Sohar University Journal of Engineering and Information Technology Innovations (SUJEITI), ISSN: 3078-7513
Volume 1, Issue 1, pp 58-67, January 2025, DOI: xxxxxxxxxxxxxxx
Received: 16/7/2024 Revised: 1/9/2024 Accepted: 30/12/2024

Ensemble Machine Learning for COVID-19 Forecasting: Enhancing Resource Planning and Pandemic Response in Oman

Wasin Al Kishri[1,*], Mohammed J. Yousif Almutoory[2]
[1]Faculty of Computer Studies, Arab Open University, Muscat, Oman.
[1]Department of Computer Science, Faculty of Science, Memorial University, Canada.

* Corresponding author: Wasin Al Kishri (wasan.k@aou.edu.om).

ABSTRACT

This present study proposes a machine-learning approach for predicting COVID-19 infection and death rates to support government resource planning in Oman. It compares three algorithms, namely Decision Tree, Random Forest, and Gradient Boost, for the best model that provides an accurate prediction of COVID-19. The Decision Tree model overfitted and gave accuracy of 99.41% in training and 53.39% in testing. On the other hand, the Random Forest model generalized better with 94.66% training accuracy versus 61.32% testing accuracy. The Gradient Boost model achieved 92.96% training accuracy and 59.44% testing accuracy but needs further tuning. A correlation analysis between the COVID-19 metrics has been presented. From the given heat map, daily new cases versus active cases represent a strong positive relation: increased active cases due to increased daily new cases. Overall case and death show a negative relation-an indication of reduced mortality rate. Detailed validation shall establish the fact that improved health services and vaccination campaigns were working in proper direction. Random Forest model was more accurate and generalized for the COVID-19 trend than the other models compared. The results from the Gradient Boost model were similar, but its performance needs further optimization. The overall findings from the study are vital in informing better public health policy improvements and effective resource management of the pandemic. This research thus contributes to the development of an efficient predictive tool to manage COVID-19 in Oman, using state-of-the-art machine learning techniques.

Keywords: Machine Learning, Optimization Models, Forecasting Models, COVID-19, Oman.

1. Introduction

According to the WHO, since the emergence of COVID-19 in 2023, millions around the world have been infected by the novel coronavirus, which has caused immense loss of life (WHO, 2024). In Oman, the pandemic of COVID-19 caused by SARS-CoV-2 started on February 24, 2020, when two arriving persons from Iran tested positive for the virus (Khamis et al., 2020). As of May 19, 2022, Oman had reported 389,473 confirmed cases, 4,260 deaths, and 384,669 recoveries (OmanvsCOVID19, 2023). The COVID pandemic's across-the-board impact on the individuals and the global economy highlighted the key concerns for governments and societies:

a. the potential recurrence of a similar pandemic,
b. transmission patterns derived from COVID-19 data,
c. the risk of implementing widespread closures, and
d. government readiness to handle future pandemics.

To address these challenges, (Boccaletti et al., 2020) identified three interdisciplinary scientific communities:

a. applied mathematicians, virologists, and epidemiologists who develop advanced diffusion models for pathogens,
b. complex systems scientists employing compartmental models, statistical mechanics, and nonlinear dynamics to study infection spread, and
c. researchers using artificial intelligence (AI) and deep learning to create precise predictive models.

Complementary studies further analyze COVID-19 societal impact by forecasting case trajectories and identifying transmission factors. Techniques include time series analysis (Becerra et al., 2020), modeling COVID-19 spread (Fanelli & Piazza, 2020), Monte Carlo decision-making (Fong et al., 2020), real-time forecasting (Roosa et al., 2020) and tracking public interest through Google Trends (Effenberger et al., 2020).

Accurate disease forecasting is important in the management of public health and, thus, enables strategic planning for addressing any future epidemic. AI models have become essential in advanced improvement regarding the accuracy of the forecast for epidemiological time series (Yousif & Saini, 2020). Vaishya (Vaishya et al., 2020) summarized AI applications in analyzing COVID-19 data and underlined its key role in public health.

