JURNAL INFOTEL
VOL. 16, NO. 4, NOVEMBER 2024, PP. 819–836. DOI:10.20895/INFOTEL.V16I3.1263

RESEARCH ARTICLE

Application of Ensemble Machine Learning for Infectious Diseases with Vaccine Intervention: A Global COVID-19 Case Study

Egi Safitri[1,*], Ruki Rizalnul Fikri [2], and Rini Nurlistiani[3]

[1,2]Department Data Science, Institute of Informatics and Business Darmajaya, Lampung, Indonesia
[3]Department of Information System, Institute of Informatics and Business Darmajaya, Lampung, Indonesia

*Corresponding email: egisafitri@darmajaya.ac.id

Received: October 29, 2024; Revised: December 08, 2024; Accepted: December 10, 2024.

Abstract: The COVID-19 pandemic has posed significant global challenges, particularly in managing vaccination campaigns and tracking the spread of active cases. Accurate prediction of daily vaccination rates and active COVID-19 cases is essential for effective pandemic control and timely decision-making. However, the complexity of pandemic-related data makes such predictions challenging, requiring advanced machine-learning models. This study utilizes global data from multiple sources to evaluate the performance of several ensemble learning algorithms, including Random Forest, Bagging, Gradient Boosting Machine (GBM), AdaBoost, and XGBoost, in predicting daily vaccination rates and active COVID-19 cases. The results reveal that Random Forest consistently outperforms other models, providing the most accurate predictions for daily vaccinations and active cases. Conversely, AdaBoost demonstrated the least effective performance in both prediction tasks. These findings underscore the importance of ensemble learning techniques in enhancing prediction accuracy. This research contributes valuable insights into the potential of machine learning for improving global pandemic response strategies, supporting policymakers in making data-driven decisions for vaccination rollout and active case monitoring.

Keywords: Covid-19, Ensemble Learning, Infectious Diseases, Vaccine

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has profoundly impacted the global population, manifesting as an acute respiratory syndrome in humans. Unlike classical diseases, this pandemic has had far-reaching effects on society. The rapid and persistent spread, coupled with high morbidity and mortality rates, has generated widespread fear and anxiety [1,2]. These conditions have disrupted the natural flow of daily life, leading to significant disturbances in individuals' psychological well-being. The fear of the disease, along with social restrictions and uncertainty, has had a detrimental impact on the mental health of many people [3,4]. Additionally, the pandemic has heightened global awareness of mortality, further exacerbating anxiety and psychological stress. This crisis has underscored human limitations in confronting global health threats and emphasized the urgent need for interventions to alleviate its psychological impact [5,6].

Infectious diseases have been a significant concern in public health throughout human history [7,8]. From the cholera outbreaks of the 19th century to the ongoing global COVID-19 pandemic, these diseases pose threats not only to individuals but also to societal, economic, and public health systems on a broad scale [9,10]. Our understanding of infectious diseases has significantly evolved with advancements in science and technology. However, controlling their spread remains complex, particularly in an era of globalization, which facilitates the rapid dissemination of infectious agents worldwide [11,12]. Infectious diseases are caused by various microorganisms such as bacteria, viruses, fungi, or parasites that can be transmitted from one individual to another or from animals to humans. These diseases can lead to local epidemics or even global pandemics [13,14]. Controlling infectious diseases is crucial for maintaining global public health [15]. Control measures include early detection, case isolation, quarantine, behavioral changes, and vaccination. Among these efforts, vaccination is one of the most effective strategies, proven to reduce morbidity and mortality rates associated with infectious diseases. Vaccines stimulate the immune system to recognize and combat disease-causing agents, establishing immunity to prevent transmission and suppress disease spread within communities [16].

Previous research on infectious diseases includes studies by Aqil et al. [17], who forecasted COVID-19 vaccination trends in Indonesia using the Facebook Prophet and ARIMA methods. The Facebook Prophet model outperformed ARIMA in predicting COVID-19 trends in Indonesia, with better RMSE, MAPE, and MAE values, indicating higher prediction accuracy. Using the Prophet model, policymakers can gain more precise insights into future vaccine needs, enabling more informed decisions regarding COVID-19 vaccine distribution. Furthermore, Emami et al. [18] compared four machine learning algorithms in predicting mortality among COVID-19 patients. Data were collected from patients admitted to five hospitals in Tehran, Iran. The results showed that the gradient boosting tree (GBT) model best predicted mortality, with 70% accuracy, 77% sensitivity, and 69% specificity. This study highlights the potential of machine learning in improving health outcomes during the pandemic by enabling timely and accurate patient mortality predictions, thereby facilitating better resource allocation and care strategies.

Zoabi et al. [19] examined COVID-19 vaccine distribution trends in Indonesia using machine-learning approaches: Facebook Prophet and ARIMA. The study found that the Facebook Prophet model was more accurate in predicting vaccine distribution than ARIMA. The predictions generated by this model can serve as a basis for government policy formulation regarding future vaccine needs. Previous related studies include work by

Wang et al. [20], who developed the Neural-SEIR model, a flexible data-driven framework for predicting epidemic diseases, including COVID-19, by considering the complex transmission mechanisms of the virus. Additionally, various other algorithms, such as Random Forest, Gradient Boosting, Adaboost, and XGBoost, have been discussed in studies by Mahmoudian et al. [21]. Similar research was conducted by Pek et al. [22], who employed Naive Bayes, Random Forest, and Decision Tree algorithms.

Although numerous studies have utilized machine learning to model COVID-19, this research distinguishes itself by integrating vaccination interventions into other machine learning algorithms. This approach enhances the assessment of disease control and allows for strategic adjustments in response to the changing dynamics of the virus and vaccine distribution. We believe that this approach will contribute significantly to the understanding of the effectiveness of vaccination in controlling the disease.

This paper is structured as follows: the next section presents the research method, describing the dataset and the ensemble machine learning models used in the study. The following section provides a detailed analysis of the models' experimental results and performance comparisons. Finally, the paper concludes with a discussion of the findings and suggestions for future research.

