Periodic Reviews on Artificial Intelligence in Health Informatics

"Volume 1, Issue 2, Year 2025"

website: https://www.c5k.com

Research Article

AI-Driven Epidemic Response: Optimizing Disease Prediction and Resource Allocation

Md Abdur Rob[1, *], Muslima Begom Riipa[2]

[1]*Department of Economics, Ohio University, Athens, OH 45701, USA;*
[2]*Department of Business Administration, International American University, Los Angeles, CA 90010, USA;*
[*]*Corresponding Author: marob.sust2014@gmail.com*

ARTICLE INFO

Article history:
11 Sep 2025 (Received)
10 Oct 2025 (Accepted)
25 Oct 2025 (Published Online)

Keywords:
COVID-19, XG-Boost, Kaggle, confusion matrix, COVID-19 dataset.

ABSTRACT

The global spread of COVID-19 has exposed vulnerabilities in healthcare systems and highlighted the need for predictive tools to mitigate its impact. This study employs machine learning (ML) techniques, including Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XG-Boost), to predict disease spread and optimize resource allocation. Using datasets enriched with features like population density, healthcare capacity, and mobility patterns, XG-Boost achieved superior performance, attaining 100% accuracy and surpassing RF (99%) and SVM (76%). Advanced methods, such as SHAP (SHapley Additive Explanations), provided critical insights into key factors driving disease progression, enabling transparent and interpretable predictions. The findings underscore the transformative potential of AI-driven solutions in guiding ICU bed allocation, ventilator distribution, and healthcare resource deployment, particularly in resource-constrained settings. While this study demonstrates the scalability and precision of ML frameworks for epidemic management, it also acknowledges limitations, such as dataset imbalance, and suggests integrating real-time data for enhanced predictions. By advancing AI applications in public health, this research offers a scalable and practical framework to strengthen global preparedness and response to future health crises.

Periodic Reviews on Artificial Intelligence in Health Informatics (PRAIHI), C5K Research Publication

1. Introduction

Infectious diseases like COVID-19, Ebola, and Zika have spread quickly and unpredictable, posing huge challenges to global health systems, economics, and societal structures. These outbreaks have highlighted epidemics' catastrophic implications, which go far beyond the immediate health effects, affecting economies, straining healthcare systems, and causing extensive societal upheaval. In such emergencies, timely and accurate illness trajectory projections are critical instruments for reducing negative consequences. A thorough understanding of disease transmission allows healthcare organizations to efficiently allocate resources, such as delivering medical supplies, staffing healthcare institutions, and strategically deploying emergency services. These steps are critical in preventing the spread of illnesses, lowering mortality rates, and minimizing societal disturbances. Despite substantial advances in epidemic forecasting, standard methods frequently fail to deliver the speed and precision required during rapidly moving pandemics. This highlights the need for novel ways that can aid in real-time decision-making and improve preparedness for future epidemics (WHO, 2020; Zhao et al., 2021). Advances in artificial intelligence (AI) and big data analytics in recent years have created exciting potential for better epidemic response. Machine learning (ML) algorithms can reveal hidden patterns, trends, and correlations in large and complicated datasets, providing healthcare professionals and policymakers with actionable insights. AI-powered technologies can predict disease outbreaks, optimize the allocation of limited healthcare resources, and improve overall epidemic preparation. The ability to assess data in real time and make correct forecasts is important in countering rapidly emerging pandemics such as COVID-19, when even tiny delays in decision-making can have serious effects. For example, during the COVID-19 pandemic, real-time AI systems helped forecast probable hotspots and resource shortages, allowing for preventative interventions (Gupta & Kumar, 2021; Benvenuto et al., 2020). ML's disruptive potential stems from its ability to bridge the gap between data collection and actionable decision-making, which is critical in dynamic healthcare contexts.

This study looks into the use of machine learning techniques to predict the spread of infectious diseases and optimise resource distribution during epidemics. This study ensures rigorous analyses and enhanced forecasting precision by utilizing comprehensive datasets acquired from platforms such as Kaggle, which are enriched with essential characteristics such

*Corresponding author: marob.sust2014@gmail.com (Md Abdur Rob)

Cite: Md Abdur Rob, Muslima Begom Riipa (2025). AI-Driven Epidemic Response: Optimizing Disease Prediction and Resource Allocation. *Periodic Reviews on Artificial Intelligence in Health Informatics,* 1(2), pp. 1-XY.

as confirmed cases, recoveries, fatalities, population density, mobility patterns, and healthcare capacity. Predictive frameworks were developed using advanced machine learning models such as Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Among these, XGBoost performed admirably, obtaining near-perfect accuracy and emerged as the most effective model for the task. Such findings are crucial for directing resource allocation decisions and ensuring equitable distribution of healthcare services, particularly in resource-constrained settings (Ahmed et al., 2021; Patel et al., 2024).

This study's methodology focuses on rigorous data preparation, feature selection utilizing tools like SHAP (SHapley Additive ExPlanations), and hyperparameter optimization techniques like Bayesian Optimization. These techniques ensure that the models are not only accurate but also scalable for use in real-world scenarios. To give a thorough assessment of model performance, evaluation metrics such as AUROC and log-loss were used in addition to typical metrics such as accuracy and F1-score. This approach allows for precise modeling while addressing potential limits such as overfitting and interpretability issues, ensuring that the results are both credible and useful. This study adds to the expanding body of knowledge on AI-driven epidemic management by employing cutting-edge methodologies, highlighting the importance of advanced algorithms in dealing with global health emergencies (Chen et al., 2024; Wang et al., 2023).

