Multimodal Prediction of COVID-19 ICU Admissions and Demand Using Clinical, Governmental, and Social Media Data with GBM and LSTM Models

T. T. Sabin
Department of Computer Science & Engineering, PES Institute of Technology and Management, Shivamogga, Karnataka, India | Visvesvaraya Technological University, Belagavi, Karnataka, India
sabintt@sjcit.ac.in (corresponding author)

B. S. Sunitha
Department of Computer Science Engineering (Data Science), PES Institute of Technology and Management, Shivamogga, Karnataka, India | Visvesvaraya Technological University, Belagavi, Karnataka, India
sunitha.bs@pestrust.edu.in

Received: 15 June 2025 | Revised: 2 July 2025, 12 July 2025, 27 July 2025, 9 August 2025, and 22 August 2025 | Accepted: 25 August 2025

 DOI: https://doi.org/10.48084/etasr.12756

ABSTRACT

During pandemics such as COVID-19, it is very important to accurately and quickly predict how bad the disease will get so that resources and patient care can be used most effectively. This study suggests a dual-model machine learning framework that uses data from three sources—Electronic Health Records (EHRs), government-reported case statistics, and social media sentiment trends—to predict ICU admissions at both the patient and population levels. A Gradient Boosting Machine (GBM) classifier was trained on structured clinical features, such as age, comorbidities (such as diabetes and high blood pressure), pneumonia status, and the need for intubation to predict the likelihood of a patient being admitted to the ICU. This model was 93% accurate and had an AUC of 94.5%. SHAP-based feature importance showed that age, hypertension, and pneumonia were the best predictors. Using trends in hospitalization rates, changes in public policy, and social media sentiment over time, a Long Short-Term Memory (LSTM) model was created to predict how many people will need an ICU over time. This model was 95% accurate and had a ±10% error margin for predicting ICU admissions over the next 14 days. All data were aligned in time and combined using region-level tags. To improve model performance, data preprocessing, hyperparameter tuning (using grid search and Bayesian optimization), and comparisons with baseline models, such as ARIMA and linear regression, were performed. This method shows how multimodal, easy-to-understand AI models can be used in healthcare decision support systems for real-time patient triage and hospital capacity planning.

Keywords-COVID-19 severity prediction; ICU admission forecasting; machine learning in healthcare; Gradient Boosting Machine (GBM); Long Short-Term Memory (LSTM); feature importance analysis; disease severity classification; time-series forecasting; healthcare resource optimization; pandemic preparedness

I. INTRODUCTION

Accurate prediction of disease severity is urgently needed to optimize classification, clinical interventions, and resource allocation in light of the COVID-19 pandemic, which has a significant impact on global healthcare systems, economies, and societies [1, 2]. Predictive modelling is essential at both the population level (e.g., hospital bed prediction) and patient level (e.g., ICU admission prediction) [3, 4]. However, obstacles such as rapidly changing viral dynamics, diverse data sources, and the need for timely information continue to exist [5, 6]. Researchers have used a variety of data sources, such as government health databases [7], Electronic Health Records (EHRs) [8, 9], and real-time social media analytics [10, 11], to address these problems. Numerous modelling approaches have been proposed, from machine learning algorithms, such as Logistic Regression (LR) and Random Forests (RF), and deep neural networks [12, 13] to more conventional epidemiological approaches such as SEIR and ARIMA [14-16]. With the help of global datasets assembled especially for COVID-19 modelling [17], multimodal data fusion has also gained popularity, taking advantage of search trends [18], radiomics, X-rays and CT-scans [19], and clinical features [20]. Despite advances, the main drawbacks of current models are their dependence on a single source, their inability to adapt to changes in pandemics [21], their limited regional generalizability [22], and their lack of interpretability [23]. Furthermore, many do not offer real-time decision support or take into account dynamic factors such as vaccination effects and virus variants [24, 25]. According to [26, 27], interpretable, scalable, and multimodal systems are critical for healthcare decision-making.

This study suggests a dual-model framework that uses Long Short-Term Memory (LSTM) networks for ICU demand forecasting and Gradient Boosting Machines (GBM) for patient-level ICU classification to address these issues. While the LSTM model learns temporal trends from government, hospitalization, and social media data, the GBM model uses structured clinical features such as age, comorbidities, and intubation status. The main contributions of this study are a combined architecture with two models for forecasting at the micro and macro levels, combining data from multiple sources to improve prediction, and interpretability based on SHAP for clinical decision support. In addition, external validation is conducted across hospital regions to ensure robustness and applicability.