The following manuscript proposes a new outbreak analytics approach for COVID-19, aiming to move beyond the traditional metrics of confirmed cases and fatalities in estimating the total number of potential infections. Using an ensemble machine-learning algorithm, this paper examines the infection and death rates in forecasting the actions that can be taken by policymakers to anticipate and dampen the impact of the pandemic across states in the Sultanate of Oman.

Section 2 of the manuscript follows with a critical review of the literature, offering a deep overview of related works and basic concepts. Further on, Section 3 goes into a detailed description of Materials and Methods, presenting data collection and preprocessing along with the description of the experimental setup. The architecture and application of Machine Learning Models used in this work are described in full detail in Section 3. Section 4.2. compares the performances of these models in terms of accuracy, robustness, and limitations. Section 4.3. discusses the Predictive Model of the study, with results and their implications. Finally, the manuscript concludes in Section 5, summarizing the results and offering recommendations for future research directions.

2. LITERATURE REVIEW

The COVID-19 pandemic has posed enormous public health challenges and disrupted the global economy to the extent that it caused a recession with high unemployment and unprecedented debt. Diagnosis and proper management of the disease continue to be very challenging due to limitations in diagnostic tools such as RT-PCR, and the absence of specific clinical or radiological findings.

Recent studies highlight various strategies examined how to mitigate the spread of COVID-19 and enhance diagnostic precision by applying advanced technologies in:

2.1. Deep Transfer Learning (DTL)

A study by Sufian (Sufian et al., 2020) discussed the potency of DTL in detecting infection during the COVID pandemic and its control. DTL is a technique of knowledge acquisition across tasks that requires less data and computational power with high accuracy. Integration of drones, IoT, and AI further enhances disease management.

COVID-19 Patient Detection Systems: Shaban et al. (2020) presented the COVID-19 Patients Detection Strategy (CPDS), a combination of Hybrid Feature Selection Methodology (HFSM) and Enhanced K-Nearest Neighbor (EKNN). The EKNN proved faster and more accurate than HFSM.

X-ray and CT Analysis: Tuncer et al. (2020) proposed a system that preprocessed the images, extracted features, and then classified them using techniques such as SVM and kNN to identify COVID-19, with accuracies reaching 100%.

Diagnostic Techniques: Yüce et al. (2021) reviewed diagnostic methods like RT-PCR, antibody detection, and antigen detection, emphasizing the importance of early and accurate identification.

2.2. Predictive Models

Deep Learning Models: Elsheikh et al. (2021) estimated a deep learning model that predicted cases, recoveries, and deaths, outperforming the ARIMA and NARANN models.
Ensemble models: The combination of ensemble learning with support vector regression has also achieved a very minimal error rate in some recent articles about the prediction of COVID-19 cases, in Ribeiro et al. (2020), which reported 0.87%.
CT Imaging Analysis: Li et al. (2020) conducted a deep learning analysis on CT scans for distinguishing COVID-19 from pneumonia, demonstrating sensitivity and specificity of over 90%.

2.3. Mathematical and Simulation Models

Sun & Wang (Sun & Wang, 2020) and Abdo (Abdo et al., 2020) used differential equation models and simulation systems to study transmission dynamics and predict unreported cases (Yousif et al., 2021). Such models are useful in decision-making processes for outbreak containment.

2.4. Advanced AI Techniques

WOCLSA Model: Su (Su et al., 2023) proposed a convolutional deep-learning model with Whale Optimization for parameter tuning, which achieved high predictive accuracy.
SCWOA Ensemble Paradigm: Qu (Qu et al., 2023) developed an ensemble neural network model optimized by a novel heuristic algorithm that outperformed comparable methods in terms of accuracy and robustness.

These studies really pinpoint the important contribution of AI, deep learning, and mathematical modeling in furthering the understanding, diagnosis, and management of COVID-19. Advanced techniques can enhance the accuracy of forecasting with better-informed public health decisions that will reduce the impact of the pandemic.

3. METHODOLOGY

It will be a two-stage process that includes preprocessing and processing. The preprocessing stage shall focus on the collection, filtration, and cleaning of data for good quality input, while processing involves the use of machine learning techniques, including but not limited to RNNs for time-series analysis, VGG-19 for classification tasks, and GRNs to model gene interaction. This research will be furthered using cleaned dataset-driven training of neural networks, evaluation via accuracy metrics, and cross-validation to ensure reliability and robustness. The methodology therefore aims at producing accurate classification and detection results from the objectives outlined in this study.