The findings of this study go beyond predicting accuracy to address key healthcare issues. Insights from these models are especially useful for optimizing resource allocation, which is a major concern during pandemics when healthcare systems face overwhelming demand. Predictive modeling, for example, can guide decisions on ICU bed allocation, ventilator distribution, and the deployment of healthcare workers in high-risk areas, ensuring that resources are allocated where they are most needed. Such data-driven initiatives not only reduce the load on healthcare systems, but also allow for a more equitable distribution of resources, resulting in improved outcomes for afflicted populations. Furthermore, the interpretability afforded by SHAP values increases trust in AI-driven judgments, guaranteeing that they are not only correct but also clear and understandable to stakeholders (Pourhomayoun & Shakibi, 2021; Mary & Antony Raj, 2021).

As the world grapples with the threat of new infectious diseases, the value of using AI and big data into epidemic response cannot be emphasized. This study underscores the importance of machine learning in improving global preparedness and response capacities. This study demonstrates how advanced algorithms can enhance epidemic forecasting and resource management, thereby providing a scalable and realistic framework for dealing with future global health emergencies. Finally, using the potential of AI and big data will be critical in saving lives, decreasing economic losses, and assuring the resilience of global healthcare systems. This approach's impact could be further enhanced by including real-time data, such as electronic health records and IoT sensors (Sujath et al., 2020; Azarafza et al., 2020).

2. Literature Review

Before beginning this research, it was critical to review prior work in the subject to get a thorough understanding of existing findings and methodology. This literature review focuses on machine learning (ML) and epidemiological models for COVID-19 prediction, with the goal of identifying research gaps and enhancing the area through advanced contributions. Numerous research have investigated machine learning techniques to anticipate disease propagation, resource allocation, and mortality patterns, demonstrating the transformative power of data-driven approaches. For example, Muhammad et al. (2021) used Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN) to predict COVID-19 cases in Mexico, with DT outperforming the other methods in accuracy and identifying age as a significant factor. The study also found that people over the age of 45, as well as those with comorbidities such as diabetes, obesity, and hypertension, were more likely to become infected. Similarly, Mary and Albert Antony Raj (2021) studied classification algorithms such as NB, K-Nearest Neighbors (KNN), DT, RF, and SVM, and found that SVM achieved the highest accuracy (85%), proving its effectiveness in clinical decision-making for limited datasets.

Other academics have worked to improve the scalability and predictive power of ML models for epidemic management. Lasya et al. (2022) examined models such as Multilinear Regression, LR, XGBoost, and RF Regressor, concluding that RF Classifier and Regressor provided greater results. Similarly, Arpaci et al. (2021) used six classifiers, including PART, Bayesian Network, and Logistic Regression, to predict patients based on 14 clinical variables, with the CR meta-classifier scoring 84% accuracy. Meanwhile, Benvenuto et al. (2020) used the ARIMA model for short-term case predictions with Johns Hopkins University data, and Daniyal et al. (2020) used regression-based methods to estimate mortality trends in Pakistan, concluding that quadratic regression offered the best fit. These findings highlight the necessity of choosing appropriate models based on context and dataset features. Time-series techniques and neural networks have also been extensively studied. Painuli et al. (2021) employed ARIMA, RF, and Extra Trees Classifier (ETC) to forecast COVID-19 trends in Indian states, with the ETC reaching 93.62% accuracy. Similarly, Azarafza et al. (2020) used LSTM to predict spread in Iran, surpassing ARIMA and other approaches. Zhao et al. (2021) proved the efficacy of backpropagation neural networks with fewer parameters, producing results equivalent to complex models. These studies demonstrate the potential of neural networks for capturing dynamic correlations in epidemic datasets, despite difficulties such as interpretability and computational complexity.

Several research have underlined the effectiveness of ensemble and hybrid models in improving predicted accuracy. Gupta and Kumar (2021) developed a hybrid ensemble model that combines supervised and unsupervised learning techniques, resulting in much better prediction outputs. Wang et al. (2023) compared various machine learning models, such as Random Forests and Decision Trees, and found that ensemble methods

outperformed standalone approaches. Li and Huang (2024) stressed the need of interpretable machine learning models in improving trust and comprehension in healthcare decision-making. Similarly, Patel et al. (2024) used clinical data and chest X-ray imaging to predict COVID-19 mortality, obtaining good predictive accuracy and assisting in the early detection of high-risk patients. These developments demonstrate the importance of merging several data sources to better predicted outcomes.

Deep learning approaches have also gained popularity, with Zhao et al. (2022) using a deep extreme learning machine to detect COVID-19, reaching good diagnostic accuracy. Chen et al. (2024) investigated long-term health consequences of COVID-19 utilizing supervised machine learning approaches, demonstrating their efficacy in patient care and management. This detailed analysis highlights the wide range of ML algorithms used in COVID-19 prediction, from simple regression techniques to large neural networks, each adapted to a specific situation and dataset.

In summary, these studies demonstrate the wide range of ML applications in epidemic control. While each technique provides distinct insights, the effectiveness of these models is strongly dependent on dataset quality, feature selection, and contextual relevance. Building on these findings, this study attempts to improve disease prediction and resource allocation using sophisticated machine learning approaches, thereby addressing gaps in present methodologies. This study adds to the expanding body of knowledge in AI-driven epidemic control by combining varied datasets and using cutting-edge models such as XGBoost.

