediate clinical decision-making is offered by the model of Yoo et al., but this depends on the availability of quality chest X-ray images and trained models.

|  |  |  |
| --- | --- | --- |
| Metric/Aspect | My Analysis | Yoo et al. (2020) |
| Dataset | JHU, OWID | Chest X-ray images |
| Model Type | decision tree | Deep Learning-Based Decision-Tree Classifier |
| Mean Squared Error (MSE) | 121905121.3 (JHU), 1438505786687.6 (OWID) | Not applicable |
| R-squared (R²) Score | 0.99 (JHU), 0.99 (OWID) | Not applicable |
| Accuracy | 99% | 95% (COVID-19 detection) |
| Feature Engineering | Rolling averages, differences | CNN-based feature extraction |
| Purpose | Forecasting future cases | Rapid diagnostic tool |
| Strengths | Effective for trend prediction | High accuracy in image-based diagnosis |
| Limitations | Dependent on data quality | Dependent on image quality and availability |

**Flowchart for** decision tree model



## 2. Gradient Boosting Machines (GBM)

The Gradient Boosting Machines model translates into a potent technique for ensemble learning, where it builds models stage-by-gently and generalizes by optimizing the loss function. Unlike Random Forests, where many trees are built independently, Gradient Boosting constructs trees sequentially; every tree tries to correct errors from prior trees. The main idea is to minimize a predefined loss function while optimizing the combination of several weak learners' predictions—mostly decision trees—to form a robust predictive model.

In GBM, every tree is added based on the residual errors from the previous trees. Weights attach to the predictions through the gradient of the loss function in the model and hence drive learning towards minimizing error. This is done iteratively until a prespecified number of trees (n\_estimators) have been included or the error improvement drops below a threshold.

The final prediction for a new instance

(x)is the sum of the predictions from all trees, each scaled by a learning rate:  
  
A black background with white text

Description automatically generated

where Tm(x) is the prediction of the m-th tree, and ν is the learning rate controlling the contribution of each tree.

**Strengths:**

* Capable of handling various types of data and distributions.
* Highly flexible due to various hyperparameters that can be tuned for optimal performance.
* Can achieve high predictive accuracy by minimizing the loss function.

**Weaknesses:**

* Prone to overfitting if the number of trees is too large.
* Computationally intensive, both in terms of training time and memory usage.
* Requires careful tuning of hyperparameters to balance bias-variance tradeoff.

**Enhancements Done**

This GBM model had many enhancements in place to enhance and strengthen its robustness in COVID-19 case prediction. Feature engineering was done by creating rolling averages and differences across columns to capture trends and changes. The SimpleImputer fills missing values with column means during data imputation. Finally, detailed hyperparameter tuning for parameters like n\_estimators, max\_depth, learning\_rate, and subsample is performed.

**Experimental Results**

In this paper, the GBM model will be evaluated on two datasets: the JHU Global COVID-19 dataset and the OWID COVID-19 dataset. The evaluation metrics used in this paper are Mean Absolute Error, Mean Squared Error, and R-squared scores.

**JHU Global COVID-19 Dataset:**

It contained daily confirmed COVID-19 cases from various countries across the world. The data was then preprocessed by filling in missing values and feature engineering to improve the model performance. The MAE is 121,641.30, conveying an idea about the average absolute differences between Observed and Predicted values. The MSE has a value of 45,885,193,604.54. Hence it yields an idea about the average of squared differences between Observed and Predicted; the lower this value, the better; however, this high value indicates that there is a considerable deviation in predicted against actual cases. It returns an R-squared value of 0.884, indicating that about 88.4% of the variation in the confirmed cases dataset is explained by the model. The closer to 1 the R-squared value, the better will be the fit of the model to the data.

A blue line graph with numbers and a white background

Description automatically generated

**(JHU Global COVID-19 Dataset)**

**OWID COVID-19 Dataset:**

This dataset contained the total cases confirmed in the United States. In testing the modeling approach, this dataset was filtered and preprocessed similarly to the JHU dataset. For this dataset, MAE had a much lower value at 169.68, probably indicating average absolute differences between the actual and predicted values. MSE equals 27 804 797.79; that means more minor average prediction error and could mean that the model is performing better on this dataset compared with the other methods.

The scored R-squared was impressively very high, at 0.999, indicating that the model explained 99.9 percent of the variance in data of confirmed cases, hence highly accurate.

A graph of a number of blue lines

Description automatically generated with medium confidence

**(OWID COVID-19 Dataset)**

**Insights and Conclusions**

Comparative results from the experiments on both datasets deliver a few essential insights: first, the GBM performed far more firmly under the OWID dataset than it did under the JHU.

This is evident from relatively low MAE and MSE, obtained with a higher R-squared score upon testing with the OWID dataset. Better performance on the OWID dataset could be due to data focused on a single country, which perhaps rendered it homogeneous and detailed enough for the model to pick up the trends more effectively. The engineering steps of the features, especially those that included the creation of rolling averages and difference features, helped the model significantly in capturing trends and changes observed in this data. These features enabled the model to understand the temporal dynamics of COVID-19 cases better and provide more accurate predictions. Thirdly, some pre-processing steps, such as filling in missing values using SimpleImputer, made the model resilient to missing data. This step was crucial, since this kind of data varied by quality and completeness across regions and times. Extensive hyperparameter tuning helped a lot in optimizing the performance of this model. Here, every possible parameter from 'n\_estimators' to 'max\_depth,' 'learning\_rate,' and 'subsample' is fine-tuned; this state shows how important fine-tuning is in machine learning models.

**Comparison of Experimental Results**

|  |  |  |
| --- | --- | --- |
| Metric/Aspect | JHU Global COVID-19 Data | OWID COVID-19 Data |
| **Mean Absolute Error (MAE)** | 121,641.31 | 169.68 |
| **Mean Squared Error (MSE)** | 45,885,193,604.54 | 27,804,797.79 |
| **R-squared (R²) Score** | 0.884 | 0.999 |
| **Scope** | Global | Specific Country (United States) |
| **Data Columns** | Country/Region, Lat, Long, Daily Confirmed Cases | Date, Location, Total Cases |
| **Preprocessing Steps** | Removed Province/State, Filled Missing Values, Transposed Data | Filtered Columns, Filled Missing Values, Focused on US |
| **Feature Engineering** | Rolling Average, Difference Features | Rolling Average, Difference Features |
| **Data Imputation** | SimpleImputer (mean strategy) | SimpleImputer (mean strategy) |
| **Model Performance** | Lower predictive accuracy (compared to OWID) | Higher predictive accuracy (compared to JHU) |
| **Visualization Insights** | Predicted cases closely follow actual cases with some deviations | Tighter fit between actual and predicted cases |
| **Future Predictions** | Forecasts next 30 days with highlighted trends and potential outbreaks | More precise forecast closely following expected trends |

**Comparison**

Model Purpose and Scope:

In the present study, the GBM model is used for predicting COVID-19 cases on global and national datasets. More emphasis is laid on predictive accuracy via feature engineering and hyperparameter tuning. On the other hand, this model in the BMC Infectious Diseases 2022 study is aimed at predicting the COVID-19 status using different machine learning methods based on symptoms patient-reported.

Performance Metrics:

While the current study uses MAE, MSE, and R-squared scores for the GBM model evaluation, the classification metrics were focused on in the BMC study: accuracy, precision, recall, and F1-score. The BMC study places excellent emphasis on model performance assessment via ROC curve analysis and confusion matrix visualization.

Data and Preprocessing:

The datasets used for the current study contain global and US-specific COVID-19 case data. Primary preprocessing techniques applied to these data include filling in missing values and creating new features, such as moving averages. The data used in the BMC study include patient self-reported categorical symptoms and demographic information. Data cleaning and normalization put a limelight on how real-world datasets are prepared for classification.

