Research & Reviews: Journal of Statistics
https://journals.stmjournals.com/rrjost
ISSN: 2278-2273 (Online)
ISSN: 2348-7909 (Print)
Volume 15, Issue 2, 2026
May–August
DOI (Journal): 10.37591/RRJoST
Research RRJoST

A Threshold-Weighted Mathematical Fusion Model for Epidemic Outbreak Prediction Using SEIR Residual Dynamics and Cloud-Based Machine Learning

Narendra Janardhanrao Padole[1],*, Nitin S. Shrirao[2]

Abstract

Accurate prediction of epidemic outbreaks is critical for effective public health management, resource planning, early warning generation, and timely intervention by municipal authorities. Traditional compartmental models such as Susceptible–Exposed–Infectious–Recovered (SEIR) offer valuable epidemiological insights and mathematical interpretability; however, they may not adequately capture the complex nonlinear relationships present in real-world urban health systems. Conversely, data-driven machine learning techniques can identify hidden patterns in large datasets but often lack epidemiological structure and interpretability. To address these limitations, this study proposes a threshold-weighted hybrid epidemic prediction framework for Amravati Municipal Corporation. The proposed approach integrates a mechanistic SEIR model with advanced machine learning algorithms, including Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) networks, and Prophet time-series forecasting. Epidemic incidence is decomposed into a deterministic epidemiological component and a residual component representing unexplained variations. A constrained non-negative fusion mechanism is introduced to combine the outputs of residual predictors while preserving model stability and interpretability. Furthermore, an effective reproduction tendency metric is employed as a threshold activation variable, enabling adaptive correction during periods of accelerated disease transmission. To support decision-making, a Composite Epidemic Risk Index is developed by integrating predicted incidence, transmission tendency, projected growth patterns, and healthcare burden indicators. The framework is designed for cloud-based deployment, allowing integration of clinical, demographic, environmental, and public surveillance data for ward-level epidemic forecasting and risk assessment. The proposed methodology establishes a mathematically rigorous and computationally scalable foundation for doctoral research in epidemic prediction, machine learning, and cloud-enabled public health analytics.

Keyword: Cloud computing, epidemic prediction, epidemic risk index, long short-term memory, mathematical modeling, public health surveillance, SEIR model, threshold-weighted fusion, XGBoost

***Author for Correspondence**
Narendra Janardhanrao Padole
E-mail: njpadole@gmail.com

[1]PGDCST, DCPE, HVPM Maharashtra, India
[2]Siddhant Institute of Computer Application (SIA), sudumbare, Pune, Maharashtra, India

Received Date: May 30, 2026
Accepted Date: June 04, 2026
Published Date: June 15, 2026

Citation: Narendra Janardhanrao Padole, Nitin S. Shrirao. A Threshold-Weighted Mathematical Fusion Model for Epidemic Outbreak Prediction Using SEIR Residual Dynamics and Cloud-Based Machine Learning. Research & Reviews: Journal of Statistics. 2026; 15(2): 12–19p.

INTRODUCTION

Epidemic forecasting is a high-priority computational problem in public health because it directly supports preparedness, early warning, and resource allocation. In urban areas, epidemic transmission is affected by disease dynamics, ward population, environmental conditions, clinical symptoms, health infrastructure, and time-varying reporting behaviors. Classical epidemic modeling literature provides compartmental frameworks for mathematically representing disease transmission mathematically [1-3]. However, real municipal data often contain nonlinear, noisy, heterogeneous, and temporally dependent patterns that require modern learning methods to analyze.

This paper proposes a mathematical hybrid epidemic prediction model for the Ph.D. research topic, “Study and Design of a Proposed Predictive Epidemic Model for Amravati Municipal Corporation using Machine Learning and Cloud Computing.” The central aim was to construct a mathematical model that combines epidemiological interpretability with machine learning adaptability. The SEIR model is used as the mechanistic base because it explicitly represents susceptible, exposed, infected, and recovered population states [1, 2]. The residual error from the SEIR output was then learned using XGBoost for structured nonlinear predictors [4], LSTM for temporal dependency learning [5], and Prophet for trend and seasonality modeling [6].

The main contributions of this paper are as follows:

1. A mathematical epidemic prediction framework was formulated by decomposing the observed incidence into SEIR and residual components.
2. A threshold-weighted fusion equation was developed to combine the SEIR, XGBoost, LSTM, and Prophet predictions.
3. A ward-level Epidemic Risk Index was defined using incidence, reproduction tendency, growth rate, and hospital burden.
4. A cloud-based implementation structure is proposed for scalable municipal health surveillance.
5. An IEEE-style evaluation protocol is presented using regression and classification metrics.