2 Research Method

2.1 Data Collection

The data used in this study was obtained from two primary sources, namely the World Health Organization and Kaggle (https://www.kaggle.com/datasets/imdevskp/corona-virus-report). This dataset includes daily information on COVID-19 cases from various countries worldwide. The dataset used in this study consists of 49.068 rows and 10 columns that record the development of COVID-19 cases in various countries and regions worldwide. The dataset includes essential information such as geographic data, number of confirmed cases, deaths, recoveries, and ongoing active cases. The Province column records the province or state in each region, with 14.664 entries filled and the rest blank. The Country column records the name of the country or region with all entries fully populated. In contrast, the Lat (Latitude) and Long (Longitude) columns record each location's latitude and longitude coordinates, both containing numeric data of float type with no missing values. The Date column represents the date the data was recorded for each entry, with all data fully populated with an object type. The Confirmed column records the number of confirmed cases of COVID-19, and the Deaths column records the number of deaths due to COVID-19, both of which have fully populated data of type int64. In addition, the Recovered column records the number of patients who have recovered from COVID-19, with all entries fully populated. The Active column records the number of active cases, calculated from confirmed cases minus the number of deaths and recoveries, and all data in this column is also fully populated with int64 type. Finally, the WHO Region column groups countries and regions based on regions defined by the WHO. Table 1 shows the first 5 rows in Covid-19 dataset.

The vaccination dataset used in this study includes 86,512 entries with 15 columns containing information about vaccination programs in various countries. Features include country name (country), ISO code (iso code), record date (date), and cumulative data such as total vaccinations (total vaccinations), people vaccinated, and people fully vaccinated.

Table 1: The first 5 rows in Covid-19 dataset

No	Prov	Country	Lat	Long	Date	C	D	R	Active	WHO Region
0	NaN	Afghan	33.93911	67.709953	22/01/02	0	0	0	0	East. Med.
1	NaN	Albania	41.1533	20.1683	22/1/02	0	0	0	0	Europe
2	NaN	Algeria	28.0339	1.6596	22/1/02	0	0	0	0	Africa
3	NaN	Andorra	42.5063	1.5218	22/1/02	0	0	0	0	Europe
4	NaN	Angola	-11.2027	17.8739	22/1/02	0	0	0	0	Africa

With C: Confirmed, D: Deaths, and R: Recovered.

In addition, there is daily vaccination data, both in raw (daily vaccinations raw) and processed (daily vaccinations) forms.

The dataset also provides per capita metrics, such as total vaccinations per hundred population (total vaccinations per hundred), people vaccinated per hundred, and fully vaccinated per hundred, as well as daily vaccinations per million population (daily vaccinations per million). Information on the type of vaccines, the source of the data (source name) and the corresponding website (source website) were also included. While the data is mostly complete, some fields such as total vaccinations, people vaccinated, and people fully vaccinated are missing, indicating imperfections in the collection of vaccination information.

2.2 Data Preprocessing

In the data preprocessing stage, several steps were taken to prepare the dataset for analysis. First step for Covid-19 dataset, blank values in the Province column were filled with "Unknown" using the fill.na() function to ensure that all data was complete and consistent, thus preventing problems during the analysis process. Next, categorical fields such as Country, WHO Region, and Province are converted into numerical form using label encoding with the LabelEncoder() function. This process is essential so that categorical data can be processed by machine learning algorithms, where a unique number represents each category. The Date column, initially a string, was converted to datetime format using pd.to datetime() to facilitate time-based analysis. New columns Year, Month, and Day were created from this column, representing the year, month, and Day of each data entry, respectively. After that, the original Date column is deleted because it is no longer needed. In addition, the Lat and Long columns, which represent the geographical coordinates, were also deleted as they were not used in further analysis. After all these processes, the dataset was ready for analysis. Table 2 shows the first 5 rows of the preprocessed Covid-19 dataset.

Table 2: The first 5 rows of the preprocessed Covid-19 dataset

Index	Prov	Country	C	D	R	Active	WHO Region	Year	Month	Day
0	72	0	0	0	0	0	2	2020	01	22
1	72	1	0	0	0	0	3	2020	01	22
2	72	2	0	0	0	0	0	2020	01	22
3	72	3	0	0	0	0	3	2020	01	22
4	72	4	0	0	0	0	0	2020	01	22

The preprocessing of vaccination data in this study involved steps to ensure data consistency and quality. First, the two datasets, COVID-19 and vaccination data, were merged by country and date. Before merging, the "Country/Region" and "Date" columns in the COVID-19 dataset were converted to "country" and "date" for uniformity. The date column was then converted to datetime format to be suitable for analysis. After that, the two datasets were merged using "inner join" on the "country" and "date" columns, ensuring that only data that had matches on both attributes were included. To handle missing values, rows with incomplete data were discarded from the merged dataset to ensure data quality and reduce potential bias. This processed dataset was checked again to ensure the merge was successful and ready for use in the analysis.

2.3 Feature Selection

This study's feature selection was based on data obtained from the WHO and Kaggle datasets, focusing on variables relevant to the COVID-19 pandemic and vaccination efforts. The selected features aim to capture essential aspects of the pandemic's progression, severity, and recovery, as well as the factors influencing vaccination distribution.

The independent variables (X) used in the prediction models include both COVID-19-related metrics and vaccination data. Specifically, the COVID-19 variables consist of the number of confirmed cases (X_1), deaths (X_2), and recoveries (X_3), which provide insights into the scale and impact of the pandemic. These variables reflect the spread of the virus and the health system's capacity to respond, making them vital for understanding the trajectory of the outbreak.

For vaccination data, the independent variables (X_4, X_5, X_6, and X_7) encompass the total number of vaccinations performed (X_4), the number of individuals vaccinated (X_5), the number of individuals fully vaccinated (X_6), and the type of vaccine used, which has been converted into a numerical representation using LabelEncoder (X_7). These vaccination-related features are essential for modelling the factors that affect the rate of daily vaccinations.