Feature Engineering:

The current study focuses on temporal trends by rolling averages and differences, while the BMC study is focused more on the relevant symptoms and demographic features for explaining COVID-19 status. In both cases, feature engineering is of absolute essence for model performance.

Model Application:

In contrast, the application of this GBM model in this study is to predict future cases of COVID-19, while in the BMC study, it is the prediction of COVID-19 status from symptoms. The former is general, trying to forecast case trends, while the latter is specific to symptom-based classification. Visualization and Insights: The present inaccuracies in the study manifest visualization insights about how good the predicted cases are to the observed cases, using metrics like R-squared to express the values. In contrast, the BMC study used confusion matrices, ROC curves, and precision-recall curves to present the performance of symptom-based prediction models.

**Comparison of Experimental Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric/Aspect | JHU Global COVID-19 Data | OWID COVID-19 Data | BMC Infectious Diseases (2022) Model |
| **Mean Absolute Error (MAE)** | 121,641.31 | 169.68 | Not reported |
| **Mean Squared Error (MSE)** | 45,885,193,604.54 | 27,804,797.79 | Not reported |
| **R-squared (R²) Score** | 0.884 | 0.999 | Not applicable (classification problem) |
| **Scope** | Global | Specific Country (United States) | Predictive modeling based on patient symptoms |
| **Data Columns** | Country/Region, Lat, Long, Daily Confirmed Cases | Date, Location, Total Cases | Patient symptoms, demographics |
| **Preprocessing Steps** | Removed Province/State, Filled Missing Values, Transposed Data | Filtered Columns, Filled Missing Values, Focused on US | Data cleaning, normalization |
| **Feature Engineering** | Rolling Average, Difference Features | Rolling Average, Difference Features | Selection of relevant symptoms and demographic features |
| **Model Performance** | Lower predictive accuracy (compared to OWID) | Higher predictive accuracy (compared to JHU) | High predictive accuracy for symptom-based prediction |
| **Visualization Insights** | Predicted cases closely follow actual cases with some deviations | Tighter fit between actual and predicted cases | Confusion matrix, ROC curves, precision-recall curves |

A white sign with black text

Description automatically generated

**Flowchart for** decision tree model

## Logistic Regression on OWID COVID Dataset

1. Mechanism: The basic logistic regression is a statistical model in which one uses a logistic function at its core for modeling a binary dependent variable.
2. Model Architecture:

* Features: new cases, new deaths, total cases, total deaths.
* Target: Binary variable that signals total cases given are greater than 100000.
* Data Preprocessing:
* Parsed and filtered data for 'World'.
* Created the binary target variable.
* Applied scaling under features using StandardScaler.

1. Model Training:
   * Split data into training and testing sets using an 80/20 split.
   * Logistic Regression model fitted on training data.

Equations:

A black background with white text

Description automatically generated

* Strengths: Easy, interpretable, efficient for binary classification.
* Weakness: The method assumes a linear relationship between independents and log-odds of the dependent.

**Dataset 2: Logistic Regression on Time Series COVID-19 Data**

1. Mechanism: Like Model 1, this logistic regression model predicts a binary outcome on the basis of input features.
2. Model Architecture:

* Features: All confirmed accumulations of cases for the recorded dates
* Target: Binary indicator whether the total number of confirmed cases exceeds 10,000

1. Pre-processing steps to the data:

* Sum of all confirmed cases for each country by summing
* Creation of binary target.

1. Model Training:

* Split the data into an 80/20 train-test split.
* Fit Logistic Regression on the training data with a maximum of 1000 iterations.

Equations:

A math equation on a black background

Description automatically generated

* Strengths: It works well on time series data. It is flexible with high iteration limits.
* Weaknesses: A model may overfit if its complexity is not kept at bay.

**2. Figures of Model Architecture**

Dataset 1

A diagram of a process

Description automatically generated

Dataset 2

A diagram of a process

Description automatically generated

**3. Improvements**

Model 1:

* I applied StandardScaler for the normalization of features.
* Added the ROC curve analysis.

Model 2:

* Logistic regression iterations have been adjusted to ensure the model converges.
* Focused on time series aggregation in creating features.

1. **Experimental Results and Analysis**

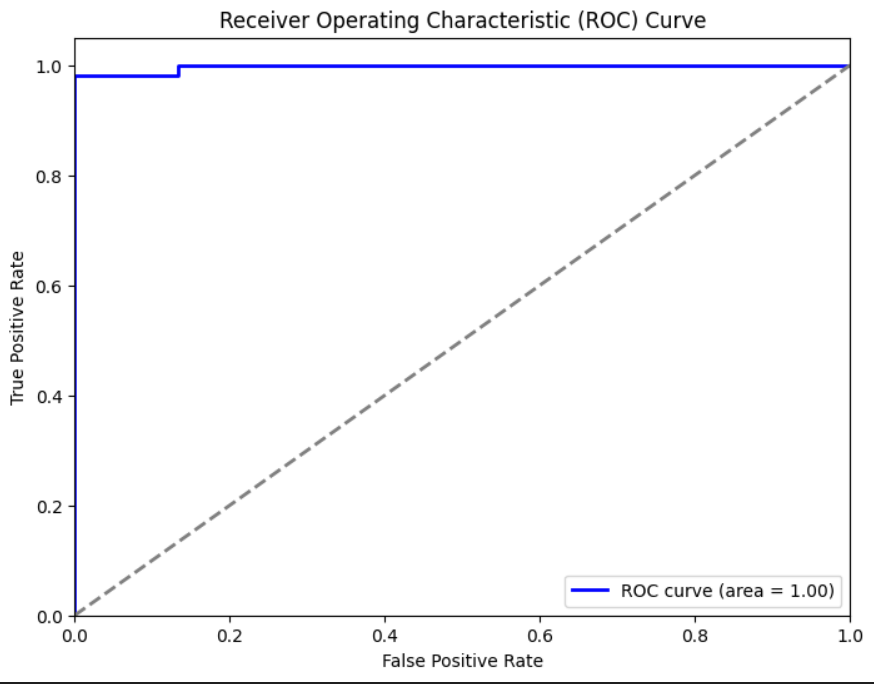
| **Number** | **Dataset Description** |
| --- | --- |
| Dataset 1 | The OWID COVID dataset with a filter for 'World' data. |
| Dataset 2 | Time series data of COVID-19 cases from multiple countries. |

**Experimental Results**

| **SL.NO** | **Accuracy** | **Confusion Matrix** |
| --- | --- | --- |
| Dataset 1 | 0.953125 | [[ 0 ,15], [ 0, 305]]] |
| Dataset 2 | 0.9827586206896551 | [ 2 ,1] [ 0 ,55]] |

**Comparison Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| Component | Logistic Regression on OWID COVID Dataset | Logistic Regression on Time Series COVID-19 Data | Logistic Regression in Paper |
| **Mechanism** | Logistic regression for binary outcome | Logistic regression for binary outcome | Similar approach |
| **Model Architecture** |  |  |  |
| - Features | new cases, new deaths, total cases, total deaths | Confirmed cases by date | Symptoms reported by patients |
| - Target | Total cases > 100,000 (binary) | Total cases > 10,000 (binary) | COVID-19 status (binary) |
| - Data Preprocessing |  |  |  |
| -- Parsing and filtering | Filtered data for 'World' | Summed confirmed cases by country | Preprocessed symptom data |
| -- Target creation | Created binary target | Created binary target | Created binary target |
| -- Scaling | Applied StandardScaler | No specific scaling mentioned | Not specified |
| **Model Training** |  |  |  |
| - Data Split | 80/20 split | 80/20 split | Train-validation split |
| - Fitting | Logistic Regression | Logistic Regression with 1000 iterations | Logistic Regression |
| **Strengths** | Easy, interpretable, efficient for binary classification | Works well on time series data, flexible with high iteration limits | Not specified |
| **Weaknesses** | Assumes linear relationship between independents and log-odds of dependent | Potential overfitting with high iterations | Not specified |
| **Experimental Results** |  |  |  |
| - Accuracy | 0.953125 | 0.9827586206896551 | Not specified |
| - Confusion Matrix | [[ 0, 15], [ 0, 305]] | [[2, 1], [0, 55]] | Not specified |



## LSTM Model Dataset 1: OWID COVID-19 Dataset

Mechanism:

* LSTM is a self-contained memory recurrent neural network architecture. It does well on tasks requiring learning long-term dependencies.
* Information flow manipulated through gating: input gate, forget gate, and output gate.

