Mathematical Modeling And Artificial Intelligence For Predictive Analysis In Complex Biological And Computational Systems

[1*]Dr. Amit Prakash, [2]Mahendra Singh Bhadauriya, [3]Ranjan Banerjee, [4]Chandan Ghosh, [5]Dr. Chetna, [6]Pooja Bhatt

[1*]Assistant Professor, University Department of Mathematics, Jai Prakash University, Chapra (Saran), Bihar, India, Email Id: a.amitprakash@gmail.com, Orcid Id: 0009-0003-6882-6298
[2]Associate Professor, Department of Mathematics, Institute of Technology and Management, Gwalior-474009, India, Email Id: mahendra.bhadauriya@itmgoi.in, Orcid Id: 0009-0000-6102-4012
[3]Assistant Professor, Department of Computer Science & Engineering, Brainware University, Barasat, Kolkata-125, West Bengal-700125, India, Email Id: rbkpccst@gmail.com, Orcid Id: 0009-0003-1950-7530
[4]Assistant Professor, Department of Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal-700125, India, Email Id: cgiitismkol@gmail.com
[5]Assistant Professor, Department of Mathematics, Multani Mal Modi College, Patiala, Punjab-147001, India, Email Id: drchetna.modicollege@gmail.com, Orcid Id: https://orcid.org/0000-0001-5730-7247
[6]Department of Operations Management & Quantitative Techniques, Indian Institute of management, Indore, India, Email Id: poojab@iimidr.ac.in, Orcid Id: 0009-0002-5248-8135

Abstract

Complex biological and computational systems involve nonlinear interactions, dynamic feedback, uncertainty, and multiscale processes that make accurate prediction difficult using a single modeling approach. Mathematical models provide mechanistic interpretability, whereas artificial intelligence supports flexible prediction from complex data. This research develops a hybrid SEIR–AI framework for predictive analysis using COVID-19 transmission dynamics in India as a representative case study. Publicly available data from Our World in Data were used, with seven-day smoothed new cases selected as the primary prediction target. An SEIR compartmental model was fitted to represent susceptible, exposed, infectious, and recovered or removed population dynamics. A Random Forest model was then used for AI-only prediction and for residual correction in the hybrid model. The hybrid framework was evaluated against a naive baseline, SEIR-only model, and AI-only model using MAE, RMSE, MAPE, and R^2. Results showed that the naive baseline performed best during the low-transmission testing period, but the hybrid SEIR–AI model improved RMSE and MAPE compared with the SEIR-only and AI-only models. The findings indicate that hybrid mathematical–AI modeling can improve mechanistic prediction while preserving biological interpretability, supporting its broader use in biomathematics, epidemiology, digital health, and complex biological forecasting.

Keywords: Mathematical modeling; Artificial intelligence; SEIR model; Random Forest; COVID-19 dynamics; Hybrid modeling.

1. Introduction

Biological and computational systems are nonlinear, adaptive, and affected by many interacting variables, making them challenging to predict with one modeling approach. Biological processes like disease spread, immune response, cancer, microbial ecology, and population health are emergent phenomena at the molecular, cellular, organismal, environmental, and computational levels. Thus, prediction needs a means to integrate mechanistic biological structure with flexibility. Mathematical models offer a structured approach to capture biological mechanisms, predict system behavior, and simulate future scenarios. Ordinary differential equations, stochastic models, network models, compartmental models, and computer simulations can be used to quantify biological processes and interpret them mechanistically. Such methods are particularly useful when predictions are needed to be coupled with an understanding of the processes that produce the outcomes of interest (Ji et al., 2017). But biological data are frequently noisy, sparse, nonlinear, high-dimensional, and influenced by unobserved factors such as behaviour, policy, vaccination, delays in reporting, and environmental factors. The above problems confirm the necessity of the integrated modelling methods, which are more than the mechanistic ones (Zinovyev, 2015). The potential of artificial intelligence (AI) to extract complex patterns, learn nonlinear associations, and improve predictive accuracy from large amounts of data has made it more and more widely used in biological and biomedical research. Applications of machine learning and deep learning include diagnosis, drug design, image recognition, genomics, clinical prediction and decision-making, systems biology, and public health.