Machine learning algorithms build mathematical models of sample data to make predictions about or decisions for that data without being explicitly programmed to perform that particular job. In enterprise applications, machine learning methods are generally part of predictive analytics. The workflow followed by the procedure of machine learning, as depicted in Algorithm 1, which is composed of the following major steps:

- Data Acquisition: Gathering relevant data is a critical first step to ensure accurate forecasting and effective model training.
- Data Formatting: The collected data is formatted or transformed into a structure understandable by machine learning models.
- Exploratory Data Analysis (EDA): EDA is conducted to uncover patterns, relationships, and insights in the data.
- Data Splitting: The dataset is divided into training data (used to train the model) and testing data (used to evaluate its performance).
- Model Training: The training data is fed into the machine learning algorithm, enabling it to learn patterns and trends in the data.
- Prediction: Once trained, the model predicts the future trends or outcomes based on new or unseen data.
- Model Evaluation: The model's performance is assessed using the testing dataset, ensuring its accuracy and reliability.

```
Algorithm: Machine Learning Workflow for Data Processing, Model Training, and Evaluation
Input: Set of data
Output: Optimized model

Step 1: Data Acquisition
  - Collect data from various sources (e.g., databases, APIs, sensors).
  - Store the acquired data in a structured format for processing.
Step 2: Data Formatting
  - Check for missing values and inconsistent data types.
  - Convert raw data into a consistent format suitable for analysis.
  - Label the dataset if it's supervised learning.
Step 3: Data Pre-processing
  - Normalize or standardize numerical features.
  - Encode categorical variables.
  - Handle missing values by using methods like imputation or removal.
  - Remove duplicates and irrelevant data points.
  - Perform feature selection or extraction if necessary.
Step 4: Splitting of Data
  - Divide the dataset into:
    - Training Set: Used for model training.
    - Testing Set: Used for model evaluation.
Step 5: Machine Learning Model - Training Phase
  - Initialize the Machine Learning model (Estimator).
  - Use the training set to fit the model:
    Model.fit(training_data, training_labels)
Step 6: Machine Learning Model - Transformation Phase
  - Apply transformations if necessary to enhance feature representation.
  - Use Transformer to improve feature space before evaluation.
Step 7: Evaluation
  - Use the testing set to evaluate the model:
    Model.evaluate(testing_data, testing_labels)
  - Calculate performance metrics such as:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
Step 8: Final Output
  - Display evaluation results.
  - Optimize the model, if necessary, based on evaluation performance.
  - Save the trained model for future predictions.

End of Algorithm
```

Figure 1 represents a structured approach where machine learning models are bound to be well-trained, better evaluated, and capable of providing actionable insights for both forecasting and decision-making.