The dependent variable (Y) in this study is twofold: for the COVID-19 model, Y represents the number of active cases, calculated as the difference between confirmed cases and the sum of deaths and recoveries ($Y = X_1 - X_2 - X_3$), indicating the current burden on the healthcare system. For the vaccination model, Y corresponds to the number of daily vaccinations, reflecting the ongoing progress in vaccination campaigns. Predicting Y for active cases helps to guide pandemic response strategies while predicting Y for daily vaccinations provides insights into the factors driving vaccination uptake and distribution patterns.

2.4 Proposed Method

The stages undertaken in this research describes by Figure 1, starting from data collection to model evaluation. The research starts with a predetermined goal and continues collecting relevant data to achieve that goal. The dataset used can come from various sources that support the research and are relevant to the problem to be solved. Once the data is collected, the next step is Data Preprocessing, where the data is cleaned and transformed to ensure it is in good condition and ready to be used in the model. This process includes handling missing values, normalization, encoding categorical variables, and other steps

needed to improve data quality. After data processing, the Feature Selection stage selects the most relevant features or variables. Feature selection aims to select a subset of all available features that contribute most significantly to the model's performance. The dataset is then divided into two parts: 80% for training data and 20% for testing data. The training data is used to train the machine learning model, while the testing data is used to evaluate the performance of the trained model. The model training process involves learning from patterns found in the data to build an accurate predictive model.

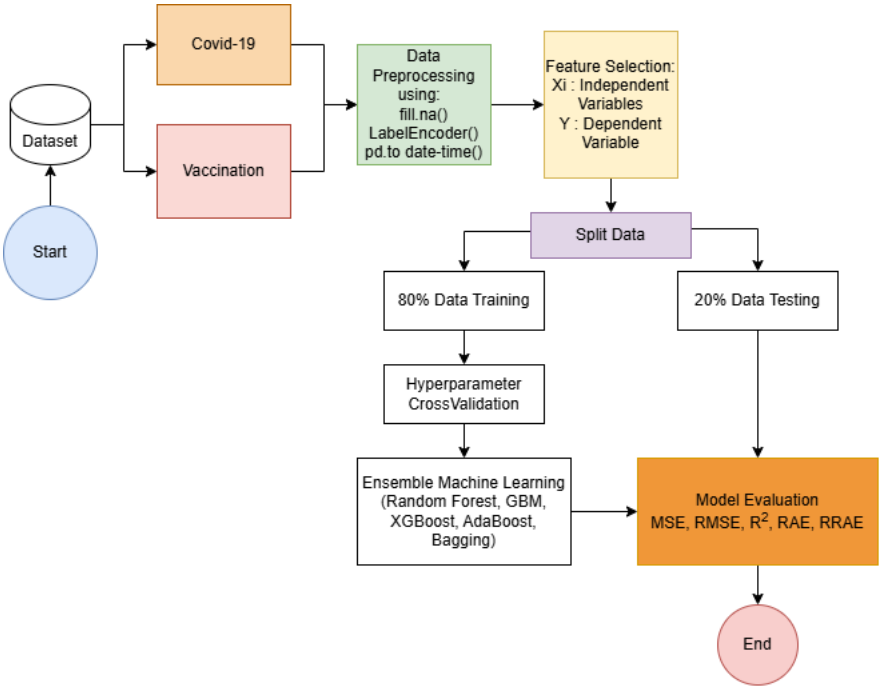

Figure 1: Proposed method for ensemble machine learning.

Once the model is trained, hyperparameter tuning and cross-validation are performed to find the optimal model parameter settings. The cross-validation technique ensures the model does not overfit and has good generalization ability on data that has never been seen before. At this stage, Ensemble Learning is applied to improve model performance. Ensemble learning combines several models to produce a more accurate and robust model. In this research, various ensemble learning methods are used, such as:

1. Random Forest is an ensemble method for classification and regression tasks. It works by building many decision trees during training and outputting a result that is the mode of the class (for classification) or the average prediction (for regression) of all the decision trees [23, 24].
2. Gradient Boosting Machine (GBM) is an ensemble learning method that combines several relatively weak predictive models to form one more robust model. This concept is used for prediction and classification [25, 26].

3. AdaBoost or Adaptive Boosting, is an ensemble method designed to improve the accuracy of classification models by combining multiple weak classifiers to form a more robust and more accurate model [27,28].
4. Bagging (Bootstrap Aggregating) is an ensemble technique used to improve the stability and accuracy of predictive models. This technique involves training multiple independent models on different subsets of the training data and then combining their predicted results to make the final prediction [29,30].
5. XGBoost is a boosting-based ensemble method that uses iteratively trained base learning models to correct the error of the previous model. XGBoost uses a gradient-boosting framework with advanced optimizations, such as regularization, to prevent overfitting and pruning techniques to reduce model complexity [31,32].

Once the model is trained with these techniques, it is tested using Data Testing to evaluate how well it works on data that has never been seen before. The results of these tests are evaluated using various evaluation metrics to measure the model's performance. The research ends once the model is evaluated, and the evaluation results are used to conclude or further implement the research according to the original objectives. This description provides a detailed overview of the research process, emphasizing the use of various ensemble learning techniques to improve model performance.

2.5 Performance

The performance of different prediction models is evaluated based on the following metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE) [33], Root Mean Squared Error (RMSE), R-squared (R^2), Relative Absolute Error (RAE), and Relative Root Absolute Error (RRAE) [34]. For a total of n samples, if y_i and $\hat{y}_i$ represent the actual and predicted values, respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$\text{RAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|} \tag{5}$$

$$\text{RRAE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

3 Results

3.1 Data Visualization

A boxplot was used to visualize the distribution of the number of deaths, confirmed cases, and recovered cases grouped by Impact Level, namely "Mild" and "Severe," to understand further the differences between areas with high and low fatality rates. This visualization aims to illustrate the variations and patterns of data distribution in each category. Figure 2 shows the distribution of the number of deaths, confirmed cases, and recovered cases by pandemic impact level.