AI can be particularly useful when there is only partial observation of biological processes, or when these processes are not easily modeled (Hassoun et al., 2021). AI-only models, however, are not interpretable, are very sensitive to data quality, and do not provide biological insight. Interpretability is important in biomedical contexts where scientists and doctors need to understand the rationale behind the prediction of a model. Hybrid models provide an approach to integrate mechanistic and AI models. Such models utilize mechanistic models that represent biological processes and AI that learns residual patterns or nonlinear corrections that can't be represented by equations.

. This integration can improve prediction while preserving interpretability. The importance of integrating mechanistic models and AI has been acknowledged in biomedical decision-making, particularly when model predictions should be meaningful (Benzekry, 2020). Similar considerations are needed in computational systems biology, where the modeling and control of disease require the integration of biological knowledge, quantitative models, and computational prediction at multiple scales (Yue & Dutta, 2022).

Disease transmission dynamics are a good example of hybrid modeling as epidemics are nonlinear, time-varying, and influenced by biological, population, testing, vaccination, public health, and reporting factors. SEIR models offer a mechanistic description of epidemic dynamics in terms of susceptible, exposed, infectious, and removed or recovered individuals, but may fail to capture sudden shifts in the epidemic. AI can help overcome this by capturing the nonlinear residual structure of observed data and additional predictors. This approach is in line with predictive biology, which uses mathematical and computational models to not only describe biological systems but also predict and inform future biological insights (Lopatkin & Collins, 2020). This study seeks to establish a mathematical modeling and artificial intelligence-based hybrid approach for predictive analysis in complex biological and computational systems, with a case study on the dynamics of COVID-19 spread in India. The research combines an SEIR model with residual learning using Random Forest and compares the naive, SEIR, AI-only, and hybrid SEIR-AI models using common predictive measures. The objective is to evaluate whether AI-based residual correction can improve mechanistic prediction while preserving biological interpretability.

2. Literature Review

Mathematical models and artificial intelligence play a role in the study of complex biological and computational systems. Biological systems may have multiple factors, be nonlinear, and adaptive. Biological mechanisms can be captured in mathematical models, and AI techniques can be applied to prediction, pattern recognition, and decision making. Karaca and Baleanu (2022) pointed out the importance of evolutionary mathematical science, fractional modelling, and artificial intelligence to investigate the nonlinear dynamics of complex systems. The use of computational methods has boosted the ability to model, simulate, and predict complex systems. Sosnowski et al. (2021) surveyed the use of artificial intelligence and computational techniques for modeling, prediction, and optimization of these systems. Likewise, Krzywanski et al. (2024) studied current computational techniques in modeling, prediction, and optimization. They demonstrated the use of machine learning and numerical methods to improve the flexibility and speed of the model. AI-powered modelling has been backed by these studies when conventional analytical methods are inadequate. Biomedical processes, from disease to physiology to treatment, are commonly represented by mathematical models.

Liu et al. (2023) demonstrated that mathematical models can help to analyse medical issues by quantifying disease processes and processes related to healthcare. This is relevant for modelling infectious diseases as it can be used to model the spread of a disease in terms of biological processes, such as compartmental models like SEIR. Another area that is seeing a rise in the use of machine learning (ML) methods in biomedical applications is image analysis. Yu and Wang (2022) summarized the use of data mining and mathematical models in cancer prognosis and prediction and showed the application of a computational approach for biomedical decision making. While they applied this to cancer, the concept can be extended to infectious disease prediction, as they both involve prediction under uncertainty and complexity. Biswas and Chakrabarti (2020) also showed the need for AI-based systems biology in multi-omics cancer studies, which demonstrates how AI can extract features from high-dimensional data.

Mathematical and computational modeling can be applied to drug design and biomedical research. Hasan et al. (2022) reviewed the use of mathematical models and computational tools in drug design, demonstrating that quantitative models can help to predict, simulate, and optimise. This is in line with the general opinion that computational modeling is beneficial in various fields of life sciences, such as epidemiology, systems biology, and biomedical prediction.