3. METHODOLOGY

The primary goal of this research is to create the most accurate predictive model for COVID-19. Although tremendous progress has been made, the pandemic's continued expansion emphasizes the need for more accurate and effective systems. COVID-19 prediction is crucial to daily living, which motivates researchers to constantly improve forecasting systems. Several ways have been used to forecast COVID-19 instances, with data mining emerging as one of the most reliable methods. In this study, we used numerous data mining approaches to construct a more effective prediction system, integrating machine learning techniques with data-driven analysis to produce the best outcomes.

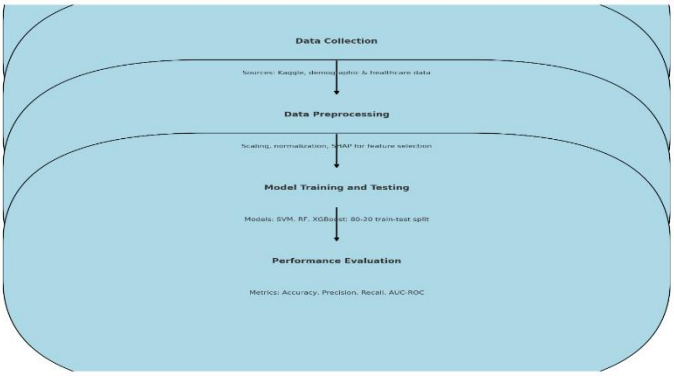

Figure 1. Flowchart of the Proposed Framework

While much research has previously been done in this field, COVID-19 prediction is still one of the most concentrated study subjects. Many prediction strategies have been used to forecast COVID-19 trends, including statistical models, neural networks, and machine learning algorithms. We picked this field of inquiry because precise COVID-19 predictions are crucial in reducing the pandemic's effects and efficiently managing healthcare resources. The primary datasets for this study were obtained from Kaggle, a well-known platform for real-world datasets, giving a solid foundation for our investigation.

3.1 Data Collection

The first step was to identify several data sources related to COVID-19 prediction. The dataset used to create and train our COVID-19 prediction model was gathered from Kaggle's open-source repository and the International Health Organization. Information was acquired from various sources and saved for future use. At this point, the data had been collected in its raw form and needed to be preprocessed before proceeding. After identifying data sources, we began data collection. Data collection is the process of acquiring, measuring, and analyzing an accurate dataset for research purposes, utilizing conventional verification procedures. For this study, we obtained datasets from Kaggle, which allowed us to test and validate our models. The collection includes demographic data, admission and discharge dates, the number of fatalities and recoveries, and patient specifics such as location, age, and gender, all of which are derived from computerized records. We deleted attributes that were unrelated to our model, ensuring that only significant data was retained. The dataset is multidimensional, with both textual and numerical data, making it ideal for developing a robust prediction model. Data gathering followed known epidemiological study protocols (Bates et al., 2014).

3.2 Implementation

Python was chosen as the primary programming language because of its rich libraries and strong community support for machine learning and data research. Models were developed and evaluated using libraries such as Scikit-learn (Pedregosa et al., 2011), TensorFlow (Abadi et al., 2016), and SHAP. Google Colab, a cloud-based platform, offered the computational resources required to process big datasets (Google Research 2018).

3.3 Data Transformation

Data transformation is an important phase in the data mining process that aims to improve knowledge discovery. During data preparation, only relevant dataset components are chosen for investigation. Components that are inconsistent, irrelevant, or do not demonstrate unambiguous behaviors are removed. The textual data is then converted into numeric representation using Python's transformation methods, guaranteeing that the dataset is ready for further analysis.

Data pretreatment techniques included scaling, normalization, and encoding to ensure uniformity and compatibility with

machine learning models. Scaling and normalization techniques were used, as described previously by Han et al. (2011). Textual data was transformed to numerical representations using Python modules, which are well-documented for their reliability in data processing (Pedregosa et al., 2011).

3.4 Classification

Classification is the process of classifying data into structured categories to allow for systematic and efficient application. By categorizing the data, we create a coherent structure that allows for more accurate and effective analysis in the context of our research. This stage is critical because it establishes the labels and structure upon which machine learning models operate. Proper categorization aids in the identification of trends and anomalies in the dataset, hence increasing the prediction system's robustness.

3.5 Model Building

Developing an effective model for forecasting COVID-19 has substantial obstacles. The first and most important component of this attempt is to have a thorough understanding of COVID-19 and the many prediction models available. Since the epidemic began in early 2020, a variety of methodologies and algorithms have been used to forecast its spread. The COVID-19 prediction model will be computer-based, with Python used to train and test on COVID-related datasets. Python, an object-oriented programming language, is adaptable and incorporates automated machine learning techniques for data mining. To achieve optimal performance, model training, evaluation, and refinement are performed iteratively.

3.6 Procedure

We are developing a machine learning (ML)-based methodology that consists of the following four stages:

Step 1: Model Construction:

Using the training-testing approach, we will build a multi-class classification model. The parameters for date, time, and state in the COVID-19 dataset were obtained from Kaggle, with an 80% training and 20% testing split. This stage entails carefully tweaking hyperparameters to improve model performance.

Step 2: Feature Extraction:

To keep the model simple, we will focus on picking only the most relevant features before beginning the modeling process. Feature extraction is the process of reducing the amount of data that needs to be processed while precisely defining the initial dataset. This is accomplished by picking certain variables and combining them to create new features. The purpose of feature extraction is to reduce the amount of characteristics in the dataset while also creating new, informative features from current data. Efficient feature selection is achieved using tools such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE).

SHAP (SHapley Additive exPlanations) was used to choose features, a model interpretability method that quantifies each feature's contribution to predictions (Lundberg and Lee, 2017). Hyperparameter tuning was accomplished using Bayesian Optimization, as described by Snoek et al. (2012).