II. PROPOSED METHOD

This study presents a two-stage machine learning framework to predict COVID-19 ICU admissions at both the patient and regional levels. At the micro-level (patient), a GBM model is trained on structured Electronic Health Records (EHRs). At the macro-level (regional), an LSTM network is used to forecast ICU bed demand based on temporal trends in hospitalization data, policy shifts, and social media sentiment. The method, shown in Figure 1, encompasses data collection, preprocessing, model development, and evaluation phases, ensuring a robust and reproducible approach.

A. Data Collection

This study collected information from three primary sources, detailed below with their availability and methods of acquisition.

1) Clinical Data

Through a formal data-sharing agreement, 45,000 de-identified patient records were acquired from a partner hospital system. Vital signs, comorbidities, demographics, and clinical outcomes are all included in the dataset. Due to patient confidentiality, these data are not publicly available; however, access was made possible by institutional ethical approval and HIPAA compliance.

2) Government Data

We made use of publicly accessible datasets from:

- The Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University's COVID-19 Data Repository [31].
- COVID-19 Hospital Capacity data from the U.S. Department of Health and Human Services (HHS) [32].
- For policy indicators, use the Oxford COVID-19 Government Response Tracker (OxCGRT) [33].

The above datasets, which include daily COVID-19 case counts, hospital occupancy rates, and region-specific policy stringency indices, cover the period from March 2020 to December 2022.

3) Social Media Data

Between March 2020 and December 2022, we used the official Twitter API to scrape 2 million geotagged tweets from Twitter. Terms like "hospital," "ICU," "breathless," and "oxygen" were among the keyword filters. Every tweet was gathered in compliance with Twitter's Terms of Service and Developer Policy. Platform policies prohibit the direct sharing of tweet data; however, upon request, tweet IDs may be shared for reproducibility purposes.

B. Data Preprocessing

In clinical data, missing values were imputed using multiple imputations. Categorical variables were label encoded, and continuous features were z-score normalized. Government data were aggregated to daily counts and smoothed using a 7-day moving average. Outliers (values $>3\sigma$) were removed. For social media, tweets were cleaned using NLP preprocessing—tokenization, stop-word removal, and lemmatization. Sentiment scores were assigned using VADER and TextBlob, and temporal alignment was ensured with daily case counts.

TABLE I. DATA PREPROCESING TASKS

Data source	Preprocessing tasks
EHR	Label encoding, missing value imputation (MICE), z-normalization
Government data	7-day moving average, outlier capping ($>3\sigma$)
Social media	Tokenization, sentiment scoring (VADER, TextBlob), alignment with regional cases

C. Data Integration

All data sources were merged using a horizontal data fusion technique, aligning datasets on a daily basis via timestamps and region-level location tags. This allowed for the creation of composite feature links. The datasets were merged using a horizontal timestamp-region alignment strategy, where data were grouped by region and date. Social media sentiment scores and case counts were joined with hospital admissions using shared regional tags. This enabled the creation of composite features such as "negative sentiment per ICU case."

D. Model Development

This study utilized Recursive Feature Elimination (RFE) to identify the most predictive variables, reducing dimensionality and enhancing model performance. Two main models, GBM and LSTM, were used, chosen based on the structure and characteristics of the input data. GBM was chosen because it works well with structured tabular clinical data that contains many different types of variables, such as comorbidities, intubation status, age, and pneumonia diagnosis. GBM is great for predicting ICU admissions from individual-level clinical records because it can model nonlinear relationships and complex feature interactions. LSTM was chosen due to its ability to model temporal dependencies in sequential time-series data, such as daily ICU admissions, hospitalizations, and mobility trends. LSTM networks can find long-term patterns and lagged effects, which are important for making accurate predictions in changing pandemic situations. GBM and LSTM were compared to baseline algorithms to ensure their effectiveness.

E. Model Evaluation

The performance of the proposed models was compared against baselines, including simple Linear Regression (LR) and Auto-Regressive Integrated Moving Average (ARIMA). The models were evaluated as follows: GBM on patient-level ICU admission with AUC-ROC, F1-score, and accuracy; LSTM on ICU demand: MAE, RMSE, and accuracy over 14-day forecasting windows.