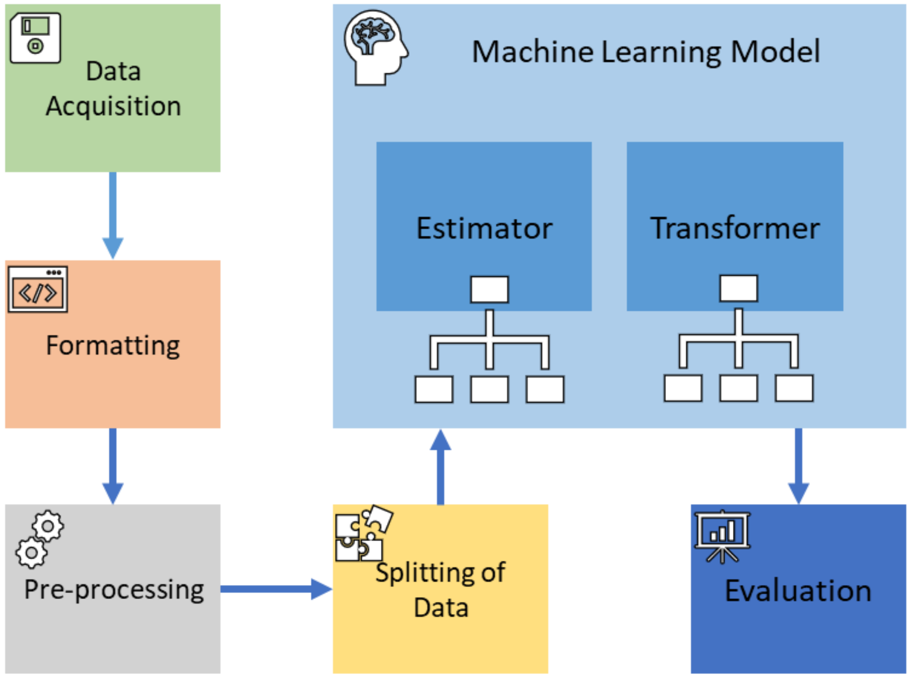

Figure 1. An Example of a Chart Represented in a Shaded Pattern.

4. RESULTS AND DISCUSSION

4.1. Analysis of the data

It elaborates on a detailed analysis of data and forecasting models used in this work, followed by insightful suggestions and recommendations based on findings; it also discusses potential future directions for further exploration and development. Exploratory Data Analysis is a crucial step in data analysis, and it acts like the very basis for further statistical or machine learning steps on a given dataset. It's about the iteration of examining and visualizing the data to uncover patterns, find anomalies, and obtain insights that are actionable in nature. This stage then essentially becomes the primer on which further analysis must rest to guarantee integrity in the models.

Figure 2 shows a schematic of this process, emphasizing its place as a preparatory step in the analytical continuum. This research work lays a very strong foundation for accurate forecasting and meaningful recommendations, since it systematically analyzed and interpreted data trends during EDA. Further directions and areas of development will expand the scope and applicability of this research.

The correlation heatmap of Figure 3 shows a correlation matrix of different metrics of COVID-19 in Oman. In this study, there was a high positive correlation between the Daily New Cases (Xnew) and Active Cases (Xactive), represented by $\rho > 0$, signifying that for any increase in the number of new daily cases, the number of active cases would be growing upwards. The values of $\rho < 0$ indicate that there is a negative correlation between the Total Coronavirus Cases, Xtotal, and the Total Coronavirus Deaths, Xdeaths. It might indicate the tendency of a decrease in mortality rate with increasing total cases, probably reflecting the efficient healthcare service and wider vaccination campaigns. However, further statistical tests must be done in order to confirm this hypothesis stringently and interpret the observed correlations validly.