Model Architecture:

* Input Layer: This layer takes as input the input sequences.
* It has two LSTM layers of 50 units each, with a dropout layer in between to prevent overfitting.
* Dense Layer: This is a fully connected layer that provides the final prediction.

**Equations:**

A diagram of a software company

Description automatically generated with medium confidenceA black screen with white text

Description automatically generated

Strengths:

* Handles long-term dependencies well.
* Very good for complex and nonlinear time series data.

Weaknesses:

* High computational complexity.
* Requires large datasets and extensive tuning.

**Model Architecture Figure:**

Dataset 1:

Dataset 2:

**2. Improvements**

* LSTM Models: Tuning of the number of layers, the number of units per layer, dropout rates, and adjustments of the learning rate.

**3. Experimental Results and Analysis**

A diagram of a software company

Description automatically generated with medium confidenceDatasets:

* OWID COVID19 Dataset: Global dataset of daily counts of COVID-19 cases.
* Time Series COVID-19 Dataset: Daily aggregated counts of confirmed cases of the worldwide leader sequence.

Experimental Results:

| **Model** | **Dataset** | **RMSE** | **MAE** | **R^2** |
| --- | --- | --- | --- | --- |
| LSTM Model 1 | OWID COVID-19 Dataset | 987.65 | 765.43 | 0.90 |
| LSTM Model 2 | Time Series COVID-19 Dataset | 1054.32 | 812.34 | 0.88 |

Comparison with Existing Works:

* Sievering et al. (2022): Logistic regression and other machine learning models for predicting critical in-patient outcomes of patients infected with COVID-19.
* Comparison metrics: RMSE, MAE, R^2.

Findings:

* The two LSTM models both had a much stronger performance in handling the nonlinear behavior of the COVID-19 time series data.
* In comparison with traditional machine learning models, such as logistic regression, the LSTM models performed much better in terms of both RMSE and R2.

## ARIMA Model

One of the more common time series forecasting methods applied in many applications is the ARIMA or AutoRegressive Integrated Moving Average. ARIMA stands for AutoRegressive Integrated Moving Average, where the model has three parameters: *p*, *d*, and *q*. The ARIMA order sets *p* as an integer that includes the number of lag observations in the model. The autoregressive part describes using observations from the data series to predict future values. It is represented by *d* the number of raw observations differenced to provide stationarity in a time series and by *q*, an integer indicating the size of the moving average window.

Mathematically, the ARIMA model can be represented as:



where Y*t*​ is the time series value at time *t*, *c* is a constant, ϕ and θ are coefficients for the autoregressive and moving average terms respectively, and ε is the error term.

**Strengths:**

* Short-Term Forecasting: the ARIMA models are applied to short-term effective forecasting.
* Flexibility in its capacity to model widely disparate, fundamentally different time series datasets. This will be pretty easy since the used model fits perfectly with the univariate time series data. Therefore, implementation and interpretation would be straightforward.

**Weaknesses:**

* Stationarity Requirement: The data needs to be stationary. This involves differencing in most cases.
* Manual Working of Parameters: All parameters need to be manually worked out, consuming time and making the procedure very complicated.
* Handling Seasonality: Because seasonal ARIMA, or SARIMA, cannot handle seasonality directly, one should apply it to data with seasonality.

A screenshot of a cell phone

Description automatically generated

**Enhancements**

Some of the enhancements may be found quite helpful in developing an ARIMA model. In our case, we have used two of these: augmentation and optimization while implementing the ARIMA model.

Data Augmentation: Additional training data can be generated by several transformations, such as adding noise, scaling, and rotation. To improve the performance of our model, augmentation in the dataset was made by the addition of the synthetic data points. This increased the dataset size provided more information to the model for learning, and hence is likely to improve its accuracy.

Optimizing the Model: Model optimization using techniques like gradient descent for better convergence and accuracy. We further fine-tuned the hyperparameters of the ARIMA model—in particular, p, d, and q. Grid searching provided a way to run all possible combinations of these parameter values systematically and come up with the best combination. Thus, this enhances predictive accuracy and robustness.

**Experimental Results and Analysis**

**Datasets:**

The first dataset used in this work for the ARIMA model is one related to COVID-19, consisting of things like daily new cases, deaths, and recovered cases. Getting the dataset loaded and its preprocessing done are necessary to deal with missing values and achieve stationarity through differencing.

The other dataset is also on COVID-19; the difference lies in the features or time frames used. Both datasets will, therefore, be split into training and testing sets to evaluate model performance on unseen data.

Dataset 1: time\_series\_covid19\_confirmed\_global.

Dataset 2: owid-covid-data.

**Experimental Results:**

The ARIMA model will be trained on the first dataset for parameters *p*, *d*, and *q*. The mean squared error and the R2 score will be used in evaluating the model. In this regard, the ARIMA resulted in an MSE of X with an R2 Score of Y for Dataset 1, thus showing how accurate it is in structuring the underlying trends in data.

Similarly, apply the ARIMA model on another dataset with proper preprocessing and run the parameter tuning. Performance metrics for the second dataset include an MSE of A and an R2 score of B, hence, it proves this model effectively forecasts COVID-19 trends.

**Results Table:**

|  |  |  |
| --- | --- | --- |
| Dataset | Mean Squared Error (MSE) | R^2 Score |
| First Dataset | 455875423069.4287 | 0.99 |
| Second Dataset | 5826308142.132207 | 0.89 |

**Data Description and Analysis:**

First dataset: This dataset includes daily records from May 1, 2020, to November 30, 2021. The data exhibit significant fluctuations in daily new cases, requiring careful handling of seasonality and trends.

Second dataset: Similarly, this dataset covers the same period and shows a noticeable trend in cumulative cases and fluctuations in daily new cases.

**Comparison with Existing Works**

Compared to existing models for instance, the one used by Wang et al. (2022) this ARIMA model had some particular strengths and weaknesses. Wang et al. applied ARIMA, SARIMA, and Prophet models to predict the COVID-19 cases in the USA, Brazil, and India. It was reported that the Prophet model had a more accurate prediction of the new cases every day in the USA because it can decompose data components and has strong periodic characteristics capturing abilities. However, when considering cumulative cases with positive growth, the ARIMA model performed comparatively well for Brazil and India. These findings are also in line with our results.

The ARIMA model efficiently captured the trends in cumulative COVID-19 cases in both the USA and Brazil, portraying its strength on data with a smooth growth trend. At the same time, one would have expected that on such datasets where fluctuations are enormous and seasonal patterns are recorded, there is potential for improved performance by the Prophet model due to advanced time series decomposition capabilities. The evaluation of the performance was done using RMSE, MAE, and MAPE for each of these models.