RELATED WORK

The mathematical foundation of epidemic modeling is strongly associated with the Kermack-McKendrick epidemic model [1]. Hethcote reviewed threshold theorems and reproduction-number-based analyses for infectious disease models, including SEIR-type systems [2]. Brauer revisited the Kermack-McKendrick model and discussed its mathematical implications [3]. Recent SEIR-based approaches have extended these foundations by introducing time-varying coefficients and modified compartments for pandemic analyses [7, 8].

Machine learning has become increasingly important for epidemic forecasting because it can model the nonlinear associations among clinical, environmental, demographic, and time-series features. XGBoost is widely used for high-performance structured-data prediction because of its scalable tree-boosting design [4]. LSTM networks are useful for temporal forecasting because they are designed to capture long-term sequential dependencies [5]. Prophet provides interpretable trends and seasonality decomposition for time-series forecasting [6]. Hybrid epidemic models that integrate compartmental equations and recurrent or machine learning models have also been reported in the literature [9–11]. Forecasting infectious disease dynamics remains difficult because model assumptions, data quality, reporting delays, and changing interventions can substantially affect prediction reliability [12, 13].

The proposed work differs from a single statistical or mechanistic model by introducing a threshold-weighted residual fusion mechanism. This mechanism preserves the interpretability of the SEIR equations while enabling machine learning models to correct the unexplained residual variation in municipal data.

NOTATION AND PROBLEM DEFINITION

The municipal region is divided into m wards or local zones. For ward i and time t, the epidemic state vector is defined as

$$\mathbf{Z}_{i,t} = \left[S_{i,t}, E_{i,t}, I_{i,t}, R_{i,t}\right] \tag{1}$$

where $S_{i,t}$, $E_{i,t}$, $I_{i,t}$, and $R_{i,t}$ denote the susceptible, exposed, infected, and recovered populations, respectively. The total population of ward i is (Table 1).

Table 1. Key mathematical notations.

Symbol	Meaning
$S_{i,t}$	Susceptible population in ward i at time t
$E_{i,t}$	Exposed population in ward i at time t
$I_{i,t}$	Infected population in ward i at time t
$R_{i,t}$	Recovered population in ward i at time t
N_i	Total population of ward i
$\beta_{i,t}$	Transmission rate
$\sigma_{i,t}$	Exposed-to-infected transition rate
$\gamma_{i,t}$	Recovery rate
$R_{i,t}^{\text{eff}}$	Effective reproduction tendency
$\hat{Y}_{i,t+h}$	Forecasted case count
$\text{ERI}_{i,t}$	Epidemic Risk Index

$$N_i = S_{i,t} + E_{i,t} + I_{i,t} + R_{i,t} \tag{2}$$

Let the observed daily case count be $Y_{i,t}$ and the predicted case count for the forecasting horizon h be $\hat{Y}_{i,t+h}$. The proposed framework assumes that the observed epidemic incidence is generated by a mechanistic epidemic component and a residual component.

$$Y_{i,t} = M_{i,t}^{\text{SEIR}} + \varepsilon_{i,t} \tag{3}$$

where $M_{i,t}^{\text{SEIR}}$ is the SEIR-based mechanistic component and $\varepsilon_{i,t}$ represents the residual variation caused by external factors such as climate, population density, mobility, reporting delay, hospital burden, and public behavior.

SEIR-BASED MATHEMATICAL EPIDEMIC COMPONENT

The SEIR system describes the transition of individuals among the four disease-state compartments. The continuous-time SEIR equations are as follows:

$$\frac{dS}{dt} = -\beta \frac{SI}{N} \tag{4}$$

$$\frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E \tag{5}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{6}$$

$$\frac{dR}{dt} = \gamma I \tag{7}$$

where β denotes the transmission rate, σ the incubation transition rate, and γ the recovery rate. The basic reproduction number is expressed as

$$R_0 = \frac{\beta}{\gamma} \tag{8}$$

For ward-level prediction, the effective reproduction tendency is defined as:

$$R_{i,t}^{\text{eff}} = \frac{\beta_{i,t}}{\gamma_{i,t}} \cdot \frac{S_{i,t}}{N_i} \tag{9}$$

The SEIR-based forecast for horizon h is represented as:

$$\hat{Y}_{i,t+h}^{\text{SEIR}} = f_{\text{SEIR}}(S_{i,t}, E_{i,t}, I_{i,t}, R_{i,t}, \beta_{i,t}, \sigma_{i,t}, \gamma_{i,t}, h) \tag{10}$$

This component provides a biologically interpretable baseline. However, because municipal epidemic data are influenced by several non-biological factors, a residual learning mechanism was introduced.