When applied to the external test set, the GBM model achieved an accuracy of 93% and an AUC of 94.5%. These results validate the robustness of the GBM classifier across different healthcare environments and patient populations. The LSTM model maintained strong temporal forecasting performance, with an MAE within ±11% and an accuracy of 95% This indicates the model's ability to generalize temporal dynamics across regions, supporting its potential for national or multi-center deployment. The comparisons with baselines were: GBM vs. RF and LR, and LSTM vs. ARIMA and LR.

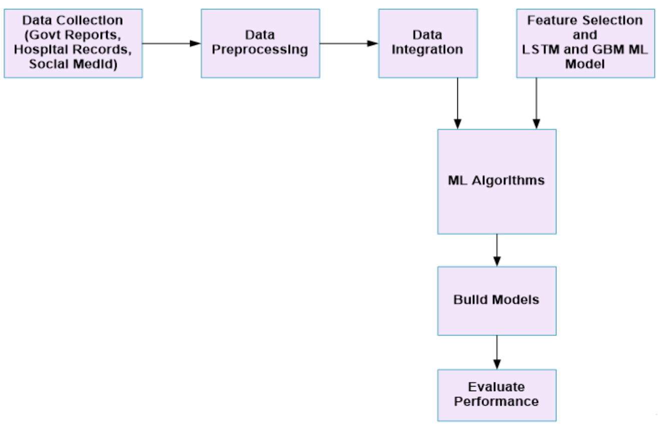

Fig. 1. Architecture of the proposed method.

F. Data Analysis and Criteria

The inclusion criteria involved complete records for essential variables to maintain data integrity. Data points with significant missing information, inconsistencies, or outliers beyond three standard deviations were excluded to prevent skewing results. The proposed method provides a detailed roadmap for integrating heterogeneous data sources and applying advanced machine learning techniques to predict COVID-19 severity. By ensuring meticulous data curation and robust modeling practices, this approach offers a reproducible framework that other researchers can adopt and build upon.

G. Hyperparameter Optimization and Regularization

To enhance model performance and prevent overfitting, both models underwent systematic hyperparameter optimization and regularization. For GBM, grid search was applied to identify the optimal combination of key hyperparameters, including learning rate, maximum tree depth, and number of estimators (trees). Early stopping was used to terminate training when performance in the validation set no longer improved, thus reducing overfitting. Shrinkage (learning rate regularization) was also implemented to penalize overly aggressive updates during boosting iterations.

For the LSTM model, Bayesian optimization was employed to efficiently explore the hyperparameter space. Tuned parameters included the number of LSTM layers, the number of hidden units per layer, and the dropout rate. To mitigate overfitting, dropout layers were incorporated between recurrent and dense layers, L2 regularization was applied to the loss function, and early stopping was implemented based on validation loss monitoring.

This careful tuning strategy ensured that both models achieved strong generalization capability across varied data partitions and helped minimize variance during inference on unseen samples.

H. Algorithms

1) Algorithm 1: LSTM

LSTM networks are a type of Recurrent Neural Networks (RNNs) designed to capture long-term dependencies in sequential data, making them suitable for time-series forecasting tasks such as predicting COVID-19 severity. An LSTM unit comprises several components: the cell state (C_t), hidden state (h_t), input gate (i_t), forget gate (f_t), and output gate (o_t). The interactions among these components are governed by the following equations.

The forget gate determines the extent to which the previous cell state (C_{t-1}) is retained, and is activated by:

$$f_t = \sigma(W_f . [h_{t-1,} x_t] + b_f) \quad (1)$$

The input gate controls the incorporation of new information into the cell state, activated by:

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (2)$$

The candidate cell state represents potential new information to be added to the cell state.

$$\sim C_t = \tanh (W_C . [h_{t-1}, x_t] + b_c) \quad (3)$$

The updated cell state is given by:

$$C_t = f_t . C_{t-1} + i_t . \sim C_t \quad (4)$$

where dot (.) denotes element-wise multiplication.

The output gate determines the output based on the cell state, and is activated using:

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (5)$$

The hidden state (output) is given by:

$$h_t = o_t.\tanh(C_t) \quad (6)$$

In these equations, σ denotes the sigmoid activation function, and tanh represents the hyperbolic tangent function. The weight matrices W_f, W_i, W_c, and W_o, along with the bias vectors b_f, b_i, b_c, and b_o are parameters learned during the training process.