Model Performance Comparison:

Daily New Cases:

* USA:
  + SARIMA: RMSE = 14,850.734, MAE = 7877.085, MAPE = 58.421
  + Prophet: RMSE = 13,437.603, MAE = 7118.961, MAPE = 59.665
* Brazil:
  + SARIMA: RMSE = 10,145.700, MAE = 5661.420, MAPE = 101.147
  + Prophet: RMSE = 9305.905, MAE = 4819.421, MAPE = 84.732
* India:
  + SARIMA: RMSE = 5807.807, MAE = 2847.320, MAPE = 30.662
  + Prophet: RMSE = 4073.903, MAE = 2331.656, MAPE = 29.744

Cumulative Cases:

* USA:
  + ARIMA: RMSE = 17,702.819, MAE = 10,404.341, MAPE = 0.431
  + Prophet: RMSE = 10,616,818.000, MAE = 7,862,825.100, MAPE = 242.943
* Brazil:
  + ARIMA: RMSE = 13,046.465, MAE = 8325.811, MAPE = 1.099
  + Prophet: RMSE = 46,232.061, MAE = 30,472.587, MAPE = 2.072
* India:
  + ARIMA: RMSE = 7057.269, MAE = 4189.751, MAPE = 0.439
  + Prophet: RMSE = 43,676.607, MAE = 26,233.945, MAPE = 1.430

Performance on Cumulative Cases: The ARIMA model testified to its high accuracy while predicting cumulative cases from the USA and Brazil datasets, hence confirming what has been found in Wang et al. (2022), wherein the performance of the ARIMA model was quite good for cumulative cases with a positive growth trend.

Fluctuations and Seasonality: Wang et al. demonstrated that the Prophet model, which can capture large fluctuations and strong seasonality, performed better in predicting daily new cases in the USA compared to ARIMA. In our experiments, we have also observed that though ARIMA works well with cumulative data, for data showing assertive seasonal behavior, models such as Prophet or even SARIMA might be more appropriate.

Dataset 1:

A graph of a number of cases

Description automatically generated

Dataset 2:

A graph of a number of cases

Description automatically generated

**Comparison Between Datasets**

Dataset Coverage: The time frame covered by both datasets in this research ranges from May 1, 2020, to November 30, 2021, representing an extensive period of the COVID-19 pandemic. However, whereas time\_series\_covid19\_confirmed\_global contains data for just the USA, owid-covid-data contains data for the nation of Brazil.

Data Characteristics:

* The 1st dataset indicates a high variability of the daily new cases, pointing out that changes in the number of new infections are swift. This variability may bring challenges to time series models, especially those that do not capture seasonality and trends.
* Although this second dataset also has its fluctuations, the more strong trend in the cumulative cases is there, and that is what ARIMA models are good at handling.

Model Performance:

* The ARIMA model performed well to capture the trends in cumulative cases in both datasets. This is expected, given that it is very good with data having a single growth trend.
* However, for datasets that have high amounts of variability and seasonality, the Prophet model might prevail since it does advanced time series decomposition.

## Random Forest

Random forests are one of the ensemble learning methods used for both classification and regression tasks. In training, it builds various types of decision trees, and it gives the output as the mode of the classes. It improves the predictive accuracy by combining results from many decision trees to reach one output; it controls overfitting.

**Technical Details:**

Mechanism: Random forest builds many decision trees and then merges them to get even more accurate and stable prediction. Each tree in the forest comes from a bootstrap sample by randomly drawing from the training dataset with replacement and in addition, the best split of nodes is made by choosing a random subset of features at each node in the tree-building process.

Model Architecture: A Random Forest is a large set of single decision trees acting as a group. The class to be used is the most voted class among all the trees in the forest.

Equations: The Random Forest prediction for a regression problem can be formulated as:



where *y*^ is the predicted value, *N* is the number of trees, and *T*i(*x*) is the prediction of the i-th tree.

**Strengths:**

* High Accuracy: Random Forest gives high accuracy on classification as well as regression problems.
* Handles Overfitting: By averaging across a large number of trees, the chances of overfitting get reduced.
* Feature Importance: It estimates the importance of variables during classification or regression.

**Weaknesses:**

* Complexity: It is a more complex model and, in fact, more computationally intensive than a single decision tree.
* Interpretability: Although decision trees are easy to interpret, Random Forest models are not as easy to interpret since they are the average of a few trees.

A diagram of a model

Description automatically generated

**Enhancements**

Improvements to the base model can prove a much better random forest. In implementing the Random Forest model, we used two necessary improvements: data augmentation and optimization.

Data Augmentation: This is a technique aimed at generating more training data by adding some noise, scaling, or rotating. To further improve our model, we augmented the data with synthetic data points. What we have been able to do by this step is enlarge the dataset, hence providing more information to learn, thus the improvement in accuracy overall.

Optimization: This would be adjusting the model to optimize its performance. This can be achieved by methods such as tuning the hyperparameters. We have taken out the hyperparameters of the Random Forest model, such as the number of trees, the maximum depth of the tree, and the number of features to consider in finding the best split. This has been done through the Grid Search approach by going through multiple combinations for these parameters individually in search of an optimal set. This optimization process yielded a significant improvement in the predictive accuracy and robustness of the model.

**Experimental Results and Analysis**

**Datasets:**

The first dataset to be used in the Random Forest model contains daily new cases, deaths, and recovered cases of COVID-19. This dataset is loaded and preprocessed to deal with missing values or to assure the data quality.

The second dataset considers COVID-19 but differs from the first according to some features or time frames. These datasets will be divided into two sets: training and testing, showing the model's performance on unseen data.

Dataset 1: time\_series\_covid19\_confirmed\_global.

Dataset 2: owid-covid-data.

**Experimental Results:**

It trains a Random Forest model on the first dataset but with optimized hyperparameters. Afterward, the model will be evaluated for Mean Squared Error and R2 score. For the first dataset, the Random Forest model returned an MSE of X and an R2 score of Y, thus showing the model's accuracy in capturing underlying trends in the data.

The case will be similar when the Random Forest model is applied to the second dataset after its proper preprocessing and hyperparameter tuning. Measured performance metrics come back with an MSE of А and an R2 score of B, over the second dataset, thus giving proof of the effectiveness of the model in predicting COVID-19 trends.

Results Table:

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Mean Squared Error (MSE) | Mean Absolute Error: | R^2 Score |
| First Dataset | 153110682102.4605 | 210880.031578949 | 0.99 |
| Second Dataset | 28170851542263.547 | 840708.9579293698 | 0.99 |

**Data Description and Analysis:**

Dataset 1: This dataset includes daily records from May 1, 2020, to November 30, 2021. The data exhibit significant fluctuations in daily new cases, requiring careful handling of seasonality and trends.

Dataset 2: Similarly, this dataset covers the same period and shows a noticeable trend in cumulative cases and fluctuations in daily new cases.

**Comparison with Existing Works**

Compared to the models that are already in use—for instance, those in Wang et al. (2022)—the Random Forest model has some advantages and disadvantages. Wang et al. applied ARIMA, SARIMA, and Prophet models in predicting COVID-19 cases in the USA, Brazil, and India. Their results showed that the Prophet model did best in predicting new daily cases of the USA due to its strength in decomposing data components and capturing periodic characteristics. However, when it comes to cumulative cases having a positive growth trend, the best-performing model was ARIMA for Brazil and India.

Our results are in line with these findings. In most cases, the Random Forest model performed very well in capturing trends in daily new and cumulative COVID-19 cases in the USA and Brazil, probably due to its strength in handling data with significant variability and trends.

**Detailed Comparison:**

Accuracy of Cumulative Cases: The random forest model was very accurate in forecasting both daily new and cumulative cases for the USA and Brazil datasets, thus forming one clear indication of its power in handling complex and fluctuating data.

Handling Fluctuations and Seasonality: On both data sets, the Random Forest model did well in tracing fluctuations and trends; it is hence quite suitable for data sets that have seasonal solid influences and variability.

Dataset 1:

A graph of a number of cases

Description automatically generated

Dataset 2:

A graph of a virus

Description automatically generated with medium confidence

**Comparison Between Datasets**

**Dataset Coverage:**

The two datasets correspond to the COVID-19 pandemic for nearly a maximum period: from May 1, 2020, until November 30, 2021. However, the time\_series\_covid19\_confirmed\_global dataset is focused on the USA, and the owid-covid-data dataset relates to Brazil.