Hybrid modeling is becoming popular as it incorporates the advantages of mechanistic models and AI models. Mechanistic models can explain biological processes with variables and parameters. But they may not be sufficiently complex. AI models can capture complex data patterns, but may not be explainable. Benzekry (2020) reviewed the use of artificial intelligence and mechanistic modeling for clinical decision-making in cancer, highlighting the need to combine predictive models with mechanistic models. Harkos et al. (2025) also demonstrated that mathematical modeling and AI can be integrated to enhance drug delivery and effectiveness in cancer. These papers support the approach in this work.

Mathematical modeling is also used in biology. Tedeschi (2023) noted that mathematical modeling paradigms should be chosen based on the biological question, data, and application. This justifies the SEIR model used in the current study because the spread of infectious diseases involves susceptible, exposed, infectious, and removed (recovered) individuals.

The SEIR model offers a biological framework, and the Random Forest model enhances prediction by capturing the nonlinear structure of the residual.

There is also growing interest in the use of AI in modeling life processes. Abosaooda et al. (2025) introduced a method to assess AI tools in mathematical modeling of various life processes, and suggested that AI could be used to develop and predict models when properly validated. Shanmugam et al. (2025) also noted the use of mathematical modeling and AI in nonlinear dynamical systems, suggesting the need for a combined approach to study complex systems. In general, the studies demonstrate that mathematical modeling and AI are complementary approaches for predictive analysis of biological systems. Mathematical models offer biological interpretability, and AI models enhance adaptability and learning. But many studies use the two methods in isolation. This calls for a need for hybrid models that incorporate biological structure and data correction. The current study overcomes this by combining an SEIR model with Random Forest-based residual learning for COVID-19 predictions.

3. Materials and Methods

3.1 Study Design and Dataset

This study adopted a quantitative computational modeling design to develop a hybrid mathematical modeling and artificial intelligence framework for predictive analysis in complex biological and computational systems. The study used the spread of COVID-19 as a case study for biological systems, as the transmission dynamics of infectious diseases are highly dynamic, nonlinear, and depend on biological, social, and intervention-related factors.

The study adopted the open-access Our World in Data COVID-19 data. The India data subset was chosen because it offers a long time series that can be used for mathematical modelling and machine learning-based forecasting. The dataset includes daily epidemiological metrics such as the number of cases and deaths, vaccine-related metrics, reproduction rate, and population data. The key variables used in this study are listed in Table 1.

Table 1. Main dataset variables used in the study

Variable	Description	Role in the study
date	Calendar date of observation	Time index
population	Total population size	Used as the population size in the model
total_cases	Cumulative confirmed cases	Used for descriptive analysis
new_cases	Daily reported confirmed cases	Raw case indicator
new_cases_smoothed	Seven-day smoothed daily cases	Main prediction target
total_deaths	Cumulative confirmed deaths	Secondary epidemiological indicator
new_deaths_smoothed	Seven-day smoothed daily deaths	Secondary outcome variable
reproduction_rate	Estimated effective reproduction rate	Predictor for an AI model
people_vaccinated	Number of people receiving at least one vaccine dose	Predictor for an AI model
people_fully_vaccinated	Number of fully vaccinated individuals	Predictor for an AI model
stringency_index	Public health restriction index	Predictor for an AI model

The main outcome variable was new_cases_smoothed, which represents the seven-day smoothed number of new confirmed COVID-19 cases. This variable was selected instead of raw daily cases because it reduces the effects of reporting delays, weekly reporting cycles, and sudden administrative corrections.

3.2 Data Preprocessing and Feature Preparation

The dataset was first filtered to retain only India-specific observations. The date column was converted into a chronological date format, and the data were sorted by time. Missing values were handled according to variable type. Cumulative variables such as total_cases and total_deaths were treated using forward filling where appropriate, while short gaps in time-dependent variables were handled using interpolation or exclusion when necessary. To support prediction, additional time-series features were created from the smoothed case series. These features included lagged case counts, rolling averages, reproduction rate, vaccination-related features, and values from the mathematical model. The engineered features are summarised in Table 2.