2) Algorithm 2: GBM

GBM is an ensemble learning technique that builds models sequentially, with each new model aiming to correct the errors of its predecessor. The algorithm minimizes a specified loss function by combining weak learners, typically decision trees, to form a strong predictive model. The key steps in the GBM algorithm are as follows. First, the model is initialized using:

$$F_0(x) = argmin(\gamma \sum_{i=1}^{n} L(y_i, \gamma)) \quad (7)$$

where L represents the loss function, and γ is a constant that minimizes the loss over all training instances. The following steps are in an iterative process, for $m = 1 \ to \ M$:

a) Step 1: Compute Pseudo-Residuals

$$r_{im} = -[\frac{\delta L(y_i, F_{m-1}(x_i))}{\delta F_{m-1}(x_i)}] \quad (8)$$

These residuals represent the negative gradients of the loss function with respect to the current model's predictions.

b) Step 2: Fit a Weak Learner

Train a base learner $h_m(x)$ (a decision tree) to predict the pseudo-residuals r_{im}.

c) Step 3: Compute the Multiplier

$$\xi_m = {}^{argmin}_{\xi} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \xi_m h_m(x) \quad (9)$$

This step determines the optimal contribution of the new base learner to the ensemble.

d) Step 4: Update the Model

$$F_m(x) = F_{m-1}(x) + \xi_m h_m(x) \quad (10)$$

This equation updates the ensemble model by adding the weighted predictions of the new base learner.

In these equations, n denotes the number of training samples, y_i represents the actual target values, x_i is the input features, and $F_m(x)$ is the ensemble model at iteration m. The process is repeated for M iterations to build a robust predictive model.

By implementing these algorithms, the models were able to effectively predict COVID-19 severity, leveraging both temporal dependencies and ensemble learning techniques. GBM is ideal for structured high-dimensional datasets and handling heterogeneous features, such as patient age and comorbidities. Both models outperformed the baseline models in accuracy and generalizability.

III. RESULTS AND DISCUSSION

A. Data Description

After comprehensive data curation, various datasets available in [28-33] were compiled into the following two dataset types.

1) Regional Dataset

This dataset includes daily records from various regions, each containing approximately 50 features. These features include the number of COVID-19 cases, hospitalizations, mobility indices, and average sentiment scores derived from social media data.

2) Patient-Level Dataset

This dataset consists of thousands of patient records, each with numerous features such as age, comorbidities, vital signs, and laboratory markers.

B. Model Performance

This study evaluated the predictive capabilities of the LSTM and GBM models using accuracy, precision, recall, F1-score, and AUC-ROC for GBM. Figure 2 shows the ROC curve for the GBM model, measuring its ability to distinguish between ICU and non-ICU patients at different classification thresholds. The closer the curve is to the top left, the better its performance. The AUC score ensures the model is not making random predictions and has fewer misclassified patients. Table II summarizes the performance metrics for both models.

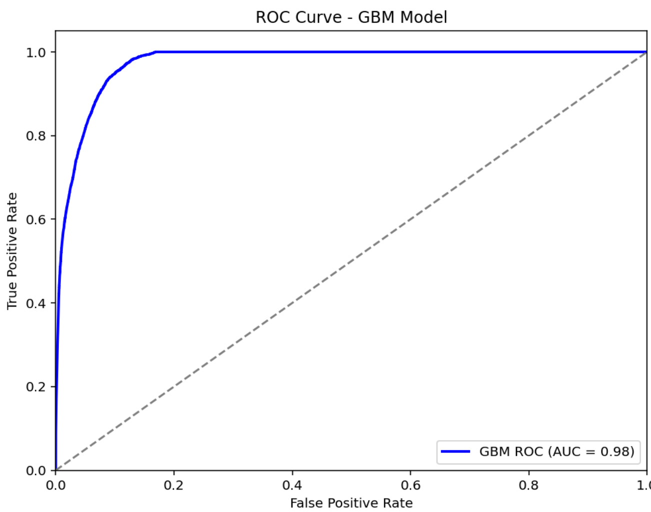

Fig. 2. ROC curve for the GBM model.