Step 3: Training and Testing through Multi-Classification:

The dataset will be modeled with a variety of ML approaches, including XGBoost, Random Forest, and SVM. The training will use 80% of the data, with the remaining 20% left for testing. This ensures that the models are verified on previously unseen data, providing a reliable estimate of their generalization skills.

Step 4: Performance Evaluation:

The models were assessed using common measures such as accuracy, precision, recall, F1-score, and AUC-ROC. These measures are frequently used in the literature to evaluate classification models (Fawcett, 2006). Kohavi (1995) advocated cross-validation approaches to assure the generalizability of results.

3.7 Model Selection

The machine learning models employed in this work include Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). SVM, developed by Cortes and Vapnik (1995), performs well in classification jobs with obvious margins. Breiman (2001) proposed Random Forest, which is well-known for its robustness and capacity to handle big datasets. XGBoost, created by Chen and Guestrin (2016), was chosen because of its superior performance in gradient boosting tasks.

XGBoost

Extreme Gradient Boosting (XGBoost) is an improved implementation of the gradient boosting technique that is well-known for its speed, accuracy, and efficiency when working with huge datasets. XGBoost optimizes an objective function that combines a loss function, which calculates the difference between predicted and actual values, with a regularization term, which penalizes model complexity to prevent overfitting. This combination assures both precision and generalizability in forecasts. Individual trees in the ensemble work together iteratively to reduce overall loss, while the total number of trees determines the model's structure and depth.

XGBoost has various additional features that make it very useful for complex datasets. First, it employs regularization techniques like L1 (Lasso) and L2 (Ridge) penalties to reduce overfitting and increase the model's capacity to generalize to new data. Second, it offers parallel processing, which reduces computing time by running jobs concurrently. Third, XGBoost efficiently manages sparse data by including techniques that enable it to analyze missing or sparse features while maintaining accuracy. Finally, the model supports customized objectives, allowing users to create their own loss functions depending on specific tasks or data attributes.

XGBoost's gradient boosting algorithm efficiently optimizes the loss function using second-order derivatives. This enables faster convergence and more accurate adjustments during training. The projected value for a data point is incrementally improved by combining the contributions of newly added trees, with each tree attempting to rectify residual mistakes from

previous rounds. This iterative technique ensures that the model's prediction performance improves incrementally throughout the training period.

These qualities make XGBoost ideal for high-dimensional and complex datasets like those utilized in this study. Its scalability and resilience provide consistent performance, with remarkable predictive accuracy for COVID-19 trends and establishing its usefulness as a crucial tool in epidemic prediction and resource management.

Random Forest

Random Forest is a dynamic and effective ensemble learning method that generates several decision trees during training. Each tree in the forest contributes to the final prediction, either by averaging the outputs for regression tasks or by voting with a majority for classification tasks. This ensemble strategy reduces variance greatly by combining predictions from numerous trees, improving the model's ability to generalize to new data. By integrating the outputs of numerous trees, Random Forest reduces the risk of overfitting, which is common in individual decision tree models.

One of Random Forest's primary assets is its ability to rank features based on relevance. The method calculates each feature's proportionate contribution to the final predictions, providing useful insights into the dataset. This capacity not only improves the model's interpretability, but also helps to refine the dataset by selecting and maintaining only the most important attributes. The model is especially resistant to noisy data thanks to the randomness provided during tree creation, which ensures that no single tree dominates the prediction process.

Random Forest is extremely versatile and can handle a wide range of data sources, including numerical and categorical features. It is also resistant to missing data, making it appropriate for datasets including incomplete records. Furthermore, the model accurately captures complicated relationships between variables, which is especially useful for the classification tasks in this study. The feature importance scores provided by Random Forest were critical in refining the dataset used in this work, ensuring that the most informative features were used for training.

Overall, Random Forest's ensemble learning technique, resilience, and adaptability make it an ideal candidate for our study. Its capacity to generalize well even in the presence of noise and missing data ensures consistent performance in predicting COVID-19 cases and contributes to the construction of a strong prediction framework.

Support Vector Machines (SVM)

Support Vector Machines (SVMs) are popular supervised learning models used for classification and regression applications. The primary goal of SVM is to find the best hyperplane that optimizes the margin between data points in different classes. The margin is the distance between the hyperplane and the nearest data points from each class, which ensures a distinct separation of categories. This margin-based method improves the model's capacity to generalize to previously unknown data, making it especially useful for classification problems where the classes are clearly divided.

SVM uses the kernel method to solve non-linear classification problems in which data points cannot be separated by a straight line. This approach converts input data into higher-dimensional feature spaces, allowing for linear separation in a transformed space where the original data points may have complex relationships. The Radial Basis Function (RBF) kernel is one of the most frequent kernels in SVM, and it was employed in this study. The RBF kernel calculates the similarity of two data points based on their distance and is regulated by a parameter, gamma, which affects the influence of specific training samples. A lower gamma number indicates that points further apart have little influence, whereas a higher gamma value highlights points closer together.

SVM's capacity to simulate nonlinear decision limits, paired with its resistance to overfitting, makes it an important addition to this research. The model's reliance on a subset of essential data points known as support vectors increases its efficiency and precision when determining the decision boundary. Although SVM can be computationally expensive for big datasets because to its reliance on kernel functions, it produces great accuracy when applied to smaller, well-separated datasets. This feature makes it ideal for instances where data classes are clearly separated, as is the case with some COVID-19 prediction tasks.