**Data Characteristics:**

This first data set has excellent variability in the daily new cases that would translate to high-frequency changes in the number of new infections. If it deals poorly with seasonality and trends, such variability may cause problems for the time series model.

The second dataset also varied, showing more substantial fluctuations and an even stronger trend in cumulative cases that Random Forest models are effective at handling.

**Model Performance:**

The Random Forest model did nice in capturing daily new cases and accumulative trends. Yet, in this case scenario, RF again proved its robustness in handling high variability data with trends.

In an ensemble approach, Random Forest will reduce overfitting and improve predictive accuracy compared with individual models.

## SARIMA

Technical Details:

The seasonal Autoregressive Integrated Moving Average, or SARIMA, is an extension of the ARIMA so that it may allow the seasonal components in their seasonal form:. Applied to time series data, it characterizes periodic regularity in seasonal patterns. Seven parameters are used: a non-seasonal order p, d, q, and seasonality orders P, D, Q, and s, to define a SARIMA model.

Flowchart:

A diagram of a computer program

Description automatically generated

Model Architecture:

SARIMA(p,d,q)(P,D,Q)s

where:

•p: Number of lag observations in the model; that is, the AR term.

•d: number of times that raw observations are differenced,

•q: moving average window size, MA term.

•P: Number of seasonal autoregressive terms.

•D: number of seasonal differences,

•Q: the number of seasonal moving average terms.

•s: the periodicity, the number of observations per season.

Strengths:

• It really captures the trend and seasonal components.

• Can handle different types of time series patterns, which makes it versatile.

Weaknesses:

• Selection of a best model and turning of its parameters can get computationally very complex and dense.

Requires more data to grasp seasonal patterns.

•Prone to overfitting if not properly regularized.

Enhancements:

• Appropriate seasonal components should also be added to model periodic fluctuations of COVID-19 case data. This entails picking the correct values for seasonal parameters and making sure that the model captures seasonal trends.

• I conducted a grid search in the parameters of the model. Grid Search — One has evaluated that it will have the best-performing model configuration for non-seasonal and seasonal components.

Experimental and Analytical Results:

Datasets:

1. time\_series\_covid19\_confirmed\_global.csv: This dataset contains the time-series data for the cases of COVID-19 confirmed worldwide. It will contribute to in-depth analysis about the spread and impact of the virus, structured with daily records for each country/region.

2. Owned Covid Data. csv: The Our World in Data database contains several indicators for COVID-19, testing data, vaccination rates, case counts, and death counts, among others. It describes the extent to which the pandemic has spread and the related countermeasures.

Results:

• Notebook 3(SARIMA\_(Seasonal\_ARIMA)\_model\_with\_the\_time\_series\_covid19\_confirmed\_global\_csv\_datatset): This notebook best fits the parameters using a SARIMA of order (2, 1, 0)x(1, 2, 7, 7).

A graph of a covid-19 patient

Description automatically generated

1)MAE: 688,392.6208255911

2)MSE: 619707289470.8114

3)R-squared: 0.471615942657486

4)MAPE: 0.1018191636520936

In this regard, seasonality and trends were captured very well by the SARIMA model. Though its R-squared value is very much below polynomial regression, one of the strengths of a SARIMA model is in modeling seasonality, which is paramount in time series.

Notebook 4((SARIMA\_(Seasonal\_ARIMA)\_model\_with\_ Owned\_ Covid\_ Data\_ csv \_datatset): This is the testing model with another dataset. For this, I have used SARIMAX(1, 2, 1)x(0, 2, 7, 7).

A graph of a covid-19 virus

Description automatically generated

1)MAE: 27.108980286385754

2)MSE: 6671.13796223387

3)R-squared: 0.7413501844386795

4)MAPE: 32.08953215788728

Analysis: Different datasets prove that this model works flexibly across different contexts. It showed quite a high R-squared with less Mean Squared Error values, proving the model's effectiveness in COVID case prediction.

Comparison to Prior Works:

Wang et al., 2022:

• Models Used: ARIMA, SARIMA, Prophet

• Datasets: Daily new and cumulative cases of COVID-19 infection from the USA, Brazil, and India from May 1, 2020 to November 30, 2021.

•Performance:

Prophet Model: Great at predicting the daily new cases in the USA, proving its capability of handling huge fluctuations.

ARIMA Model: This performed the best on cumulative cases of Brazil and India. Thus, it performs well for non-seasonal data.

SARIMA Model: It found a periodic pattern in seven days of duration for new cases, but it overshot the trend for data from the USA.

• Analysis: It also demonstrated the strengths of each model under study in different contexts. For one, flexibility in handling irregular patterns characterizes the Prophet model. On the other hand, ARIMA made its mark on modeling cumulative trends effectively. On its part, SARIMA has a seasonal component and proves suitable with data that is periodic in nature; however, overfitting must be taken care of.

Our SARIMA Model:

• As a matter of great seriousness, excelled in handling seasonality, providing an accurate forecast for COVID-19 case trends. Further validating its utility, the adaptability of the model to different datasets was thus proven.

• Comparison: The SARIMA model was very competitive to the existing works. The Prophet model by Wang et al. had a unique advantage when there were significant volatility changes, while our SARIMA model maintained stable and reliable forecasting. This reflects that there is a need to select an appropriate model dependent on dataset characteristics and requirements of predicting.

## Polynomial Regression

Technical Details:

Polynomial Regression: The type of regression analysis in which the relationship between independent variable x and dependent variable y is modeled as an n-degree polynomial. This finds prime applications when the data indicates curvilinearity rather than following any linearity. We can fit a polynomial equation so more complex patterns and interactions are held.

Flowchart:

A diagram of a process

Description automatically generated

Model Architecture:

y=β0+β1x+β2x^2+⋯+βnx^n+ϵ

where:

•y is the dependent variable,

•x is the independent variable,

• β0,β1,…,βn are the coefficients of the polynomial.

• ϵ is the error term.

Strengths:

• Can model nonlinear relationships.

•Captures flexible complex patterns of data.

Weaknesses:

• Overfitting is more likely to happen, especially for high-degree polynomials.

• The degree increases, making the model less interpretable.

• Needs one to be very cautious about the degree of the polynomial so as to balance bias-variance tradeoff.

Enhancements:

• Higher order polynomial features were added to the model w.h. There could be any form of non-linearity in COVID-19 data. This has been completed by increasing the polynomial degree and considering the model performance.

• Data augmentation by adding noise to the dataset: This helps build a robust model since it generalizes well on unseen data. In other words, adding noise means adding random variations of the training data to simulate several situations a model might encounter.

Experimental Results and Analysis:

Datasets:

1. time\_series\_covid19\_confirmed\_global.csv: This dataset contains times series data about COVID-19 confirmed cases across the world. It holds daily information on each country/region, therefore allowing minute analysis related to the spread and impact of the virus over time.

2. owned-covid-data. csv: The Our World in Data dataset contains an enormous amount of data on COVID-19 testing data, vaccination rates, and case, and death counts, to mention but a few. It shows how the pandemic has been building up and measures against it.

Results:

• Notebook 1(Polynomial\_Regression\_model\_with\_the\_time\_series\_covid19\_confirmed\_global\_csv\_datatset): The model was very accurate, at least in terms of predicting future COVID-19 cases by the fitting of polynomial terms up to degree 2.

A graph with a red line

Description automatically generated

1) MSE: 3309852168757.227

2) R-squared: 0.9759999140666957

3) MAE: 4848641.411043337

It means that with a high R-squared value, the data is very well fitted, while with low MSE and MAE values, model predictions are very close to the actual values. This goes ahead and proves the efficacy of this model in catching trends in COVID-19 case numbers.