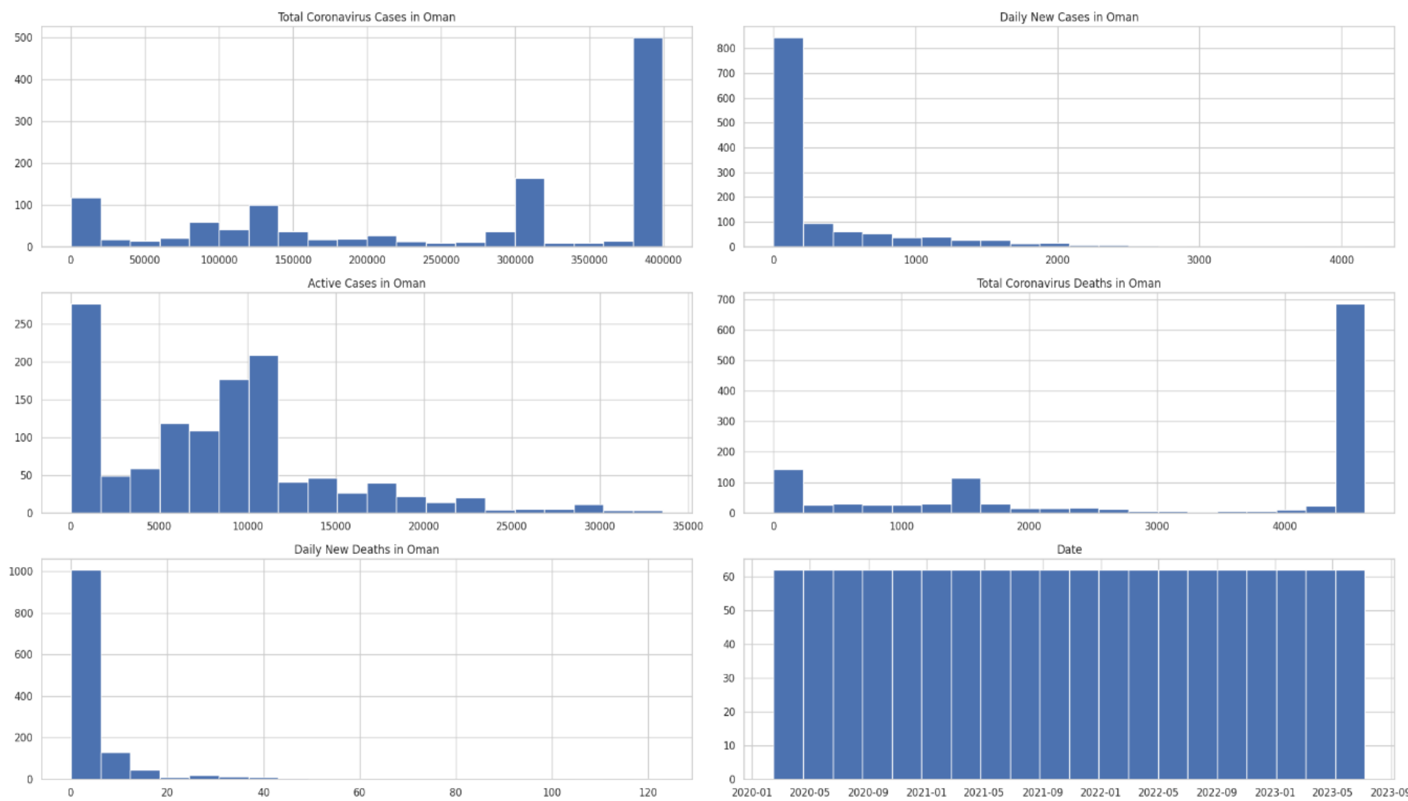

Figure 2. EDA explained using sample Data set.