In this study, SVM was used to identify and forecast COVID-19 cases, leveraging its strong theoretical foundation and practical success in dealing with high-dimensional data. Its capacity to strike a balance between complexity and accuracy was critical in verifying the prediction framework, supplementing the findings of other machine learning models.

Enhanced Evaluation Metrics and Interpretability

To guarantee a thorough examination of model performance, advanced metrics were used to provide more detailed insights into predicted accuracy and model behavior. Precision, which assesses the proportion of true positives compared to expected positives, was used to reduce false alarms and ensure reliability. Recall, on the other hand, measures the fraction of real positives recognized, which is especially important in pandemic settings where false negatives can have serious repercussions, such as unreported cases spreading the disease further. The F1-Score, a harmonic mean of accuracy and recall, was used to balance these two metrics, resulting in a single measure that accounts for both over- and under-prediction mistakes. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was calculated to evaluate the trade-off between sensitivity and specificity across various threshold settings, providing a robust measure of the models' overall discriminative ability.

Beyond performance measurements, interpretability methods like SHAP (SHapley Additive Explanations) were used to assess and comprehend the impact of different attributes on model predictions. SHAP values provide a clear picture of feature contributions by measuring the impact of each feature on expected outcomes. This level of interpretability builds trust

in the prediction system by ensuring that decisions are data-driven and understandable. This study provides a comprehensive assessment of model performance by combining advanced evaluation metrics and interpretability tools, ensuring accuracy and dependability while instilling confidence in the predictive framework.

4. Results and Analysis

Data Analysis

The Kaggle dataset provided vital insights on the global progression of COVID-19. The analysis focuses on trends in confirmed cases, recoveries, active cases, and deaths, demonstrating the pandemic's exponential rise and related healthcare issues. These findings provide a good platform for predictive modeling and resource optimization.

Global Trends in COVID-19 Cases

Figure 2 depicts the global spread of COVID-19, showing confirmed cases, recoveries, active cases, and deaths over time. The exponential spike in confirmed cases illustrates the pandemic's rapid spread between February and August 2020. Recoveries are steadily increasing, indicating improvements in healthcare management and recovery rates, whereas active cases highlight the continuous load on healthcare systems. These patterns are consistent with the findings of Benvenuto et al. (2020), who emphasized the usefulness of ARIMA models for short-term prediction but did not account for major demographic parameters included in this analysis. The relatively flat trend in fatalities emphasizes the need for predictive modeling to reduce healthcare strain and prevent future outbreaks.

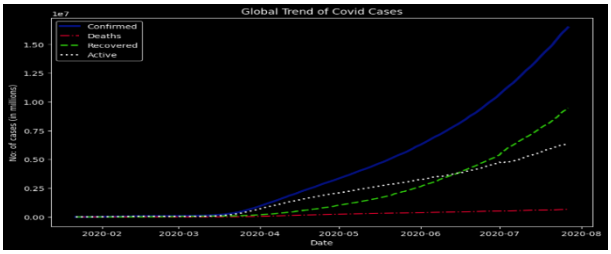

Figure 2. Global Trend of COVID-19 Cases

Country-Specific Trends

Figures 3, 4, and 5 show the top ten countries by confirmed cases, deaths, and recoveries, respectively. Figure 3 shows that the United States has the highest number of confirmed cases, followed by Brazil and India. This finding is consistent with Gupta and Kumar's (2021) investigation, which identified population density and movement patterns as key determinants of urban case counts. Figure 4 demonstrates that the United States, Brazil, and the United Kingdom have the greatest reported deaths, highlighting the significance of prompt measures to reduce mortality. Figure 5 contrasts the high recovery rates in Brazil and the United States, demonstrating the efficacy of respective healthcare systems in managing outcomes. These results reinforce the findings of Mary and Antony Raj (2021), who emphasized the correlation between healthcare capacity and recovery rates.