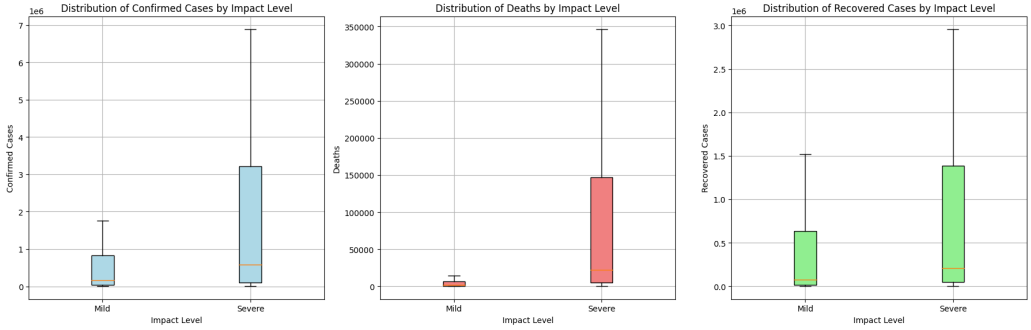

Figure 2: The distribution of confirmed cases, deaths and recovered cases by impact level.

The distribution of deaths shows that severe regions have a much more comprehensive range of deaths than Mild regions. This indicates that regions with higher fatality rates also experience a more significant number of deaths, with very wide variations between them. In contrast, Mild regions tend to have fewer and more consistent deaths. The distribution of confirmed cases shows a similar pattern to the distribution of deaths. Severe regions have a very high and variable number of cases, while Mild regions have fewer cases with a smaller range. It suggests that regions with high fatality rates also face a more significant number of COVID-19 cases. The distribution of the number of recovered cases also shows that Severe regions have a broader variation in recoveries than Mild regions. Some Severe regions recorded a high number of recoveries, but the distribution was highly variable, while Mild regions showed a lower and more uniform number of recoveries.

Figure 3 shows the 10 countries or regions with the highest fatality rates from COVID-19. The fatality rate is calculated as the percentage of deaths out of the total number of confirmed cases in each country. This visualization is important to understand how the pandemic has impacted some countries more severely than others by examining the significant differences in fatality rates.

Figure 3 shows that Yemen has the highest fatality rate, reaching 0.26% of the total confirmed cases. It indicates that Yemen has a significant mortality burden relative to the number of cases, possibly due to limited health infrastructure or other factors that exacerbate the impact of the pandemic in the country. Belgium, the UK, and France each have a fatality rate of 0.15%, which is also high. These Western European countries may face similar challenges in terms of the high number of deaths relative to total cases, even though they have more advanced health systems. Italy (0.14%), one of the first countries in Europe to be

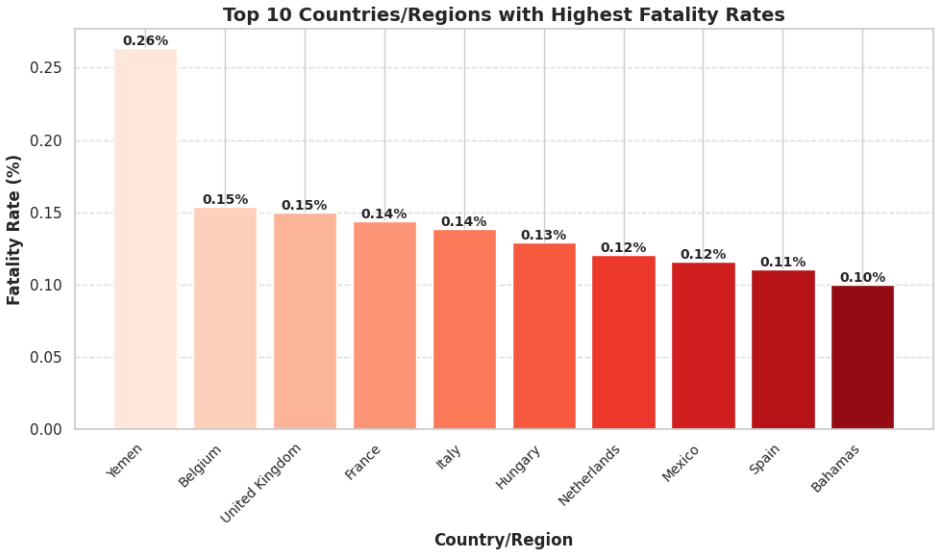

Figure 3: Top 10 country/regions with highest fatality rate.

severely affected by COVID-19, still has a high fatality rate. Hungary and the Netherlands are also ranked next, with fatality rates of 0.13% and 0.12%, respectively. Countries such as Mexico, Spain, and the Bahamas are also on this list, with fatality rates ranging from 0.12% to 0.10%. This shows that the pandemic has a severe impact not only on developed countries but also on other regions with different challenges.

3.2 Covid-19 with Ensemble Learning

This study aims to evaluate the performance of various ensemble learning algorithms in predicting COVID-19 cases by using various metrics to measure the accuracy and effectiveness of the model. Table 3 compares the five algorithms' performance, while Figure 4 visualizes the performance differences between the algorithms more clearly.

Table 3: Metric performance for each algorithm

Algorithm	MAE	MSE	RMSE	R^2	RAE	RRAE
Random Forest	166.0185	17667317.6222	4203.2508	0.9970	0.0121	0.0551
XGBoost	293.8263	6183726.9276	2486.7101	0.9989	0.0214	0.0326
AdaBoost	15306.9045	447684150.1286	21158.5479	0.9230	0.1124	0.2775
Bagging	194.5251	14008659.3193	3742.8144	0.9976	0.0141	0.0491
GBM	898.7925	23835204.8159	4882.1312	0.9959	0.0653	0.0640

Table 3 and Figure 4 show the performance evaluation results of the ensemble learning algorithm in predicting COVID-19 cases and show significant performance differences between algorithms. Based on the metrics used, namely Mean Absolute Error (MAE), Mean

Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), Relative Absolute Error (RAE), and Relative Root Absolute Error (RRAE), it can be seen that the Random Forest and XGBoost algorithms provide the best performance.