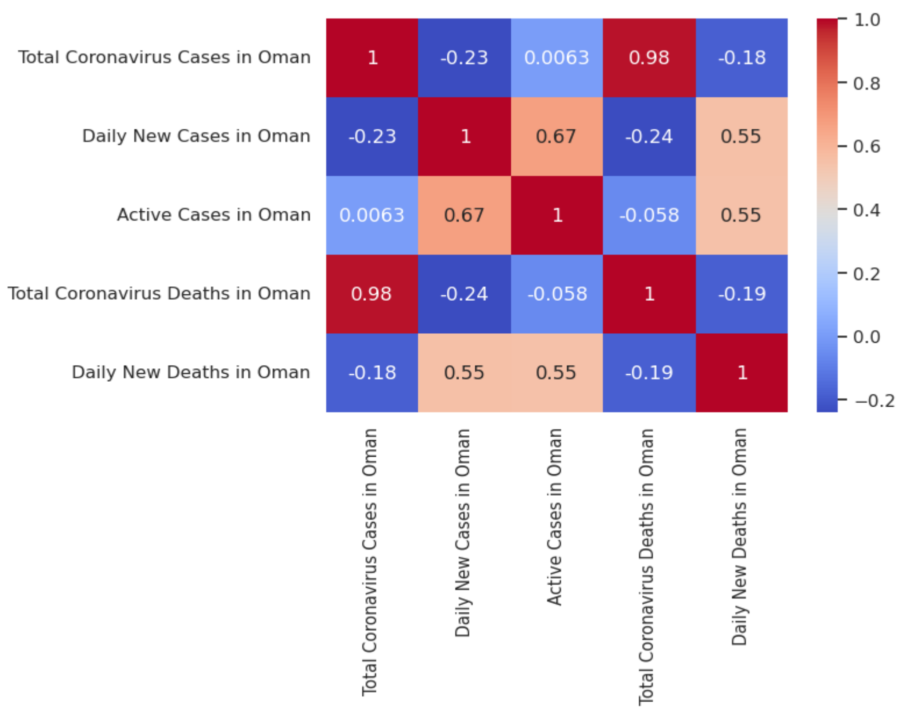

Figure 3. Correlation Heat Map

4.2. Models Comparison

Figure 4 shows the comparisons among the three machine learning algorithms: The Decision Tree model is overly fit since its training accuracy reaches a very high of 99.41%, whereas its test accuracy went down to as low as 53.39%. While the Random Forest model yields a slightly lower training accuracy of 94.66%, its testing accuracy increases to 61.32%, which would indicate better generalization and less overfitting. The Gradient Boost model reaches an accuracy of 92.96% on training and 59.44% on testing, making it sit in between the Decision Tree and Random Forest models for both training and testing performance. Although performing quite well on training data, this model has to be further optimized or regularized for better generalization because of the drop in testing accuracy.
Overall, the Random Forest model outperforms the Decision Tree and Gradient Boost models on testing data, showing better generalization capability. The Gradient Boost model is close to Random Forest in performance but with a little lower testing accuracy, making it a competitive alternative with further potential for tuning.

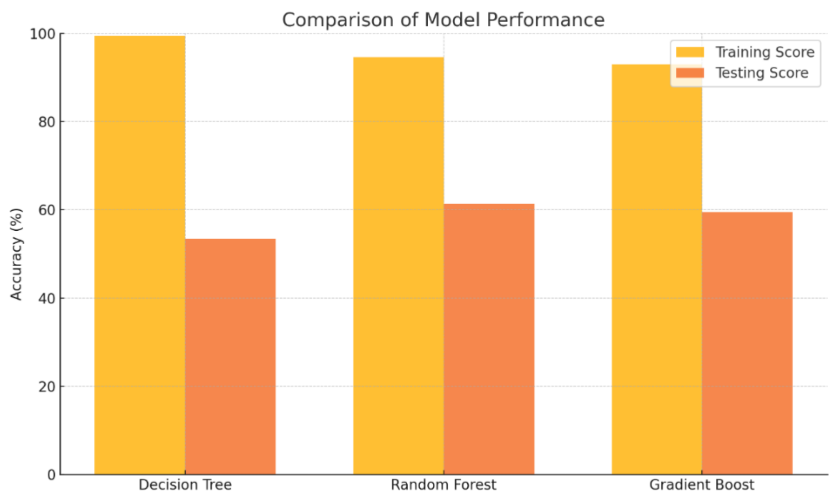

Figure 4. Comparing the Results of the Three Algorithms

4.3. Predictive Models

Comparing the models, it can be observed that among all the models, the Decision Tree Regressor is much better in terms of accuracy. Predictive analytics were performed on estimating whether an event will occur or not based on historical data by both Random Forest Regressor and Decision Tree Regressor.
In this work, the models were applied on a certain input dataset to predict an outcome: [813, 86, 679, 4]. The Random Forest Regressor predicted the value 0.2, indicating low likelihood of the predicted outcome. In contrast, the Decision Tree Regressor predicted a value of approximately 1.32, which is a higher likelihood for the outcome to happen, as depicted in Figure 5.
Model selection is one of the most important aspects of predictive analysis, as the accuracy and reliability of the predictions depend a lot on the model selected. Based on the results for this dataset, the Decision Tree Regressor outperforms the Random Forest Regressor. Again, this emphasizes that model selection needs to be specific to data and the objective of the prediction.
These insights can then be applied in strategic domains like finance, medicine, or engineering, therefore supporting better-informed decision-making processes due to more accurate predicted outcomes.