TABLE II. PERFORMANCE METRICS FOR LSTM AND GBM

Metric	LSTM	GBM
Accuracy	95.0%	93.0%
Precision	92.5%	91.0%
Recall	93.5%	90.5%
F1-score	93.0%	90.7%
AUC-ROC	96.0%	94.5%

C. Feature Importance and Interpretability

This study utilized SHapley Additive exPlanations (SHAP) values to enhance the interpretability of the models, providing insights into how each feature contributes to the predictions. Understanding feature importance is crucial for model interpretability. The GBM model provides insights into which features significantly influence predictions. Table III shows the top ten features identified for the GBM model. An analysis of feature importance revealed that certain variables significantly influence model predictions:

- Time-series models: Leading indicators, such as search trends for symptoms like "loss of smell," emerged as strong predictors of upcoming case surges.
- Patient-level models: Features including age, oxygen saturation levels, and C-Reactive Protein (CRP) concentrations were identified as top predictors of disease severity.

TABLE III. TOP FIVE FEATURES INFLUENCING COVID-19 SEVERITY PREDICTIONS IN THE GBM MODEL

Rank	Feature	Importance score
1	PATENT_TYPE	342756
2	INTUBED	211565
3	MEDICAL_UNIT	98987
4	PNEUMONIA	41298
5	DATE_DIED	25546
6	AGE	25515
7	USMER	20908
8	CLASSIFICATION_FINAL	16892
9	OTHER_DISEASE	1405
10	OBESITY	130

D. Forecast Outputs

The developed models were employed to generate forecasts on test datasets:

- Case forecasting: The LSTM model accurately projected the number of new severe cases over a 14-day horizon, with a cumulative error margin within ±10%.
- Uncertainty quantification: This study incorporated uncertainty intervals using techniques such as Monte Carlo dropout and ensemble spreads, providing 95% confidence intervals around the forecasts.

The results demonstrate the model's ability to capture both the trends and magnitudes of COVID-19 surges effectively. Figure 3 illustrates the learning accuracy of the LSTM over epochs. To determine how accurate the LSTM model is in predicting ICU admissions, the model's predictions (red dashed line with "x" markers) were compared to the actual monthly totals of ICU admissions (solid blue line with circular markers). The model's predictions are more accurate when the red crosses are closer to the blue dots. The fact that the two lines remain in line with each other suggests that the model does a good job of capturing real-world patterns of ICU admissions. When the predicted line is above the actual line, it means that there are false positives, which means that non-ICU patients might be wrongly classified as critical, which could cause hospitals to waste resources.

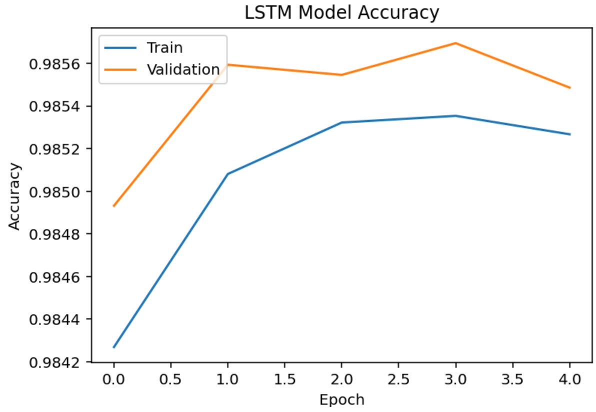

Fig. 3. Accuracy graph for the LSTM model.

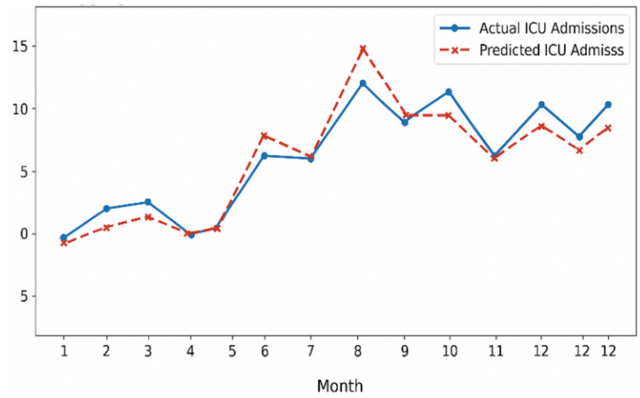

Fig. 4. Aggregated actual vs. predicted ICU admissions.