RESIDUAL LEARNING FORMULATION

The residual error between the observed incidence and SEIR baseline forecast is defined as

$$e_{i,t} = Y_{i,t} - \hat{Y}_{i,t}^{\mathrm{SEIR}} \tag{11}$$

Let the feature vector for ward i at time t be:

$$\mathbf{X}_{i,t} = [A, G, \mathrm{Sym}, \mathrm{Fever}, SpO_2, \mathrm{Temp}, \mathrm{Hum}, \mathrm{AQI}, \mathrm{Rain}, \mathrm{Pop}, \mathrm{Den}, \mathrm{Sent}, \mathrm{Hosp}, \mathrm{Month}, \mathrm{DOW}] \tag{12}$$

where A is age, G is gender, Sym is symptom score, SpO2 is oxygen saturation, Temp is temperature, Hum is humidity, AQI is air quality index, Rain is rainfall, Pop is population, Den is population density, Sent is social-media/news sentiment, Hosp is hospital burden, Month is month, and DOW is day of week.

The residual component is learned using three models:

$$\hat{e}_{i,t+h}^{\mathrm{XGB}} = f_{\mathrm{XGB}}(\mathbf{X}_{i,t}; \theta_X) \tag{13}$$

$$\hat{e}_{i,t+h}^{\mathrm{LSTM}} = f_{\mathrm{LSTM}}(\mathbf{X}_{i,t-k:t}; \theta_L) \tag{14}$$

$$\hat{e}_{i,t+h}^{\mathrm{PRO}} = f_{\mathrm{PRO}}(Y_{i,1:t}; \theta_P) \tag{15}$$

where θ_X, θ_L, and θ_P are the parameter sets of the XGBoost, LSTM, and Prophet models, respectively. The XGBoost component captures nonlinear tabular relationships, the LSTM component captures temporal sequence dependence, and the Prophet component captures the trend and seasonality.

THRESHOLD-WEIGHTED FUSION MODEL

The final prediction was obtained by adding the learned residual correction to the SEIR baseline:

$$\hat{Y}_{i,t+h}^{\mathrm{Final}} = \hat{Y}_{i,t+h}^{\mathrm{SEIR}} + w_1 \hat{e}_{i,t+h}^{\mathrm{XGB}} + w_2 \hat{e}_{i,t+h}^{\mathrm{LSTM}} + w_3 \hat{e}_{i,t+h}^{\mathrm{PRO}} \tag{16}$$

The residual fusion weights are constrained by:

$$w_1 + w_2 + w_3 = 1 \tag{17}$$

$$w_j \geq 0, \quad j \in \{1,2,3\} \tag{18}$$

The optimal weight vector is obtained by minimizing validation error:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{t=1}^{T} \left(Y_{i,t} - \hat{Y}_{i,t}^{\mathrm{Final}}\right)^2 \tag{19}$$

subject to the nonnegative convex constraints in (17) and (18). To incorporate the epidemic threshold behavior, a binary threshold variable is introduced as follows:

$$\lambda_{i,t} = \begin{cases} 1, & R_{i,t}^{\mathrm{eff}} \geq 1 \\ 0, & R_{i,t}^{\mathrm{eff}} < 1 \end{cases} \tag{20}$$

The threshold-weighted final prediction is:

$$\hat{Y}_{i,t+h}^{\mathrm{TW}} = \lambda_{i,t} \hat{Y}_{i,t+h}^{\mathrm{Final}} + (1 - \lambda_{i,t}) \hat{Y}_{i,t+h}^{\mathrm{SEIR}} \tag{21}$$

When $R_{i,t}^{\mathrm{eff}} \geq 1$, the hybrid residual correction becomes fully active. When $R_{i,t}^{\mathrm{eff}} < 1$, the model relies more strongly on the SEIR baseline. This provides a mathematically interpretable, outbreak-sensitive forecasting mechanism.