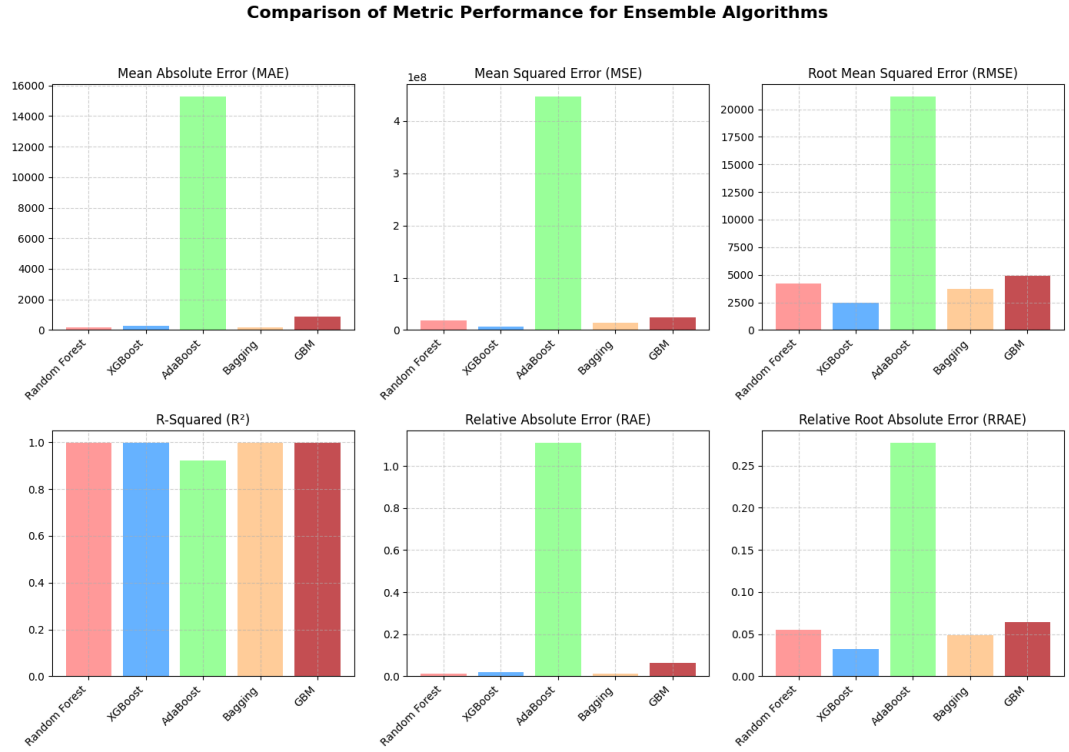

Figure 4: Comparison of metric performance for ensemble algorithms.

The low MAE values in Random Forest and XGBoost indicate that these two algorithms can provide predictions with minimal absolute error. This is supported by the low MSE and RMSE values, which indicate that these two algorithms produce predictions that are closer to the actual values. In contrast, the AdaBoost algorithm shows much higher MAE, MSE, and RMSE values, indicating poor prediction performance. The R^2 values of XGBoost and Random Forest are close to 1, meaning both algorithms are very good at explaining data variability. The Bagging algorithm also shows a reasonably good (R^2) value but is still below Random Forest and XGBoost. Meanwhile, AdaBoost has a lower (R^2) value, indicating that this model is less effective in capturing data variation. On the relative error metrics, namely RAE and RRAE, Random Forest and XGBoost again showed the best results with the smallest values, confirming that these two algorithms are effective in predicting and reliable in minimizing the relative error. On the other hand, AdaBoost again showed the worst performance with the highest RAE and RRAE values.

Overall, Random Forest and XGBoost were the most effective models for COVID-19 data analysis, providing the most accurate and reliable predictions. These models consistently showed low errors and an excellent ability to explain data variability, making them

the top choice for further prediction and modelling related to the COVID-19 pandemic. In contrast, AdaBoost showed inadequate performance, while Bagging and Gradient Boosting offered better alternatives to AdaBoost but were still inferior to Random Forest and XGBoost.

3.3 Vaccination with Ensemble Learning

This stage evaluates the performance of various ensemble learning algorithms in predicting daily vaccination cases using various metrics to measure the accuracy and effectiveness of the model. Table 4 compares the performance of the five algorithms, while Figure 5 visualizes the performance differences between algorithms more clearly.

Table 4: Metric performance for vaccine intervention

Model	MSE	MAE	RMSE	R^2 Score	RAE	RRAE
AdaBoost	5.54E+10	106228.6	235349.8	0.867338	0.427076	0.364228
Random Forest	4.7E+9	16971.07	68557.25	0.988743	0.06823	0.1061
GBM	2.22E+10	59049.98	148897.7	0.9469	0.237402	0.230435
Bagging	4.78E+9	17039.81	69123.19	0.988556	0.068506	0.106975
XGBoost	4E+10	54778.25	200057.7	0.904142	0.220228	0.30961

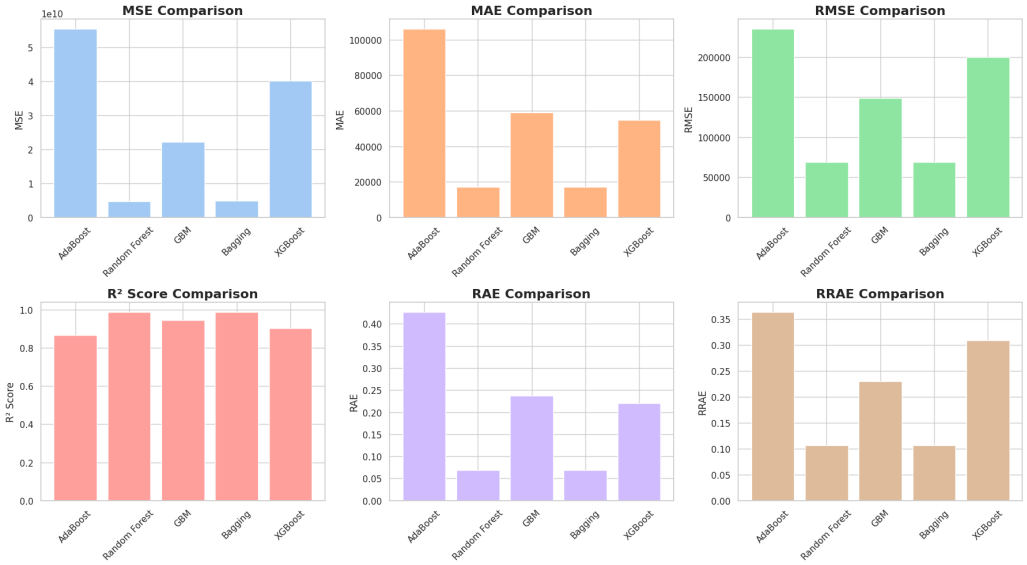

Figure 5: Comparison of metric performance for vaccine intervention.