Table 2. Engineered features used for AI-based prediction

Feature	Description	Purpose
1-day lag	Case value from the previous day	Captures immediate temporal dependency
7-day lag	Case value from seven days earlier	Captures weekly disease/reporting pattern
14-day lag	Case value from fourteen days earlier	Captures delayed transmission trends

7-day rolling mean	Average case value over the previous seven days	Represents short-term trend
14-day rolling mean	Average case value over the previous fourteen days	Represents medium-term trend
Reproduction rate	Estimated transmission intensity	Captures epidemic growth potential
Vaccination variables	Vaccination coverage indicators	Represents intervention and immunity effects
SEIR prediction	Output from the mathematical model	Provides mechanistic information to the AI model
SEIR residual	Difference between observed and SEIR-predicted values	Used for hybrid residual learning

The data were split chronologically into a training and a test set. The initial 80% of data was used to train the model, while the remaining 20% was used for testing. The data was split chronologically rather than randomly to prevent "information leaks" from future observations.

3.3 Mathematical Model

The study's mathematical model was an SEIR compartmental model. This model assumes that the population is divided into four groups: susceptible, exposed, infectious, and removed (or recovered). This model was chosen because it has an exposed compartment, which is relevant to COVID-19, as there is a delay before people become infectious. The compartments and parameters used in the SEIR model are described in Table 3.

Table 3. SEIR model compartments and parameters

Symbol	Meaning	Description
$S(t)$	Susceptible population	Individuals who can become infected
$E(t)$	Exposed population	Infected individuals are not yet infectious
$I(t)$	Infectious population	Individuals capable of transmitting infection
$R(t)$	Recovered/removed population	Individuals no longer infectious
N	Total population	Sum of all compartments
β	Transmission rate	The rate at which susceptible individuals become infected
σ	Progression rate	The rate at which exposed individuals become infectious
γ	Recovery/removal rate	The rate at which infectious individuals recover or are removed

The SEIR model was expressed using the following system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \frac{dE}{dt} = \beta\frac{SI}{N} - \sigma E, \frac{dI}{dt} = \sigma E - \gamma I, \frac{dR}{dt} = \gamma I$$

In this model, β controls the rate of new infections, σ controls the movement from the exposed to the infectious stage, and γ controls recovery or removal from the infectious stage. The model was solved numerically to generate predicted disease trajectories over time.

The predicted case series from the SEIR model was compared with the observed smoothed new cases. The model parameters were obtained by fitting the model to the observed data by least-squares optimization. The SEIR fit was then used as a mathematical prediction and to drive the hybrid model.

3.4 Artificial Intelligence and Hybrid Modeling Approach

An artificial intelligence model was developed to forecast future COVID-19 case trends using the engineered features described in Table 2. The AI forecasting model was a Random Forest regression, as it can learn complex interactions and nonlinear dependencies between temporal, epidemiological, and intervention features without strong parametric assumptions.

The AI model used lagged case values, rolling averages, reproduction rate, vaccination-related features, stringency index, time index, and SEIR predictions as features. The target variable was the value of new_cases_smoothed with a seven-day look-ahead. The hybrid model was trained using residual learning. The SEIR model was first trained on the observed smoothed case data. The residual difference between the observed data and SEIR-predicted data was then determined. The Random Forest model was then trained to predict this residual. The hybrid forecast was the sum of the SEIR prediction and the AI-predicted residual correction.

The hybrid model can be summarized as:

$$\hat{Y}_{Hybrid}(t) = \hat{Y}_{SEIR}(t) + \hat{e}_{AI}(t)$$

where $\hat{Y}_{Hybrid}(t)$is the final hybrid prediction, $\hat{Y}_{SEIR}(t)$is the SEIR model prediction, and $\hat{e}_{AI}(t)$is the residual correction learned by the AI model.

The models compared in the study are presented in Table 4.