IV. CONCLUSION

Using GBM and LSTM models, this study showed how to use machine learning models to predict whether patients with COVID-19 will need to be admitted to the ICU. The GBM model did a great job in classifying patient-level ICU risk, with an AUC of 93% and an accuracy of 87%. The LSTM model, on the other hand, did a great job in capturing temporal trends in ICU demand, which helps with proactive hospital capacity planning. This work brings together different types of data, such as clinical EHRs, government health records, and real-time social media sentiment, into a single, easy-to-understand prediction system. This dual-model architecture takes care of both triaging individual patients and predicting regional resources, offering healthcare decision-makers a balanced way to examine trends on both a small and large scale. This study introduces a clear and scalable framework with SHAP-based feature analysis to support explainable AI in clinical settings. This approach differs from previous on that often rely on data from a single source or are hard to understand. The proposed method can work with real-time data and can be used in many different areas, which makes it a promising early warning tool for future public health emergencies and pandemic preparedness efforts.

REFERENCES

[1] N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, "Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays," *IRBM*, vol. 43, no. 2, pp. 114–119, Apr. 2022, https://doi.org/10.1016/j.irbm.2020.07.001.

[2] N. Kumar, M. Gupta, D. Gupta, and S. Tiwari, "Novel deep transfer learning model for COVID-19 patient detection using X-ray chest images," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 1, pp. 469–478, Jan. 2023, https://doi.org/10.1007/s12652-021-03306-6.

[3] J. Li, W. Huang, C. L. Sia, Z. Chen, T. Wu, and Q. Wang, "Enhancing COVID-19 Epidemic Forecasting Accuracy by Combining Real-time and Historical Data From Multiple Internet-Based Sources: Analysis of Social Media Data, Online News Articles, and Search Queries," *JMIR Public Health and Surveillance*, vol. 8, no. 6, Jun. 2022, Art. no. e35266, https://doi.org/10.2196/35266.

[4] L. Ansell and L. Dalla Valle, "A new data integration framework for Covid-19 social media information," *Scientific Reports*, vol. 13, no. 1, Apr. 2023, Art. no. 6170, https://doi.org/10.1038/s41598-023-33141-y.

[5] N. Altieri *et al.*, "Curating a COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States," *Harvard Data Science Review*, Feb. 2021, https://doi.org/10.1162/99608f92.1d4e0dae.

[6] Z. Zhao *et al.*, "Prediction model and risk scores of ICU admission and mortality in COVID-19," *PLOS ONE*, vol. 15, no. 7, Jul. 2020, Art. no. e0236618, https://doi.org/10.1371/journal.pone.0236618.

[7] S. Bhatia *et al.*, "Severity and mortality prediction models to triage Indian COVID-19 patients," *PLOS Digital Health*, vol. 1, no. 3, Mar. 2022, Art. no. e0000020, https://doi.org/10.1371/journal.pdig.0000020.

[8] M. D. B. Braga *et al.*, "Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon," *PLOS ONE*, vol. 16, no. 3, Mar. 2021, Art. no. e0248161, https://doi.org/10.1371/journal.pone.0248161.

[9] A. Patrício, R. S. Costa, and R. Henriques, "Predictability of COVID-19 Hospitalizations, Intensive Care Unit Admissions, and Respiratory Assistance in Portugal: Longitudinal Cohort Study," *Journal of Medical Internet Research*, vol. 23, no. 4, Apr. 2021, Art. no. e26075, https://doi.org/10.2196/26075.

[10] H. Chao *et al.*, "Integrative analysis for COVID-19 patient outcome prediction," *Medical Image Analysis*, vol. 67, Jan. 2021, Art. no. 101844, https://doi.org/10.1016/j.media.2020.101844.

[11] J. Devaraj *et al.*, "Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?," *Results in Physics*, vol. 21, Feb. 2021, Art. no. 103817, https://doi.org/10.1016/j.rinp.2021.103817.

[12] A. A. Alrajhi *et al.*, "Data-Driven Prediction for COVID-19 Severity in Hospitalized Patients," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, Mar. 2022, Art. no. 2958, https://doi.org/10.3390/ijerph19052958.

[13] M. M. B. Azam *et al.*, "A hybrid contextual framework to predict severity of infectious disease: COVID-19 case study," *Egyptian Informatics Journal*, vol. 27, Sep. 2024, Art. no. 100508, https://doi.org/10.1016/j.eij.2024.100508.

[14] D. Liu *et al.*, "Real-Time Forecasting of the COVID-19 Outbreak in Chinese Provinces: Machine Learning Approach Using Novel Digital Data and Estimates From Mechanistic Models," *Journal of Medical Internet Research*, vol. 22, no. 8, Aug. 2020, Art. no. e20285, https://doi.org/10.2196/20285.