Table 4 and Figure 5 The performance comparison of several ensemble learning algorithms in predicting the number of daily vaccinations is shown in Table 4 and Figure 5, using evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score, Relative Absolute Error (RAE), and Relative

Root Absolute Error (RRAE). Random Forest and Bagging performed the best, achieving the lowest MSE and RMSE values. The relatively smaller MAE indicates that these models have minimal average daily prediction errors. Additionally, their R^2 Score, which is very close to 1, suggests that they can explain almost all the variability in the data very well. Both models demonstrated high efficiency, as evidenced by their low RAE and RRAE values.

In contrast, AdaBoost performed the worst, with the highest MSE, RMSE, and MAE, indicating much larger prediction errors than the other models. Although its R^2 Score still reflects the ability to explain most of the data variability, the high RAE and RRAE values point to lower prediction efficiency. XGBoost also underperformed relative to Random Forest and Bagging, with relatively high MSE and RMSE values. The MAE suggests a significant average prediction error. While the R^2 Score of XGBoost indicates it can still explain much of the data variability, its RAE and RRAE values show that it is less efficient than the top-performing models.

Overall, based on the evaluation metrics, Random Forest and Bagging proved to be the most reliable and efficient algorithms for predicting daily vaccination rates. These models provided more accurate predictions with lower errors than AdaBoost, GBM, and XGBoost, making them highly recommended for further analysis in the context of daily vaccination prediction.

3.4 Discussion

The COVID-19 pandemic has presented unparalleled challenges globally, making vaccination a critical tool in controlling the virus spread. Our study investigates the patterns of COVID-19 impact in different regions, evaluates ensemble learning algorithms for predicting COVID-19 cases, and assesses their effectiveness in forecasting daily vaccination rates. This discussion interprets the key findings and compares them with existing literature while exploring the implications, limitations, and future research directions.

The results from our analysis show a marked difference in the distribution of COVID-19 cases and fatalities across regions with varying impact levels. Severe regions demonstrated a wider variation in death tolls and case numbers, with significant disparities in outcomes. Conversely, regions categorized as Mild showed more stable and consistent figures in terms of deaths, confirmed cases, and recoveries. These findings align with previous studies indicating that high fatality rates often correlate with limited healthcare infrastructure, delayed interventions, and population vulnerability, particularly in lower-income countries [35]. Furthermore, our evaluation of ensemble learning algorithms highlights the superior performance of Random Forest and XGBoost in predicting both COVID-19 case numbers and daily vaccination rates. Both models consistently outperformed other algorithms, such as AdaBoost and GBM, in terms of prediction accuracy, as evidenced by their low MAE, MSE, RMSE, and high R^2 scores. It supports the growing body of research advocating for the use of ensemble learning techniques, particularly Random Forest and XGBoost, for complex pandemic-related data forecasting [36].

Our findings largely align with prior studies in the field of COVID-19 forecasting and vaccination modeling. For instance, Random Forest and XGBoost have been previously shown to outperform other machine learning models in pandemic forecasting, owing to their ability to handle large, complex datasets and account for non-linear relationships in the data [37]. In contrast, AdaBoost, which underperformed in our study, has been noted in other research to be less effective in scenarios where model stability and accuracy are

paramount. The performance of Bagging in vaccination prediction is particularly noteworthy. Its effectiveness, comparable to Random Forest, further corroborates findings from related studies showing that ensemble methods, especially those based on decision trees, are reliable for predicting vaccine rollouts and their efficacy. However, the relatively poorer performance of XGBoost in this domain suggests that while it excels in predicting COVID-19 case numbers, its ability to forecast vaccination rates may require further refinement, perhaps through the integration of more granular demographic or policy-related variables.

The results of this study do not merely confirm existing models but contribute to refining our understanding of how ensemble learning can be applied to public health forecasting. By demonstrating the effectiveness of Random Forest and XGBoost in pandemic modeling, our findings extend current machine learning theory, particularly in the context of time-series prediction and epidemiological forecasting. These findings also underscore the potential of ensemble methods in real-time crisis management, as seen in the rapid deployment of predictive models during the COVID-19 pandemic. The integration of ensemble learning models into public health decision-making could lead to more accurate predictions of case trajectories and vaccination demand, which can directly inform policy interventions. In this sense, our study contributes to the applied theory of machine learning in public health by testing its efficacy in the context of the COVID-19 crisis and enhancing theoretical frameworks regarding predictive analytics in epidemiology.

Despite the promising results, this study has several limitations. First, the quality and availability of data across different regions were variable, which may have introduced biases in the predictions, particularly for countries with underreporting or inconsistent data collection practices. The reliance on globally aggregated data also masked regional disparities in vaccination campaigns and pandemic responses. Additionally, the use of certain algorithms, such as AdaBoost, which showed suboptimal performance, suggests that hyperparameter tuning and further model refinement could improve prediction accuracy. Another limitation pertains to the scope of our analysis, which was focused on a specific set of ensemble learning algorithms. While these algorithms have demonstrated effectiveness, other machine learning models, such as deep learning architectures, might offer additional insights into COVID-19 and vaccination prediction.

Given the current limitations, future research could explore several avenues to extend and refine our findings. First, future studies could include a more diverse set of machine learning algorithms, including deep learning approaches such as LSTM (Long Short-Term Memory Networks), which have shown promise in handling time-series data with greater predictive accuracy [38, 39]. Integrating more demographic and policy-related variables could also improve model performance, particularly in predicting regional variations in vaccination rates and case numbers. Additionally, expanding the scope of our analysis to include real-time prediction models that adapt to changing epidemiological conditions could provide valuable insights for pandemic response teams. Studies on the societal and behavioral factors influencing vaccination uptake would also be beneficial, integrating these variables into predictive models for more accurate and actionable predictions.