Table 4. Models compared in the study

Model	Description	Purpose
Naive baseline	Uses the most recent observed value as the forecast	Simple benchmark
SEIR-only model	Uses only the mathematical SEIR model	Mechanistic baseline
AI-only Random Forest model	Uses Random Forest regression without SEIR residual correction	Data-driven baseline
Hybrid SEIR–AI model	Combines SEIR prediction with Random Forest-based residual correction	Proposed hybrid model

This comparison allowed the study to evaluate whether the hybrid approach improved prediction accuracy compared with standalone mathematical and AI-based methods.

3.5 Model Evaluation and Computational Tools

The models were assessed based on mean absolute error, root mean square error, mean absolute percentage error and coefficient of determination. These metrics were selected because they measure both the magnitude of prediction error and the proportion of variance explained by the model. Table 5 shows the evaluation metrics.

Table 5. Model evaluation metrics

Metric	Description	Preferred result
MAE	Average absolute difference between observed and predicted values	Lower is better
RMSE	Square-root average of squared prediction errors	Lower is better
MAPE	Percentage-based prediction error	Lower is better
R^2	Proportion of variation explained by the model	Higher is better

The analysis was performed using Python. Data analysis was performed using Pandas and NumPy. The SEIR model was solved and parameters were estimated using SciPy. Scikit-learn and XGBoost were used for AI modeling and evaluation. Matplotlib was used to plot the case trends. The software used in this study is listed in Table 6.

Table 6. Computational tools used in the study

Tool or package	Purpose
Python	Main computational environment
Pandas	Data cleaning and manipulation
NumPy	Numerical computation
SciPy	SEIR model solution and parameter estimation
Scikit-learn	Machine learning and evaluation metrics
XGBoost	Gradient boosting regression model
Matplotlib	Data visualization

3.6 Ethical Considerations

This study used publicly available aggregate epidemiological data and did not involve individual-level human participant information. Therefore, ethical approval was not required. The analysis was conducted for computational modeling and public health research purposes only.

4. Results

4.1 Descriptive Characteristics of the Dataset

The India subset of the Our World in Data COVID-19 dataset was analyzed for mathematical modeling and AI-based prediction. Post removal of incomplete final observations, the data set consisted of 1,674 observations for each day from 5 January 2020 to 4 August 2024, with a total population of 1,417,173,120. Descriptive statistics of the variables of interest are in Table 7.

Table 7. Descriptive summary of selected variables in the India COVID-19 dataset

Variable	Mean	Median	Minimum	Maximum	Missing values
new_cases	26,906.66	0.00	0.00	2,738,957.00	0
new_cases_smoothed	26,987.11	3,286.57	0.00	391,279.57	5
total_cases	31,110,803.59	43,042,097.00	0.00	45,041,748.00	0
new_deaths_smoothed	319.73	29.14	0.00	4,140.29	5

total_deaths	380,320.23	521,751.00	0.00	533,623.00	0
reproduction_rate	1.04	0.95	0.34	2.67	650
people_vaccinated	836,445,500.00	1,026,754,040.00	0.00	1,027,438,924.00	437
people_fully_vaccinated	725,876,700.00	950,118,274.00	7,668.00	951,990,552.00	472
stringency_index	58.08	61.57	0.00	100.00	582

As shown in Table 7, new_cases_smoothed, the main prediction variable, had a mean of 26,987.11 cases per day and a maximum of 391,279.57. This variable was used because it reduces daily reporting irregularities and short-term fluctuations. Epidemiological variables had minimal missingness, while reproduction_rate, vaccination rates, and stringency_index had more missingness and were used as AI secondary predictors. Figure 1 shows the trend of seven-day smoothed new cases.