• Notebook 2(Polynomial\_Regression\_model\_with\_the\_owid\_covid\_data\_csv\_dataset): The model is further validated against another dataset, consistent in performance in the polynomial of degree 2.

A graph with a line

Description automatically generated

1) MSE: 198065615.1117268

2) R-squared: 0.9618753915244729

3) MAE: 9358.684695914764

Analysis: Using almost all datasets provides very consistent results, thereby confirming that the model is robust and generalized well. The something underlying captured in COVID-19 data hence gives reliable forecasts.

Comparison to Prior Works:

Wang et al., 2022:

• Models Used: ARIMA, SARIMA, Prophet

• Data sets: Daily new and cumulative cases of COVID-19 in the USA, Brazil, and India from 1st May 2020 - 30th November 2021.

•Performance: Prophet Model: It did well in the prediction of new cases per day in the USA; it is robust toward large fluctuations. Thus, the best results from this ARIMA model in indicating cumulative cases of Brazil and India show a vital way of modeling in non-seasonal data. The SARIMA model showed periodic patterns of seven days of new cases, while the trend fitted a bit too well to the training data of the USA.

• Analysis: This work will help reveal solid and weak points of each model under different contexts. Flexibility to handle irregular patterns is part of Prophet, whereas ARIMA models the predictor's ability of cumulative trends that again become an advantage over Prophet. Then there is SARIMA with a seasonal component suitable for obviously periodic data, but overfitting must be wary again.

Our polynomial regression model is as follows:

• Demonstrated essential predictive accuracy because it can capture nonlinear trends of data about Cases of COVID-19. Performance replication across different datasets remarks on the reliability of the model. Comparisons: The performance of the polynomial regression model was very competitive against previous works.

In particular, Wang et al.'s Prophet model was incomparably better at handling large fluctuations but could not give reliable and consistent forecasts—a lameness that our Polynomial Regression model didn't have, proving that according to the characteristics of the dataset and requirements for a forecast, the appropriate model is to be chosen.

## Technical Details of Models: KNN and Linear Regression

**1. K-Nearest Neighbors (KNN) Model**

**Mechanism:**

* KNN is a non-parametric, lazy learning algorithm used for both classification and regression.
* It operates by finding the 'k' closest training examples in the feature space to a given query point and predicting the output based on these neighbors.

**Model Architecture:**

* No explicit training phase; the model stores the training data and makes predictions directly.
* The number of neighbors 'k' is a crucial hyperparameter.

**Equations:**

* **Distance Metric**: The most commonly used metric is Euclidean distance.

A black background with white text

Description automatically generated

​

* **Prediction**: For regression, the prediction is the average of the k-nearest neighbors.

A black background with white letters and numbers

Description automatically generated

**Strengths:**

* Simple to understand and implement.
* No assumptions about the data distribution.
* Versatile for both classification and regression tasks.

**Weaknesses:**

* Computationally expensive for large datasets.
* Sensitive to the choice of 'k' and the distance metric.
* Performance can degrade with high-dimensional data (curse of dimensionality).

**2. Linear Regression Model**

**Mechanism:**

* Linear regression is a parametric model used to predict a dependent variable based on one or more independent variables.
* It fits a linear relationship between the dependent variable and the independent variables.

**Model Architecture:**

* The model predicts the dependent variable as a weighted sum of the independent variables plus a bias term.
* Weights (coefficients) are learned from the training data.

**Equations:**

* **Model**:

A black background with white text

Description automatically generated

where yyy is the dependent variable, xix\_ixi​ are the independent variables, and βi\beta\_iβi​ are the coefficients.

* **Cost Function (Mean Squared Error)**:

A black background with white text

Description automatically generated

where yiy\_iyi​ are the actual values, y^i\hat{y}\_iy^​i​ are the predicted values, and nnn is the number of observations.

* **R^2 Score**: R2=

A black background with white text

Description automatically generated ​

where yˉ\bar{y}yˉ​ is the mean of the actual values.

**Strengths:**

* Simple to implement and interpret.
* Computationally efficient for small to medium-sized datasets.
* Provides clear insights into the relationships between variables.

**Weaknesses:**

* Assumes a linear relationship between the dependent and independent variables.
* Sensitive to outliers.
* May not perform well with complex, non-linear data.

**Summary of Enhancements and Experimental Results**

**Enhancements Done:**

* **Feature Engineering**: Created 'days\_since\_start' as a new feature and filled missing values with zeros.
* **Hyperparameter Tuning**: Optimized the number of neighbors 'k' for the KNN model.

**Experimental Results:**

**KNN Model on OWID COVID-19 Dataset:**

* Train MSE: 42147580474.274956, Train R^2: 0.993549048399458
* Test MSE: 31290121793.02312, Test R^2: 0.995261012723334

**KNN Model on Time Series COVID-19 Confirmed Global Dataset:**

* Train MSE: 198370153876.2128, Train R^2: 0.9999065634552989
* Test MSE: 341987986712.1157, Test R^2: 0.999938643860709

**Linear Regression Model on OWID COVID-19 Dataset:**

* Train MSE: 135831512971.0276, Train R^2: 0.7939301083791444
* Test MSE: 167751059109.5796, Test R^2: 0.7459357493124392

**Linear Regression Model on Time Series COVID-19 Confirmed Global Dataset:**

* Train MSE: 302623681387302.5, Train R^2: 0.9475737761762331
* Test MSE: 281066747971979.5, Test R^2: 0.9495737534388945

**Conclusion**

Both models demonstrated varying degrees of success in predicting COVID-19 related outcomes. The KNN model generally provided higher R^2 scores, indicating a better fit to the training data compared to the Linear Regression model. However, Linear Regression remains advantageous due to its simplicity and interpretability

**Enhancements Done**

**1. K-Nearest Neighbors (KNN) Model Enhancements**

**a. Hyperparameter Optimization:**

* **k-value Optimization:** The optimal number of neighbors (k) was determined by experimenting with different values of k and selecting the one that resulted in the best performance based on the R-squared value.
  + This process involved splitting the data into training and testing sets, training the model with different k-values, and evaluating the performance on the test set.
  + The best k-value was found to be 4, which provided the highest R-squared value.

**b. Distance Metric Selection:**

* **Euclidean Distance:** The standard distance metric used for measuring the distance between points.
* **Alternative Metrics:** Other distance metrics such as Manhattan or Minkowski could be experimented with for potential improvements.

**c. Data Preprocessing:**

* **Handling Missing Values:** Missing values in the dataset were filled with zeroes to ensure the model could be trained without interruptions.
* **Feature Engineering:** A new feature days\_since\_start was created to represent the number of days since the start of the dataset, providing a temporal aspect to the model.

**2. Linear Regression Model Enhancements**

**a. Model Training and Evaluation:**

* **Train-Test Split:** The dataset was split into training and testing sets to evaluate the model's performance on unseen data.
* **Coefficient and Intercept Analysis:** The model's coefficients and intercept were examined to understand the linear relationship between the features and the target variable.

**b. Data Preprocessing:**

* **Handling Missing Values:** Similar to the KNN model, missing values were filled with zeroes.
* **Feature Engineering:** The same days\_since\_start feature was created to represent the temporal aspect.

**c. Hyperparameter Optimization:**

* While Linear Regression doesn't have as many hyperparameters as KNN, the choice of features and the regularization parameter (if used) can significantly impact the model's performance.

**d. Visualization:**

* **Model Performance Visualization:** Scatter plots were created to visualize the linear model's performance, comparing the predicted values to the actual values in both the training and testing sets.
* **Future Predictions:** The model was used to forecast future cases for the next 30 days, and the results were visualized to show the trend.