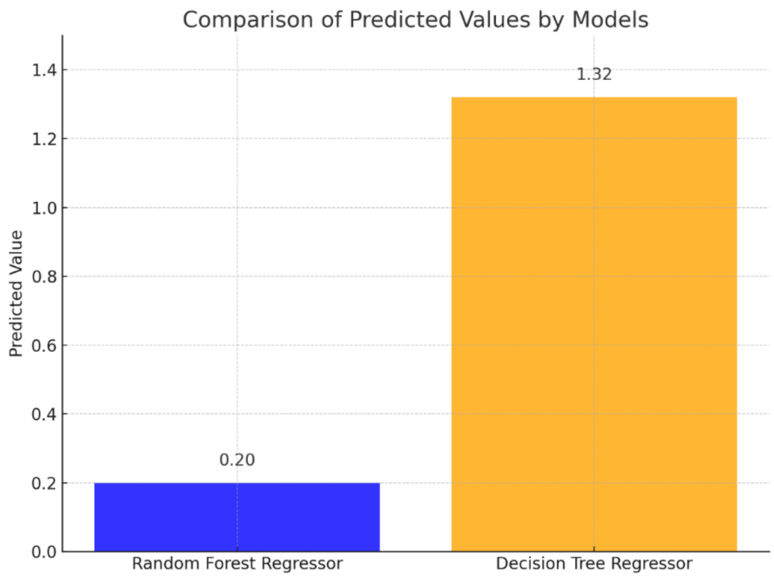

Figure 5: Comparing of predicted values by models

5. CONCLUSION

The COVID-19 pandemic has badly hurt the economy globally; recession, high unemployment, and high debt levels were recorded as results of the pandemic. Some of these challenges are being handled by innovative methods to enhance pandemic management: Deep Transfer Learning, Hybrid Feature Selection, and other models using deep learning. Machine learning will, therefore, be used in outbreak analytics for predicting the number of future cases and improving disease detection. It can be used with equal efficiency in both offline and online systems. This study presents an improved machine-learning framework for predicting COVID-19 cases and mortality rates in Oman, focusing on practical applications of advanced models to support government planning and healthcare resource allocation. It compares three machine learning models: Decision Tree, Random Forest, and Gradient Boost; the results show that the Random Forest model has the best generalization performance on testing data, making it more reliable in real-world applications.

This correlation analysis points out the main patterns in the COVID-19 metrics, ranging from high positive daily new cases and active cases to negative total cases and deaths. This could therefore mean that Omani people are experiencing improvement in health status. The important thing here is a combined approach using the machine learning predictions and detailed statistical analyses in developing actionable insights on the COVID-19 pandemic.

The optimization of the Gradient Boost model should be continued in further works, and additional feature sets could include vaccination rates or regional healthcare capacities to enhance the predictive power of such models. Integration of deeper complex architectures with real-time updates may further enhance the reliability of the predictions for more effective and timely responses against the dynamically changing public health challenges. The findings of this study, therefore, shall be important for the government agencies and healthcare institutions in developing a basis to make better informed decision-making processes in combating the spread and effects of COVID-19 within Oman.

Acknowledgements: The authors express their gratitude to their colleagues at the Faculty of FCIT at Sohar University for their valuable guidance.

Author contribution: All authors have contributed, read, and agreed to the published version of the manuscript results.

Conflict of interest: The authors declare no conflict of interest.

References

[1]. Abdo, M. S., Shah, K., Wahash, H. A., & Panchal, S. K. (2020). On a comprehensive model of the novel coronavirus (COVID-19) under Mittag-Leffler derivative. Chaos, Solitons & Fractals, 135, 109867. https://doi.org/10.1016/j.chaos.2020.109867

[2]. Becerra, M., Jerez, A., Aballay, B., Garcés, H. O., & Fuentes, A. (2020). Forecasting emergency admissions due to respiratory diseases in high variability scenarios using time series: A case study in Chile. Science of the Total Environment, 706, 134978. https://doi.org/10.1016/j.scitotenv.2019.134978

[3]. Boccaletti, S., Ditto, W., Mindlin, G., & Atangana, A. (2020). Modeling and forecasting of epidemic spreading: The case of Covid-19 and beyond. Chaos, Solitons & Fractals, 135, 109794. https://doi.org/10.1016/j.chaos.2020.109794

[4]. Effenberger, M., Kronbichler, A., Shin, J. I., Mayer, G., Tilg, H., & Perco, P. (2020). Association of the COVID-19 pandemic with internet search volumes: A Google TrendsTM analysis. International Journal of Infectious Diseases, 95, 192-197. https://doi.org/10.1016/j.ijid.2020.03.034