In conclusion, this study highlights the significant role of ensemble learning algorithms in predicting COVID-19 cases and vaccination trends. Random Forest and XGBoost emerged as the most reliable models, offering robust predictive capabilities and minimal error across various metrics. Our findings align with existing literature while also extending the theory of machine learning in public health. Although there are limitations, particularly with data quality and the scope of algorithms tested, our research provides a solid foun-

dation for future studies aiming to enhance pandemic prediction models and vaccination strategies.

We recommend further refinement of the models, the incorporation of more diverse algorithms, and an exploration of real-time adaptive predictive models to improve public health responses in future pandemics.

4 Conclusion

This study demonstrates that ensemble machine learning models, specifically Random Forest and XGBoost, provide the most accurate predictions for COVID-19 case numbers. At the same time, Random Forest and Bagging are particularly effective for predicting daily vaccination rates. These models excel in minimizing prediction errors, as reflected by their low Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), along with R^2 values approaching 1. In contrast, AdaBoost showed poor performance across all evaluation metrics, making it less suitable for these tasks. While Random Forest and Bagging are recommended for vaccination predictions, further optimization, and research are needed for XGBoost and AdaBoost to improve their accuracy, particularly in vaccine forecasting. One limitation of the study is its reliance on a set of ensemble learning models without incorporating deep learning techniques that may capture more complex patterns in pandemic data. Therefore, future studies should consider integrating deep learning models to further improve prediction accuracy, especially for real-time, large-scale data.

Future Work

Future research should explore the application of deep learning techniques, such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), to capture intricate, non-linear relationships in COVID-19 and vaccination data. Additionally, integrating more localized, real-time data and considering socio-political factors could refine model performance across diverse global contexts. Further work should also investigate hybrid models combining ensemble learning and deep learning approaches, as they may offer enhanced predictive capabilities in forecasting both disease dynamics and vaccination trends. This research could significantly contribute to more precise and actionable public health decision-making.

Acknowledgments

The authors express their most profound appreciation to the Directorate General of Higher Education (DRTPM) and the Institute for Research and Community Service (LPPM) of Darmajaya Institute of Informatics and Business for the support and funding through the Beginner Lecturer Research Grant (PDP). This support is beneficial in conducting research that contributes to developing science and publishing practical scientific papers. Our thanks also go to all parties who have contributed to the success of this research.

References

[1] R. Mousa and P. K. Ozili, "Reimagining financial inclusion in the post COVID-19 world: the case of Grameen America," *International Journal of Ethics and Systems*, vol. 39, no. 3, 2023, doi: 10.1108/IJOES-12-2021-0230.

[2] Y. Rosilawati, R. Bustami, M. Fajar, and M. I. Khatami, "The role of university social responsibility during Covid-19 outbreaks: Analysis of Indonesia and Malaysia," *Multidisciplinary Science Journal*, vol. 6, no. 3, 2023, doi: 10.31893/multiscience.2024035.

[3] N. Xiong, K. Fritzsche, Y. Pan, J. Löhlein, and R. Leonhart, "The psychological impact of COVID-19 on Chinese healthcare workers: a systematic review and meta-analysis," *Soc Psychiatry Psychiatr Epidemiol*, vol. 57, no. 8, 2022, doi: 10.1007/s00127-022-02264-4.

[4] N. Salari et al., "Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis," *Global Health*, vol. 16, no. 1, 2020, doi: 10.1186/s12992-020-00589-w.

[5] U. Phalswal, V. Pujari, R. Sethi, and R. Verma, "Impact of social media on mental health of the general population during Covid-19 pandemic: A systematic review," *J Educ Health Promot*, vol. 12, no. 1, 2023, doi: 10.4103/jehp.jehp_460_22.

[6] S. Araki, "S-42-6: COVID-19 MORTALITY CORRELATES WITH CHRONIC DISEASE BURDEN AMONG COUNTRIES' ONGOING SOCIAL DEVELOPMENT," *J Hypertens*, vol. 41, no. Suppl 1, 2023, doi: 10.1097/01.hjh.0000913640.08435.81.

[7] A.-M. Wu et al., "Global, regional, and national burden of neck pain, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021," *Lancet Rheumatol*, vol. 6, no. 3, 2024, doi: 10.1016/S2665-9913(23)00321-1.

[8] C. E. Stauber et al., "Mobile Health Technologies Are Essential for Reimagining the Future of Water, Sanitation, and Hygiene," *Am J Trop Med Hyg*, vol. 106, no. 4, 2022, doi: 10.4269/ajtmh.21-1040.

[9] M. Raynaud et al., "Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production," *BMC Med Res Methodol*, vol. 21, no. 1, 2021, doi: 10.1186/s12874-021-01404-9.

[10] V. Recchia, A. Aloisi, and A. Zizza, "Risk management and communication plans from SARS to COVID-19 and beyond," *Int J Health Plann Manage*, vol. 37, no. 6, 2022, doi: 10.1002/hpm.3545.

[11] D. Zhang et al., "Ecological Barrier Deterioration Driven by Human Activities Poses Fatal Threats to Public Health due to Emerging Infectious Diseases," *Engineering*, vol. 10, 2022, doi: 10.1016/j.eng.2020.11.002.

[12] L. Li et al., "Governing public health emergencies during the coronavirus disease outbreak: Lessons from four Chinese cities in the first wave," *Urban Studies*, vol. 60, no. 9, 2023, doi: 10.1177/00420980211049350.

[13] V. N. E. Malange et al., "The perinatal health challenges of emerging and re-emerging infectious diseases: A narrative review," *Front Public Health*, vol. 10, 2023, doi: 10.3389/fpubh.2022.1039779.

[14] R. A. Karron et al., "Assessment of Clinical and Virological Characteristics of SARS-CoV-2 Infection Among Children Aged 0 to 4 Years and Their Household Members," *JAMA Netw Open*, vol. 5, no. 8, 2022, doi: 10.1001/jamanetworkopen.2022.27348.