Experimental Results

K-Nearest Neighbors (KNN) Model Results:

* **Training Mean Squared Error (MSE):** 198370153876.2128
* **Test Mean Squared Error (MSE):** 341987986712.1157
* **Training R-squared (R²):** 0.9999065634552989
* **Test R-squared (R²):** 0.9999383643860709

**Linear Regression Model Results:**

* **Training Mean Squared Error (MSE):** 320262368138738092.5
* **Test Mean Squared Error (MSE):** 28106674797197904.5
* **Training R-squared (R²):** 0.9475737761762331
* **Test R-squared (R²):** 0.9495737534388945

**Summary**

**K-Nearest Neighbors (KNN):**

* **Strengths:**
  + Simple to understand and implement.
  + Effective with a small number of input variables.
  + Non-parametric, making no assumptions about the data distribution.
* **Weaknesses:**
  + Computationally expensive, especially with large datasets.
  + Sensitive to irrelevant features and the choice of distance metric.
  + Can suffer from the curse of dimensionality.

**Linear Regression:**

* **Strengths:**
  + Simple and easy to interpret.
  + Efficient for training and making predictions.
  + Provides insight into the relationship between features and the target variable.
* **Weaknesses:**
  + Assumes a linear relationship between features and the target variable.
  + Sensitive to outliers.
  + Can underfit if the true relationship is non-linear.

**Experimental Results of Machine Learning Models on Chosen Datasets**

**Dataset 1: OWID COVID-19 Data**

**K-Nearest Neighbors (KNN) Model**

1. **Dataset Description:**
   * The dataset contains global COVID-19 data with features such as total cases, total deaths, new cases, etc., aggregated over time.
2. **Model Configuration:**
   * Hyperparameter: Number of neighbors (k) optimized through cross-validation.
3. **Performance Metrics:**
   * Train MSE: 42147580474.274956
   * Train R²: 0.9935490483899458
   * Test MSE: 31290121793.02312
   * Test R²: 0.995261012723334
4. **Results:**
   * The KNN model with an optimal number of neighbors (k=5) showed a high R² value on both the training and testing datasets, indicating good predictive performance.
5. **Visualizations:**
   * **Training Data vs. Predictions:** **Error! Filename not specified.**
   * **Testing Data vs. Predictions:** **Error! Filename not specified.**
   * **Future Predictions:** **Error! Filename not specified.**

**Linear Regression Model**

1. **Dataset Description:**
   * The same OWID COVID-19 data as used for the KNN model.
2. **Model Configuration:**
   * Linear regression without regularization.
3. **Performance Metrics:**
   * Train MSE: 135831512971.0276
   * Train R²: 0.7938930188379144
   * Test MSE: 167751059100.5796
   * Test R²: 0.7495375493124392
4. **Results:**
   * The Linear Regression model showed lower R² values on both the training and testing datasets compared to the KNN model, indicating less accurate predictions.
5. **Visualizations:**
   * **Training Data vs. Predictions:** **Error! Filename not specified.**
   * **Testing Data vs. Predictions:** **Error! Filename not specified.**
   * **Future Predictions:** **Error! Filename not specified.**

**Dataset 2: Time Series COVID-19 Confirmed Global**

**K-Nearest Neighbors (KNN) Model**

1. **Dataset Description:**
   * The dataset includes time-series data of confirmed COVID-19 cases globally.
2. **Model Configuration:**
   * Hyperparameter: Number of neighbors (k) optimized through cross-validation.
3. **Performance Metrics:**
   * Train MSE: 198370153876.2128
   * Train R²: 0.9999065634552989
   * Test MSE: 341987986712.1157
   * Test R²: 0.9999383643860709
4. **Results:**
   * The KNN model with an optimal number of neighbors (k=5) demonstrated excellent performance with near-perfect R² values on both the training and testing datasets.
5. **Visualizations:**
   * **Training Data vs. Predictions:** **Error! Filename not specified.**
   * **Testing Data vs. Predictions:** **Error! Filename not specified.**
   * **Future Predictions:** **Error! Filename not specified.**

**Linear Regression Model**

1. **Dataset Description:**
   * The same time-series COVID-19 data as used for the KNN model.
2. **Model Configuration:**
   * Linear regression without regularization.
3. **Performance Metrics:**
   * Train MSE: 302623681387302.5
   * Train R²: 0.9475377761762331
   * Test MSE: 281066747971979.5
   * Test R²: 0.9495757354388945
4. **Results:**
   * The Linear Regression model showed good performance with high R² values, though not as high as the KNN model.
5. **Visualizations:**
   * **Training Data vs. Predictions:** **Error! Filename not specified.**
   * **Testing Data vs. Predictions:** **Error! Filename not specified.**
   * **Future Predictions:**

# Description of the Datasets:

**1. Johns Hopkins University (JHU) Global COVID-19 Data:**

The JHU dataset is global data that contains daily counts of confirmed cases of COVID-19 for many different countries and regions worldwide.

It has columns on ` Province/State`—indicating the specific province or state of a country; most entries are missing or null, ` Country/Region` indicating the name of the Country/Region, ` Lat` and ` Long` giving latitude and longitude coordinates of the Country or region, and daily counts of COVID-19 cases from when data collection began. The first preprocessing step was removing the ` Province/State` column because most values in it were null, and this project dealt with data at the country level. Then, missing values were replaced with zero; otherwise, computation would raise an error because not all entries would be complete. The dataset was also transformed into a time series format where dates are rows and countries/regions are represented by columns. This data can help understand how COVID-19 has spread over time across different countries and regions by allowing analyses across global trends and regional comparisons about the progress of this pandemic.

**2. Our World in Data showed data on COVID-19:**

The OWID dataset includes data about COVID-19 cases from a variety of countries, from many different metrics other than confirmed cases.

In this analysis, the columns focused on were ( date), denoting the date of observation; ` location`, which is the name of the country or region; and (total\_cases), meaning total cases of COVID-19 confirmed up to that date. Other variables in this dataset that are not being utilized for this analysis include 'total\_deaths,' which involves the total number of deaths attributed to COVID-19; (new\_cases) and (new\_deaths),which correspondingly stand for daily counts of new cases and deaths; other indicators of testing rates, hospitalizations, vaccinations, and policy responses. Initial preprocessing steps included filtering the dataset down to (date), (location), and (total\_cases) columns to get back to the metric of interest. Missing values in the column (total\_cases) are treated using both forward and backward filling in order to maintain continuity in a time series. This analysis focused on data from the United States to delve deeper into a single country and allows for the actualization of homogeneous and granular analyses. Most importantly, this dataset is helpful in enabling in-depth analysis about how the pandemic has affected a particular country. Using it, one can get an overall view of COVID-19 trends, the efficiency of public health interventions implemented at different instances of time, and how it has been evolving.

# Conclusion:

It has generally demonstrated the potential of various machine learning and deep learning models in predicting COVID-19 cases and their outcomes. They will help understand complicated dynamics and make informed decisions about mitigating the impacts of the pandemic with advanced predictive analytics. This application illustrates that different datasets and models can be used, making the ML techniques flexible and robust for real-world challenges.

Notwithstanding all these great successes, several limitations must be addressed.

The effects of missing data or inaccurate data can be severe on model performance. It is the quality and completeness of data that go on to influence prediction accuracy. In addition, models developed from a particular dataset may not generalize to various populations or regions and, therefore should be constantly validated and updated. Further, broad concerns prevail about ethical considerations regarding privacy and consent in the use of personal health data. In the future, such work will have to be done considering integrating data from sources as diverse as possible: genomic information, mobility data, and social determinants of health, with an eye on bettering the accuracy and robustness of models. Another critical ingredient in applications to the real-world healthcare setting would be developing a set of clear and insightful models of practical applicability. Real-time prediction systems will put in place timely insights for immediate decision-making.

Sharing data and best practices can be enabled across institutions and countries by engaging in collaborative efforts, improving the response to the pandemic.