[5]. Elsheikh, A. H., Saba, A. I., Abd Elaziz, M., Lu, S., Shanmugan, S., Muthuramalingam, T., Kumar, R., Mosleh, A. O., Essa, F. A., & Shehabeldeen, T. A. (2020). Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. Process Safety and Environmental Protection, 149, 223-233. https://doi.org/10.1016/j.psep.2020.09.030

[6]. Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 134, 109761. https://doi.org/10.1016/j.chaos.2020.109761

[7]. Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Applied Soft Computing, 93, 106282. https://doi.org/10.1016/j.asoc.2020.106282

[8]. Khamis, F., Al Rashidi, B., Al-Zakwani, I., Al Wahaibi, A. H., & Al Awaidy, S. T. (2020). Epidemiology of COVID-19 infection in Oman: Analysis of the first 1304 cases. Oman Medical Journal, 35 (3), 1-4. https://doi.org/10.5001/omj.2020.60

[9]. Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., & Cao, K. (2020). Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. Radiology, 296 (2), E65-E71. https://doi.org/10.1148/radiol.2020200905

[10]. Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Medicine and Infectious Disease, 37, 101742. https://doi.org/10.1016/j.tmaid.2020.101742

[11]. Oman COVID - Coronavirus Statistics - Worldometer. (n.d.). https://www.worldometers.info/coronavirus/country/oman/. Accessed January 10, 2024

[12]. OmanvsCOVID19. (2024). Online resource. Retrieved January 10, 2024, from https://opendata.om/2021/covid-cases/

[13]. Qu, Z., Li, Y., Jiang, X., & Niu, C. (2023). An innovative ensemble model based on multiple neural networks and a novel heuristic optimization algorithm for COVID-19 forecasting. Expert Systems with Applications, 212, 118746. https://doi.org/10.1016/j.eswa.2022.118746

[14]. Ribeiro, M. H., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 135, 109853. https://doi.org/10.1016/j.chaos.2020.109853

[15]. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., & Chowell, G. (2020). Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5, 256-263. https://doi.org/10.1016/j.idm.2020.03.004

[16]. Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowledge-Based Systems, 205, 106270. https://doi.org/10.1016/j.knosys.2020.106270

[17]. Su, X., Sun, Y., Liu, H., Lang, Q., Zhang, Y., Zhang, J., Wang, C., & Chen, Y. (2023). An innovative ensemble model based on deep learning for predicting COVID-19 infection. Scientific Reports, 13 (1), 12322. https://doi.org/10.1038/s41598-023-38982-5

[18]. Sufian, A., Ghosh, A., Sadiq, A. S., & Smarandache, F. (2020). A survey on deep transfer learning to edge computing for mitigating the COVID-19 pandemic. Journal of Systems Architecture, 108, 101830. https://doi.org/10.1016/j.sysarc.2020.101830

[19]. Sun, T., & Wang, Y. (2020). Modeling COVID-19 epidemic in Heilongjiang province, China. Chaos, Solitons & Fractals, 138, 109949. https://doi.org/10.1016/j.chaos.2020.109949
[20]. Tuncer, T., Dogan, S., & Ozyurt, F. (2020). An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image. Chemometrics and Intelligent Laboratory Systems, 203, 104054. https://doi.org/10.1016/j.chemolab.2020.104054
[21]. Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14 (4), 337-339. https://doi.org/10.1016/j.dsx.2020.04.012
[22]. WHO. (2023). Press conferences on COVID-19 and other global health issues. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[23]. WHO. (2024). Trends in COVID-19 cases, World. https://data.who.int/dashboards/covid19/cases?n=c
[24]. Yousif, J. H., & Saini, D. K. (2020). Big Data Analysis on Smart Tools and Techniques. In Cyber Defense Mechanisms (pp. 111-130). CRC Press. https://doi.org/10.1201/9781003032247
[25]. Yousif, J. H., Khan, F. R., Al Jaradi, S. N., & Alshibli, A. S. (2021). Exploring the influence of social media usage for academic purposes using a partial least squares approach. Computation, 9 (6), 64. https://doi.org/10.3390/computation9060064
[26]. Yüce, M., Filiztekin, E., & Özkaya, K. G. (2021). COVID-19 diagnosis—A review of current methods. Biosensors and Bioelectronics, 172, 112752. https://doi.org/10.1016/j.bios.2020.112752