[15] N. Zheng et al., "A Novel Linear B-Cell Epitope on the P54 Protein of African Swine Fever Virus Identified Using Monoclonal Antibodies," *Viruses*, vol. 15, no. 4, 2023, doi: 10.3390/v15040867.

[16] J. Liu et al., "CD8 T cells contribute to vaccine protection against SARS-CoV-2 in macaques," *Sci Immunol*, vol. 7, no. 77, 2022, doi: 10.1126/sciimmunol.abq7647.

[17] A. F. Aqil, H.-C. Lee, and S. I. Wardani, "Forecasting COVID-19 Vaccination Trends in Indonesia using Machine Learning," *Indonesian Scholars Scientific Summit Taiwan Proceeding*, vol. 3, 2021, doi: 10.52162/3.2021118.

[18] H. Emami, R. Rabiei, S. Sohrabei, and A. Atashi, "Predicting the mortality of patients with Covid-19: A machine learning approach," *Health Sci Rep*, vol. 6, no. 4, 2023, doi: 10.1002/hsr2.1162.

[19] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *NPJ Digit Med*, vol. 4, no. 1, 2021, doi: 10.1038/s41746-020-00372-6.

[20] H. Wang, X. Qiu, J. Yang, Q. Li, X. Tan, and J. Huang, "Neural-SEIR: A flexible data-driven framework for precise prediction of epidemic disease," *Mathematical Biosciences and Engineering*, vol. 20, no. 9, 2023, doi: 10.3934/mbe.2023749.

[21] A. Mahmoudian, N. Tajik, M. M. Taleshi, M. Shakiba, and M. Yekrangnia, "Ensemble machine learning-based approach with genetic algorithm optimization for predicting bond strength and failure mode in concrete-GFRP mat anchorage interface," *Structures*, vol. 57, 2023, doi: 10.1016/j.istruc.2023.105173.

[22] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, and R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2022.3232984.

[23] A. N. Khan et al., "A New Method for Determination of Optimal Borehole Drilling Location Considering Drilling Cost Minimization and Sustainable Groundwater Management," *ACS Omega*, vol. 8, no. 12, 2023, doi: 10.1021/acsomega.2c06854.

[24] S.-J. Lee et al., "Random RotBoost: An Ensemble Classification Method Based on Rotation Forest and AdaBoost in Random Subsets and Its Application to Clinical Decision Support," *Entropy*, vol. 24, no. 5, 2022, doi: 10.3390/e24050617.

[25] Z. Liang, "Predict Customer Churn based on Machine Learning Algorithms," *Highlights in Business, Economics and Management*, vol. 10, 2023, doi: 10.54097/hbem.v10i.8051.

[26] Y. Baashar et al., "Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/5849995.

[27] X. Li and K. Li, "Imbalanced data classification based on improved EIWAPSO-AdaBoost-C ensemble algorithm," *Applied Intelligence*, vol. 52, no. 6, 2022, doi: 10.1007/s10489-021-02708-5.

[28] S. B. Oyong, U. O. Ekong, and O. U. Obot, "Dynamic analysis of malware intrusion in mobile devices using Adaboost Algorithm, KNN and SVM base classifiers," *World Journal of Applied Science & Technology*, vol. 15, no. 1, 2023, doi: 10.4314/wojast.v15i1.78.

[29] E. Kesriklioğlu, E. Oktay, and A. Karaaslan, "Predicting total household energy expenditures using ensemble learning methods," *Energy*, vol. 276, 2023, doi: 10.1016/j.energy.2023.127581.

[30] K. Kim et al., "Pre-diagnosis of flooding and drying in proton exchange membrane fuel cells by bagging ensemble deep learning models using long short-term memory and convolutional neural networks," *Energy*, vol. 266, 2023, doi: 10.1016/j.energy.2022.126441.

[31] N. Kati and F. Ucar, "An Intelligent Model for Supercapacitors with a Graphene-Based Fractal Electrode to Investigate the Cyclic Voltammetry," *Fractal and Fractional*, vol. 7, no. 3, 2023, doi: 10.3390/fractalfract7030218.

[32] J.-Y. Zhu et al., "Ultrasound-based radiomics analysis for differentiating benign and malignant breast lesions: From static images to CEUS video analysis," *Front Oncol*, vol. 12, 2022, doi: 10.3389/fonc.2022.951973.

[33] S. Mohapatra, R. Mukherjee, A. Roy, A. Sengupta, and A. Puniyani, "Can Ensemble Machine Learning Methods Predict Stock Returns for Indian Banks Using Technical Indicators?," *Journal of Risk and Financial Management*, vol. 15, no. 8, 2022, doi: 10.3390/jrfm15080350.

[34] M. Das and E. Akpinar, "Investigation of Pear Drying Performance by Different Methods and Regression of Convective Heat Transfer Coefficient with Support Vector Machine," *Applied Sciences*, vol. 8, no. 2, 2018, doi: 10.3390/app8020215.

[35] R. M. Barber, N. Fullman, R. J. D. Sorensen, T. Bollyky, M. McKee, E. Nolte, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, A. M. Abdulle, A. A. Abdurahman, S. F. Abera, B. Abraham, G. F. Abreha, K. Adane, A. L. Adelekan, I. M. O. Adetifa, ..., and C. J. L. Murray, "Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990-2015: A novel analysis from the global burden of disease study 2015," *The Lancet*, vol. 390, no. 10091, 2017. doi: 10.1016/S0140-6736(17)30818-8.

[36] F. Özen, "Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey," *Heliyon*, vol. 10, no. 4, 2024, doi: 10.1016/j.heliyon.2024.e25746.

[37] O. Shahid et al., "Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance," *J Biomed Inform*, vol. 117, 2021, doi: 10.1016/j.jbi.2021.103751.

[38] M. U. Tariq and S. B. Ismail, "Deep learning in public health: Comparative predictive models for COVID-19 case forecasting," *PLoS One*, vol. 19, no. 3, 2024, doi: 10.1371/journal.pone.0294289.

[39] M. O. Alassafi, M. Jarrah, and R. Alotaibi, "Time series predicting of COVID-19 based on deep learning," *Neurocomputing*, vol. 468, 2022, doi: 10.1016/j.neucom.2021.10.035.