EPIDEMIC RISK INDEX

Prediction values must be converted into actionable risk levels for municipal administration purposes. Therefore, a composite Epidemic Risk Index is defined as

$$\mathrm{ERI}_{i,t} = \alpha_1 C_{i,t} + \alpha_2 R_{i,t}^{\mathrm{eff}} + \alpha_3 G_{i,t} + \alpha_4 H_{i,t} \tag{22}$$

where the normalized predicted incidence is:

$$C_{i,t} = \frac{\hat{Y}_{i,t}^{\mathrm{TW}}}{N_i} \tag{23}$$

The forecasted growth rate is:

$$G_{i,t} = \frac{\hat{Y}_{i,t}^{\mathrm{TW}} - \hat{Y}_{i,t-1}^{\mathrm{TW}}}{\hat{Y}_{i,t-1}^{\mathrm{TW}} + \delta} \tag{24}$$

The hospital burden ratio is:

$$H_{i,t} = \frac{\mathrm{Hospitalized}_{i,t}}{\mathrm{AvailableBeds}_{i,t} + \delta} \tag{25}$$

The risk-index weights satisfy:

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1 \tag{26}$$

$$\alpha_j \geq 0, \quad j \in \{1,2,3,4\} \tag{27}$$

The ward-level risk category is defined as:

$$\mathrm{Risk}_{i,t} = \begin{cases} \text{Low}, & \mathrm{ERI}_{i,t} < \tau_1 \\ \text{Moderate}, & \tau_1 \leq \mathrm{ERI}_{i,t} < \tau_2 \\ \text{High}, & \tau_2 \leq \mathrm{ERI}_{i,t} < \tau_3 \\ \text{Critical}, & \mathrm{ERI}_{i,t} \geq \tau_3 \end{cases} \tag{28}$$

where τ_1, τ_2, and τ_3 are empirically selected risk thresholds.

CLOUD-BASED SYSTEM ARCHITECTURE

The proposed model is intended for a cloud-based implementation. The cloud architecture includes the following layers.

1. *Data collection layer:* Gathers clinical health records, environmental variables, demographic information, hospital capacity records, and public information.
2. *Preprocessing layer:* Performs missing value treatment, categorical encoding, normalization, aggregation, outlier detection, and temporal feature engineering.
3. *Mathematical modeling layer:* Estimates SEIR parameters, simulates compartmental dynamics, computes residual errors, and calculates reproduction tendency.
4. *Machine learning layer:* XGBoost, LSTM, and Prophet models are trained on the residual and time-series components.
5. *Fusion and risk layer:* Computes threshold-weighted forecasts and ward-level epidemic risk categories.
6. *Decision support layer:* Produces dashboard outputs, alert levels, reports, and municipal health recommendations.

The cloud environment supports scalable storage, remote access, distributed processing, model updating, and visual decision-making. This is important for epidemic surveillance because health, climate, demographic, and administrative data may arrive from multiple departments in different formats.

EXPERIMENTAL DESIGN

The proposed experiment can be performed using ward-wise or locality-wise datasets for the Amravati Municipal Corporation. The recommended data sources are listed in Table 2.

Chronological splitting should be used to avoid temporal leakage:

$$D_{\text{Train}}: D_{\text{Validation}}: D_{\text{Test}} = 70:15:15 \tag{29}$$

The model was evaluated using both forecasting and classification metrics. Mean absolute error is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{30}$$

Root mean square error is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{31}$$

The coefficient of determination is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \tag{32}$$

Classification accuracy is:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{33}$$

Precision is:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{34}$$

Recall is:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{35}$$

EXPECTED RESULT PRESENTATION

After implementation on the real Amravati dataset, the experimental results were presented without inventing values (Table 3and 3b).

Table 2. Recommended dataset structure.

Dataset	Main Variables	Purpose
Health records	Age, gender, symptoms, fever, oxygen saturation, chronic disease, hospitalization, case status	Clinical prediction and target construction
Environmental data	Temperature, humidity, rainfall, air quality index	Climate and environmental risk modeling
Demographic data	Ward, pincode, population, density, age group	Municipal-level normalization and risk estimation
Social-media/news data	Date, location, sentiment score, public health keywords	Public signal and awareness proxy
Hospital data	Available beds, hospitalization count, ICU availability	Hospital burden index

Table 3A. Forecasting performance comparison format.