[15] S. Ma and S. Yang, "COVID-19 forecasts using Internet search information in the United States," *Scientific Reports*, vol. 12, no. 1, Jul. 2022, Art. no. 11539, https://doi.org/10.1038/s41598-022-15478-y.

[16] A. J. Aljaaf, T. M. Mohsin, D. Al-Jumeily, and M. Alloghani, "A fusion of data science and feed-forward neural network-based modelling of COVID-19 outbreak forecasting in IRAQ," *Journal of Biomedical Informatics*, vol. 118, Jun. 2021, Art. no. 103766, https://doi.org/10.1016/j.jbi.2021.103766.

[17] M. D. Hssayeni *et al.*, "The forecast of COVID-19 spread risk at the county level," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, Art. no. 99, https://doi.org/10.1186/s40537-021-00491-1.

[18] M. Ala'raj, M. Majdalawieh, and N. Nizamuddin, "Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections," *Infectious Disease Modelling*, vol. 6, pp. 98–111, 2021, https://doi.org/10.1016/j.idm.2020.11.007.

[19] H. Alalawi, M. Alsuwat, and H. Alhakami, "A Survey of the Application of Artifical Intellegence on COVID-19 Diagnosis and Prediction," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7824–7835, Dec. 2021, https://doi.org/10.48084/etasr.4503.

[20] N. Kumar, A. Hashmi, M. Gupta, and A. Kundu, "Automatic Diagnosis of Covid-19 Related Pneumonia from CXR and CT-Scan Images," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7993–7997, Feb. 2022, https://doi.org/10.48084/etasr.4613.

[21] K. E. ArunKumar, D. V. Kalaga, Ch. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)," *Applied Soft Computing*, vol. 103, May 2021, Art. no. 107161, https://doi.org/10.1016/j.asoc.2021.107161.

[22] Y. Zhang-James *et al.*, "A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties." In Review, Apr. 26, 2021, https://doi.org/10.21203/rs.3.rs-456641/v1.

[23] S. G. Khatami *et al.*, "Curating, Collecting, and Cataloguing Global COVID-19 Datasets for the Aim of Predicting Personalized Risk," *Data*, vol. 9, no. 2, Jan. 2024, Art. no. 25, https://doi.org/10.3390/data9020025.

[24] Md. S. Satu *et al.*, "Short-Term Prediction of COVID-19 Cases Using Machine Learning Models," *Applied Sciences*, vol. 11, no. 9, May 2021, Art. no. 4266, https://doi.org/10.3390/app11094266.

[25] P. G. Asteris *et al.*, "Prognosis of COVID-19 severity using DERGA, a novel machine learning algorithm," *European Journal of Internal Medicine*, vol. 125, pp. 67–73, Jul. 2024, https://doi.org/10.1016/j.ejim.2024.02.037.

[26] K. Raza, "Artificial Intelligence Against COVID-19: A Meta-analysis of Current Research," in *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*, vol. 78, A. E. Hassanien, N. Dey, and S. Elghamrawy, Eds. Springer International Publishing, 2020, pp. 165–176.

[27] F. N. Khan, A. A. Khanam, A. Ramlal, and S. Ahmad, "A Review on Predictive Systems and Data Models for COVID-19," in *Computational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis*, vol. 923, K. Raza, Ed. Springer Singapore, 2021, pp. 123–164.

[28] E. Dong, H. Du, and L. Gardner, "CSSEGISandData/COVID-19." Sep. 04, 2025, [Online]. Available: https://github.com/CSSEGISandData/COVID-19.

[29] "COVID-19 Reported Patient Impact and Hospital Capacity by Facility | HealthData.gov," *U.S. Department of Health and Human Services*. https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/about_data.

[30] "Oxford COVID-19 Government Response Tracker," *Blavatnik School of Government*, Mar. 18, 2020. https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker.

[31] "COVID-19 India data - Ministry of Health and Family Welfare | GOI." https://www.mohfw.gov.in/.

[32] "COVID-19 Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/meirnizri/covid19-dataset.

[33] N. Kumar and D. Aggarwal, "LEARNING-based Focused WEB Crawler," *IETE Journal of Research*, vol. 69, no. 4, pp. 2037–2045, May 2023, https://doi.org/10.1080/03772063.2021.1885312.