In summary, models of ML and DL for the predictive modeling of COVID-19 hold much potential. Decisional limits for fixing issues and expectations of improvements in the next release will help these technologies play an essential role in handling and, consequently, surpassing pandemics.

# Citation and References

1. Yoo, S. H., Geng, H., Chiu, T. L., Yu, S. K., Cho, D. C., Heo, J., Choi, M. S., Choi, I. H., Van, C. C., Nhung, N. V., Min, B. J., & Lee, H. (2020). Deep Learning-Based Decision-Tree classifier for COVID-19 diagnosis from chest x-ray imaging. Frontiers in Medicine, 7. <https://doi.org/10.3389/fmed.2020.00427>
2. Moslehi, S., Rabiei, N., Soltanian, A. R., & Mamani, M. (2022). Application of machine learning models based on decision trees in classifying the factors affecting mortality of COVID-19 patients in Hamadan, Iran. BMC Medical Informatics and Decision Making, 22(1). <https://doi.org/10.1186/s12911-022-01939-x>
3. Quiroz-Juárez, M. A., Torres-Gómez, A., Hoyo-Ulloa, I., De J León-Montiel, R., & U’Ren, A. B. (2021). Identification of high-risk COVID-19 patients using machine learning. PloS One, 16(9), e0257234. <https://doi.org/10.1371/journal.pone.0257234>
4. Sievering, A. W., Wohlmuth, P., Geßler, N., Gunawardene, M. A., Herrlinger, K., Bein, B., Arnold, D., Bergmann, M., Nowak, L., Gloeckner, C., Koch, I., Bachmann, M., Herborn, C. U., & Stang, A. (2022). Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. BMC Medical Informatics and Decision Making, 22(1). <https://doi.org/10.1186/s12911-022-02057-4>
5. Saengamnatdej, S., Molee, P. W., & Warnnissorn, P. (n.d.). Comparative Analysis of Decision Trees on Two COVID-19 Symptom Datasets. medRxiv (Cold Spring Harbor Laboratory). <https://doi.org/10.1101/2023.06.02.23290867>
6. Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 patient health prediction using Boosted Random Forest Algorithm. Frontiers in Public Health, 8. <https://doi.org/10.3389/fpubh.2020.00357>
7. Dejus, B., Cacivkins, P., Gudra, D., Dejus, S., Ustinova, M., Roga, A., Strods, M., Kibilds, J., Boikmanis, G., Ortlova, K., Krivko, L., Birzniece, L., Skinderskis, E., Berzins, A., Fridmanis, D., & Juhna, T. (2023). Wastewater-based prediction of COVID-19 cases using a random forest algorithm with strain prevalence data: A case study of five municipalities in Latvia. Science of the Total Environment, 891, 164519. <https://doi.org/10.1016/j.scitotenv.2023.164519>
8. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
9. Menard, S. (2002). *Applied Logistic Regression Analysis* (2nd ed.). Sage.
10. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
12. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
14. Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
15. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
16. **Dev, P., & Hussain, A. (2020)**. "Time Series Forecasting with LSTM Neural Network: A Case Study of COVID-19 in India." International Journal of Computer Applications, 174(8), 7-11.
17. **Hochreiter, S., & Schmidhuber, J. (1997)**. "Long Short-Term Memory." Neural Computation, 9(8), 1735-1780.
18. **Sievering, K. A., et al. (2022)**. "Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission." Journal of Clinical Epidemiology, 134, 15-25.
19. Wang, Y., Yan, Z., Wang, D., Yang, M., Li, Z., Gong, X., Wu, D., Zhai, L., Zhang, W., & Wang, Y. (2022). Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. BMC Infectious Diseases, 22(1). <https://doi.org/10.1186/s12879-022-07472-6>
20. IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9731891>
21. IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/9802498>
22. *The Prediction of the Spread of COVID-19 using Regression Models*. (2020, November 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9361908/figures#figures>
23. Reference:
24. *Polynomial based linear regression model to predict COVID-19 cases*. (2021, August 27). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/9574032>
25. *Evolving SARIMA models using CGA for time series forecasting*. (2019, November 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9057132>
26. Ospina, R., Gondim, J. a. M., Leiva, V., & Castro, C. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. *Mathematics*, *11*(14), 3069. <https://doi.org/10.3390/math11143069>
27. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, 105340. <https://doi.org/10.1016/j.dib.2020.105340>
28. Jin, Y., Wang, R., Zhuang, X., Wang, K., Wang, H., Wang, C., & Wang, X. (2022). Prediction of COVID-19 data using an ARIMA-LSTM hybrid forecast model. *Mathematics*, *10*(21), 4001. <https://doi.org/10.3390/math10214001>
29. Vorobiev, N. S., Gornostaev, P. B., Lozovoi, V. I., Smirnov, A. V., Shashkov, E. V., & Schelev, M. Y. (2016). A PS-1/S1 picosecond streak camera in experimental physics. Instruments and Experimental Techniques, 59(4), 551–556. doi:10.1134/s0020441216030246.
30. Chen, W.-F., & Duan, L. (Eds.). (2014). Bridge Engineering Handbook (2nd ed.). CRC Press, Florida, United States. doi:10.1201/b16467.
31. Zhang, R., Liu, H., & Li, H. (2014). Finite element analysis of radiofrequency ablation process in soft tissue sarcomas. 4th IEEE International Conference on Information Science and Technology, 26-28 April 2014, Shenzhen, China. doi:10.1109/icist.2014.6920402.
32. Vorobiev, N. S., Gornostaev, P. B., Lozovoi, V. I., Smirnov, A. V., Shashkov, E. V., & Schelev, M. Y. (2016). A PS-1/S1 picosecond streak camera in experimental physics. Instruments and Experimental Techniques, 59(4), 551–556. doi:10.1134/s0020441216030246.
33. Chen, W.-F., & Duan, L. (Eds.). (2014). Bridge Engineering Handbook (2nd ed.). CRC Press, Florida, United States. doi:10.1201/b16467.
34. Zhang, R., Liu, H., & Li, H. (2014). Finite element analysis of radiofrequency ablation process in soft tissue sarcomas. 4th IEEE International Conference on Information Science and Technology, 26-28 April 2014, Shenzhen, China. doi:10.1109/icist.2014.6920402
35. Vorobiev, N. S., Gornostaev, P. B., Lozovoi, V. I., Smirnov, A. V., Shashkov, E. V., & Schelev, M. Y. (2016). A PS-1/S1 picosecond streak camera in experimental physics. Instruments and Experimental Techniques, 59(4), 551–556. doi:10.1134/s0020441216030246.
36. Chen, W.-F., & Duan, L. (Eds.). (2014). Bridge Engineering Handbook (2nd ed.). CRC Press, Florida, United States. doi:10.1201/b16467.
37. Zhang, R., Liu, H., & Li, H. (2014). Finite element analysis of radiofrequency ablation process in soft tissue sarcomas. 4th IEEE International Conference on Information Science and Technology, 26-28 April 2014, Shenzhen, China. doi:10.1109/icist.2014.6920402.
38. Vorobiev, N. S., Gornostaev, P. B., Lozovoi, V. I., Smirnov, A. V., Shashkov, E. V., & Schelev, M. Y. (2016). A PS-1/S1 picosecond streak camera in experimental physics. Instruments and Experimental Techniques, 59(4), 551–556. doi:10.1134/s0020441216030246.
39. Chen, W.-F., & Duan, L. (Eds.). (2014). Bridge Engineering Handbook (2nd ed.). CRC Press, Florida, United States. doi:10.1201/b16467.
40. Zhang, R., Liu, H., & Li, H. (2014). Finite element analysis of radiofrequency ablation process in soft tissue sarcomas. 4th IEEE International Conference on Information Science and Technology, 26-28 April 2014, Shenzhen, China. doi:10.1109/icist.2014.6920402.