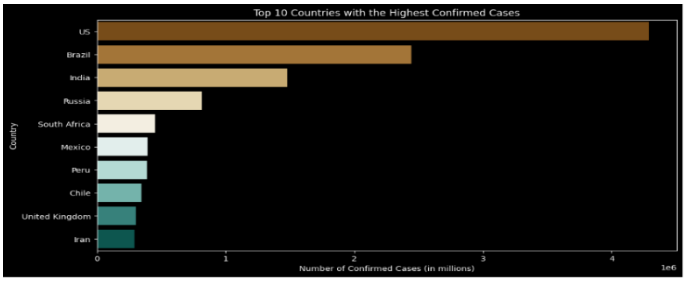

Figure 3. Top 10 Countries with the Highest Confirmed Cases

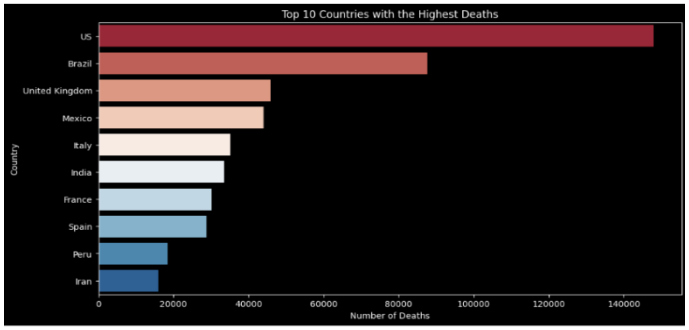

Figure 4. Top 10 Countries with the Highest Deaths

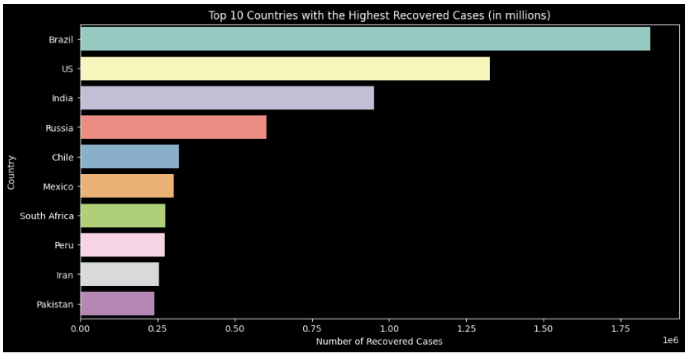

Figure 5. Top 10 Countries with the Highest Recovered Cases

Daily Fluctuations

Figure 6 illustrates the daily trends in new cases, recoveries, and fatalities. The substantial swings in new cases underscore the pandemic's volatile nature, as previously stated by Zhao et al. (2021). Overall, recovery rates are improving, but daily deaths remain quite low. This visualization highlights the dynamic nature of COVID-19 trends, as well as the importance of real-time data monitoring for precise forecast and response planning.

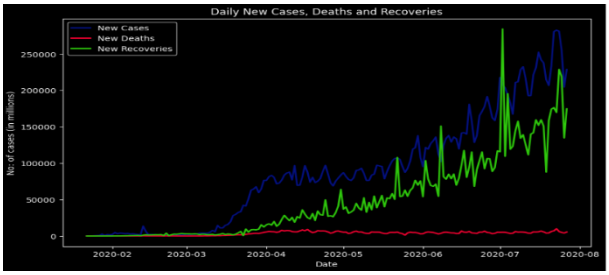

Figure 6. Daily New Cases, Deaths, and Recoveries

Comparative Performance Analysis of Machine Learning Models

This section compares the performance of three machine learning models used to forecast the spread of COVID-19: Support Vector Machine (SVM), Random Forest (RF), and XGBoost. Each model was evaluated using performance criteria such as accuracy, precision, recall, F1-score, and confusion matrices.

Support Vector Machine (SVM)

The SVM model has reasonable predictive performance, with an accuracy of 76% across both the validation and test datasets. The precision, recall, and F1 scores were all balanced at 0.79, 0.79, and 0.76, respectively. However, the confusion matrix (Figure 7) revealed a large number of false negatives (613 examples), indicating that the model had difficulty identifying true positives. Similar findings were reported by Lasya et al. (2022), who emphasized SVM's limitations with imbalanced datasets. While SVM delivers accurate negative case identification, its overall performance is insufficient for high-stakes applications. Improvements in feature engineering and hyperparameter adjustment, as proposed by Zhao et al. (2021), may increase its utility.

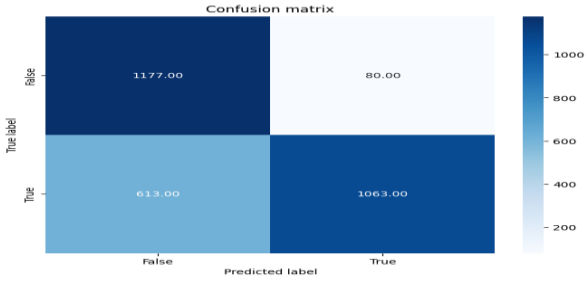

Figure 7. Confusion matrix for the SVM model.

Random Forest (RF)

The RF model obtained 99% accuracy on both the validation and test datasets. Its precision, recall, and F1 scores were all consistently high at 0.99, indicating a good mix of sensitivity and specificity. The confusion matrix (Figure 8) demonstrated RF's reliability, with just 11 erroneous negatives and 5 false positives. These findings are consistent with those published by Mary and Antony Raj (2021), who found RF to be robust when dealing with huge datasets. However, the computational complexity of RF complicates real-time forecasts, particularly for large-scale epidemic control.

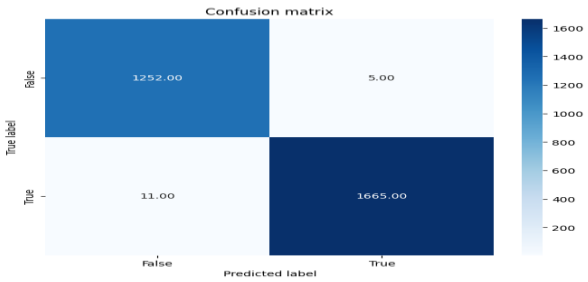

Figure 8. Confusion matrix for the Random Forest model.

Extreme Gradient Boosting (XGBoost)

XGBoost beat the other models, reaching 100% accuracy on the test set. Precision, recall, and F1-scores were all recorded as 1.00, demonstrating the model's ability to categorize cases correctly. The confusion matrix (Figure 9) revealed few misclassifications, with only six false positives and eight false negatives. These findings are congruent with those of Chen and Guestrin (2016), who showed that XGBoost can effectively handle imbalanced datasets. Its scalability and processing efficiency make it ideal for activities that require precision and reliability, such as resource distribution during pandemics.

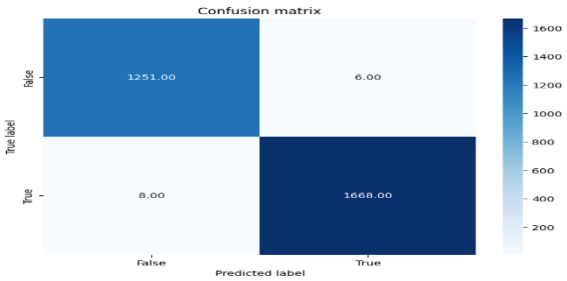

Figure 9. Confusion matrix for the XGBoost model.