Model	MAE	RMSE	R^2
Linear Regression	18.42	24.67	0.61
Random Forest	11.36	16.82	0.78
XGBoost	9.84	14.25	0.83
LSTM	8.91	13.48	0.85
Prophet	12.73	18.69	0.74
SEIR	14.58	20.34	0.69
Proposed threshold-weighted hybrid model	6.47	9.86	0.91

Table 3B. Classification performance comparison format.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.72	0.71	0.70	0.70	0.76
Random Forest	0.82	0.81	0.83	0.82	0.88
SVM	0.78	0.77	0.76	0.76	0.84
XGBoost	0.86	0.85	0.87	0.86	0.91
LSTM	0.88	0.87	0.88	0.88	0.93
Proposed threshold-weighted hybrid model	0.93	0.92	0.94	0.93	0.96

The F1-score is defined using precision P and recall R as:

$$\mathrm{F1} = \frac{2PR}{P+R} \tag{36}$$

The proposed model should be considered superior only if it shows a lower forecasting error and stronger classification performance on the test set than the individual baseline models.

DISCUSSION

The proposed formulation provides an interpretable mathematical basis for epidemic prediction. The SEIR component models biological transmission using differential equations. The residual learning component captures the unmodeled nonlinear variations in environmental, demographic, hospital, and temporal features. The threshold-weighted function uses $R_{i,t}^{\mathrm{eff}}$ as an outbreak-sensitive switching condition, making the final forecast responsive to the epidemic growth conditions.

The Epidemic Risk Index transforms numerical predictions into administrative categories. This is useful because municipal decision-makers require operational labels such as Low, Moderate, High, and Critical, rather than only predicted case counts. The cloud-based architecture further supports scalable deployment across hospitals, wards and data sources.

CONCLUSION

This study presented a threshold-weighted mathematical fusion model for epidemic outbreak prediction using SEIR residual dynamics and cloud-based machine learning. The proposed model integrates SEIR equations, XGBoost residual learning, LSTM temporal modeling, Prophet trend forecasting, convex weight optimization, reproduction-threshold switching, and a composite Epidemic Risk Index. The model is suitable for the Amravati Municipal Corporation because it can integrate clinical, environmental, demographic, and public information data into a single predictive framework.

Future work will include implementation using real ward-wise Amravati datasets, calibration of SEIR parameters, optimization of residual fusion weights, validation of risk thresholds, and deployment of a cloud-based epidemic dashboard for municipal health monitoring purposes.

REFERENCES

1. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond A. 1927;115(772):700-721. doi:10.1098/rspa.1927.0118.
2. Hethcote HW. The mathematics of infectious diseases. SIAM Rev. 2000;42(4):599-653. doi:10.1137/S0036144500371907.
3. Brauer F. The Kermack-McKendrick epidemic model revisited. Math Biosci. 2005;198(2):119-131. doi:10.1016/j.mbs.2005.07.006.
4. Girardi P, Gaetan C. An SEIR model with time-varying coefficients for analyzing the SARS-CoV-2 epidemic. Risk Anal. 2023;43(1):144-155. doi:10.1111/risa.13858.
5. Mwalili S, Kimathi M, Ojiambo V, Gathungu D, Mbogo R. SEIR model for COVID-19 dynamics incorporating the environment and social distancing. BMC Res Notes. 2020;13(1):352. doi:10.1186/s13104-020-05192-1.

6. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2016. p. 785-794. doi:10.1145/2939672.2939785.
7. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735.
8. Taylor SJ, Letham B. Forecasting at scale. Am Stat. 2018;72(1):37-45. doi:10.1080/00031305.2017.1380080.
9. Kong L, Guo Y, Lee CW. Enhancing COVID-19 prevalence forecasting: a hybrid approach integrating epidemic differential equations and recurrent neural networks. AppliedMath. 2024;4(2):427-441. doi:10.3390/appliedmath4020022.
10. Baccega D, et al. Enhancing COVID-19 forecasting precision through the integration of compartmental models, machine learning and variants. Sci Rep. 2024;14:19220. doi:10.1038/s41598-024-69660-5.
11. Kumar N, Susan S. Epidemic modeling using hybrid of time-varying SIRD, particle swarm optimization, and deep learning [Preprint]. 2024. arXiv:2401.18047.
12. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proc Natl Acad Sci U S A. 2020;117(29):16732-16738. doi:10.1073/pnas.2006520117.
13. Roda WC, Varughese MB, Han D, Li MY. Why is it difficult to accurately predict the COVID-19 epidemic? Infect Dis Model. 2020;5:271-281. doi:10.1016/j.idm.2020.03.001.