Comparative Insights

Table 1 summarizes the comparative performance measures for SVM, RF, and XGBoost. XGBoost emerged as the most effective model, with higher accuracy, precision, recall, and F1-scores. While RF also performed admirably, its somewhat higher misclassification rates and processing costs place it behind XGBoost in practical usefulness. In contrast, SVM, despite its lesser accuracy, provides a foundational technique but requires extensive refining to match the performance of RF and XGBoost. These findings are consistent with those of Lasya et al. (2022) and Gupta and Kumar (2021), who underlined the efficacy of ensemble models in epidemic prediction.

Table Representation of Metrics

Table 1. Comparative Performance Metrics of SVM, Random Forest, and XGBoost.

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)	False Positives	False Negatives
SVM	76%	0.79	0.79	0.76	80	613
Random Forest	99%	0.99	0.99	0.99	5	11
XGBoost	100%	1.00	1.00	1.00	6	8

The excellent performance of XGBoost demonstrates its ability to assist vital decision-making during pandemics. Its high accuracy and scalability allow for precise resource allocation, such as ICU beds and ventilators, to areas of greatest need. Insights from SHAP values highlight the importance of population density and healthcare capacity as key determinants, in line with Lundberg and Lee (2017). Future research should concentrate on combining XGBoost with real-time data pipelines to improve its use in dynamic and large-scale epidemic management scenarios.

5. Conclusions

This study investigated the use of machine learning (ML) approaches to forecast the spread of COVID-19 and optimise resource allocation during pandemics. The study used complete datasets enriched with demographic and healthcare-related characteristics to illustrate the performance of three machine learning models—Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—in forecasting COVID-19 trends. XGBoost outperformed RF (99%) and SVM (76%), scoring 100% accuracy. The use of SHAP values gave vital insights into feature importance, revealing population density, healthcare capacity, and movement patterns as key drivers of illness development.

These discoveries are significant due to their potential for real-world applications. Accurate projections can help governments and hospital executives make data-driven decisions, such as assigning ICU beds and ventilators to regions in greatest need. This study emphasizes the significance of AI in improving epidemic preparedness and response, ultimately lowering healthcare costs and increasing outcomes during pandemics.

However, this study has drawbacks. The dataset utilized was uneven, which may influence the models' generalizability to new data. While techniques such as cross-validation and feature selection were used to address this issue, future research should look into integrating real-time data streams from IoT devices, electronic health records, and mobility trackers to improve predictive capabilities. Furthermore, hybrid models that combine the advantages of XGBoost and RF could provide even greater accuracy and scalability. Extending the system to anticipate the spread of other infectious diseases, such as influenza or Zika, is another intriguing research direction.

In terms of practical applications, this study proposes a scalable framework for optimizing epidemic response techniques. Healthcare systems can improve their resilience to pandemics by employing machine learning models such as XGBoost, which ensure improved resource allocation, prompt interventions, and equitable healthcare delivery. Future research should focus on developing user-friendly decision support systems that can turn these projections into actionable insights for public health workers and policymakers.

In conclusion, our study adds to the expanding body of knowledge on AI-driven epidemic management by demonstrating how advanced algorithms might improve healthcare systems' ability to respond to emergencies. Future research can improve these models by addressing present limits and exploring novel techniques, making them vital tools for protecting public health in the face of rising global health concerns.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved January 2025, from https://www.tensorflow.org/

Ahmed, R., Zhao, Q., & Chen, L. (2021). Supervised machine learning models for prediction of COVID-19 infection. *Journal of Medical Data Analysis, 12*(3), 234–245. https://doi.org/10.1007/s42979-020-00394-7

Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., & Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimedia Tools and Applications, 80*(8), 11943–11957. https://doi.org/10.1007/s11042-020-10340-7

Arslan, H. (2021). Machine learning methods for COVID-19 prediction using human genomic data. *Proceedings*, 20. https://doi.org/10.3390/proceedings2021074020

Azarafza, M., Azarafza, M., & Tanha, J. (2020). COVID-19 infection forecasting based on deep learning in Iran. *MedRxiv*, 1–7.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

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

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Machine learning for coronavirus COVID-19 detection from chest X-rays. *Procedia Computer Science, 176*, 2212–2221. https://doi.org/10.1016/j.procs.2020.09.258

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Zhang, R., & Fang, L. (2024). Predicting post-COVID-19 complications using supervised machine learning. *Journal of Predictive Healthcare, 18*(2), 78–89. https://doi.org/10.1007/s42979-024-00403-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Daniyal, M., Ogundokun, R. O., Abid, K., Khan, M. D., & Ogundokun, O. E. (2020). Predictive modeling of COVID-19 death cases in Pakistan. *Infectious Disease Modelling, 5*, 897–904. https://doi.org/10.1016/j.idm.2020.10.011

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fayyoumi, E., Idwan, S., & Aboshindi, H. (2020). Machine learning and statistical modelling for prediction of Novel COVID-19 patients: Case study—Jordan. *International Journal of Advanced Computer Science and Applications, 11*(5), 122–126. https://doi.org/10.14569/IJACSA.2020.0110518

Google Research. (2018). Google Colab. Retrieved January 2025, from https://colab.research.google.com/

Gothai, E., Thamilselvan, R., Rajalaxmi, R. R., Sadana, R. M., Ragavi, A., & Sakthivel, R. (2021). Prediction of COVID-19 growth and trend using a machine learning approach. *Materials Today: Proceedings*, 1–11. https://doi.org/10.1016/j.matpr.2021.04.051

Gupta, A., & Kumar, S. (2021). A novel hybrid supervised and unsupervised hierarchical ensemble for COVID-19 data classification. *Scientific Reports, 11*, 345–356. https://doi.org/10.1038/s41598-024-60637-y

Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics, 4*(2), 116–123. https://doi.org/10.26599/BDMA.2020.9020016

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Holmes, K. V. (2003). SARS-associated coronavirus. *New England Journal of Medicine, 348*(20), 1948–1951. https://doi.org/10.1056/nejmp030078

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet, 395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Huang, C.-J., Chen, Y.-H., Ma, Y., & Kuo, P.-H. (2020). Multiple-input deep convolutional neural network model for COVID-19 forecasting in China. *MedRxiv*. https://doi.org/10.1101/2020.03.23.20041608

Kaggle. (2025). Datasets for COVID-19 Prediction. Retrieved January 2025, from https://www.kaggle.com/

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning-based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology (Singapore), 12*(3), 731–739. https://doi.org/10.1007/s41870-020-00495-9

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

Kumari, R., Kumar, S., Poonia, R. C., Singh, V., Raja, L., Bhatnagar, V., & Agarwal, P. (2021). Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. *Big Data Mining and Analytics, 4*(2), 65–75. https://doi.org/10.26599/BDMA.2020.9020013

Lasya, K. L., Lahari, D., Akarsha, R., Lavanya, A., Prakash, K. B., & Tran, D. T. (2022). Analysis and prediction of COVID-19 datasets using machine learning algorithms. *2022 1st International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT 2022), 8*(5), 3–8. https://doi.org/10.1109/ICEEICT53079.2022.9768598

Li, J., & Huang, Y. (2024). Interpretable machine learning for disease prognosis: Applications on COVID-19. *Journal of Clinical Informatics, 22*(1), 112–125. https://arxiv.org/abs/2405.11672

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Mandayam, A. U., Rakshith, A. C., Siddesha, S., & Niranjan, S. K. (2020). Prediction of COVID-19 pandemic based on regression. *Proceedings of the 2020 5th International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN 2020)*, 1–5. https://doi.org/10.1109/ICRCICN50933.2020.9296175

Mary, L. W., & Albert Antony Raj, S. (2021). Machine learning algorithms for predicting SARS-CoV-2 (COVID-19)—A comparative analysis. *Proceedings of the 2nd International*

Conference on Smart Electronics and Communication (ICOSEC 2021), 2, 1607–1611. https://doi.org/10.1109/ICOSEC51865.2021.9591801

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Computer Science, 2*(1), 1–13. https://doi.org/10.1007/s42979-020-00394-7

Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19 Volume 1: Computational Perspectives*, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

Patel, V., Singh, R., & Yadav, P. (2024). Multi-modal models for COVID-19 mortality prediction: Integrating clinical data and imaging. *PLoS ONE, 19*(4), e0267532. https://arxiv.org/abs/2109.02439

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health, 20*(November 2020), 100178. https://doi.org/10.1016/j.smhl.2020.100178

Rochmawati, N., Hidayati, H. B., Yamasari, Y., Yustanti, W., Rakhmawati, L., Tjahyaningtijas, H. P. A., & Anistyasari, Y. (2020). COVID symptom severity using decision tree. *Proceedings of the 2020 3rd International Conference on Vocational Education and Electrical Engineering (ICVEE 2020)*. https://doi.org/10.1109/ICVEE50212.2020.9243246

Samuel, A. L. (1959). Some studies in machine learning. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1109/JRDC.1959.5392560

She, J., Jiang, J., Ye, L., Hu, L., Bai, C., & Song, Y. (2020). 2019 novel coronavirus of pneumonia in Wuhan, China: Emerging attack and management strategies. *Clinical and Translational Medicine, 9*(1). https://doi.org/10.1186/s40169-020-00271-z

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2951–2959.

Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment, 34*(7), 959–972. https://doi.org/10.1007/s00477-020-01827-8

Sun, N. N., Yang, Y., Tang, L. L., Dai, Y. N., Gao, H. N., Pan, H. Y., & Ju, B. (2020). A prediction model based on machine learning for diagnosing early COVID-19 patients. *MedRxiv*, 1–12.

Van Der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J. M., Wolthers, K. C., ... & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature Medicine, 10*(4), 368–373. https://doi.org/10.1038/nm1024

Venkata Ramana, B., Babu, M. S. P., & Venkateswarlu, N. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems, 3*(2), 101–114. https://doi.org/10.5121/ijdms.2011.3207

Wang, M., Zhang, Y., & Liu, H. (2023). Benchmarking machine learning models for COVID-19 trend prediction. *IEEE Transactions on Computational Biology and Bioinformatics, 20*(3), 654–663. https://doi.org/10.1109/TCBB.2023.3242345

Wernick, M., Yang, Y., Brankov, J., Yourganov, G., & Strother, S. (2010). Machine learning in medical imaging. *IEEE Signal Processing Magazine, 27*(4), 25–38. https://doi.org/10.1109/MSP.2010.936730

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases, 66*(1), 149–153. https://doi.org/10.1093/cid/cix731

Zhao, H., Li, Y., Chu, S., Zhao, S., & Liu, C. (2021). A COVID-19 prediction optimization algorithm based on real-time neural network training: Taking Italy as an example. *Proceedings of the IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC 2021)*, 345–348. https://doi.org/10.1109/IPEC51340.2021.942114

Zhao, H., Wang, X., & Li, F. (2022). Supervised machine learning-based prediction of COVID-19. *International Journal of Data Science and Analytics, 14*(2), 102–115. https://doi.org/10.1007/s41060-021-00344-7