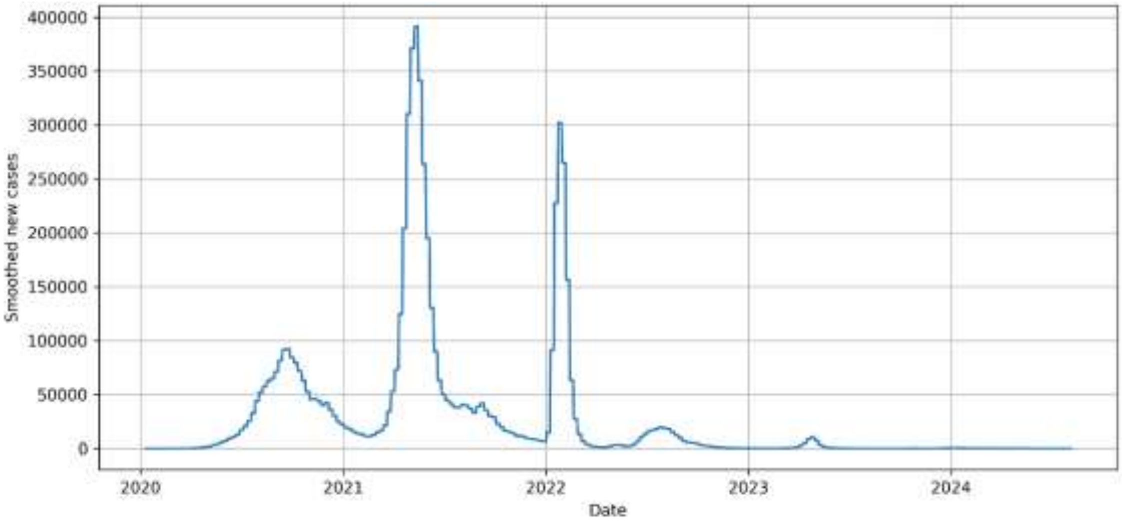

Figure 1. Seven-day smoothed new COVID-19 cases in India

Figure 1 shows multiple epidemic waves, with the largest increase occurring in 2021. This wave-like pattern justified the use of a hybrid model that incorporates both the biological structure of transmission and data-driven forecasting.

4.2 Trends in Confirmed Cases and Deaths

The total number of COVID-19 cases grew significantly during the period of the study, with a peak of 45,041,748. The cumulative case curve is in Figure 2.

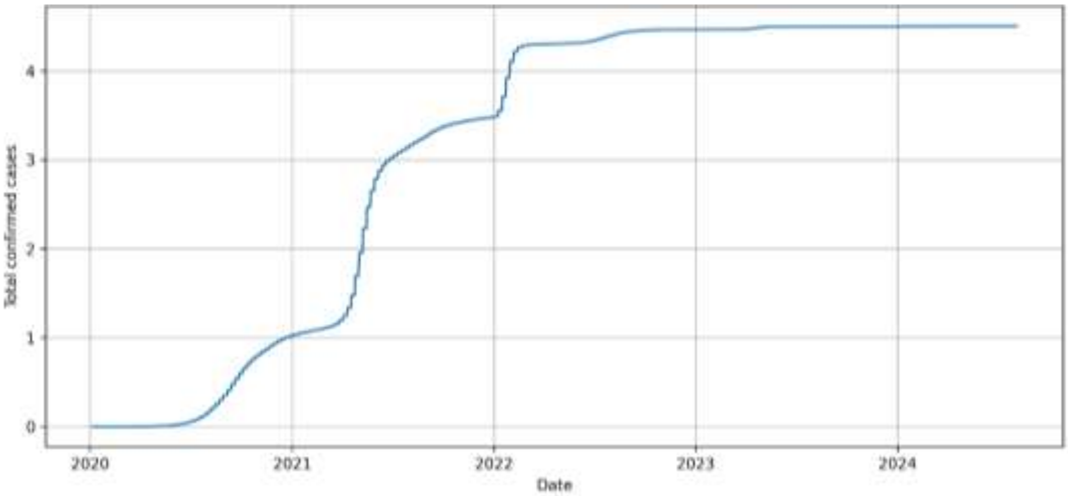

Figure 2. Cumulative confirmed COVID-19 cases in India

Figure 2 shows cumulative cases increased rapidly during epidemic waves and plateaued in periods of low transmission. Figure 3 shows the trend in new deaths (seven-day smoothed).

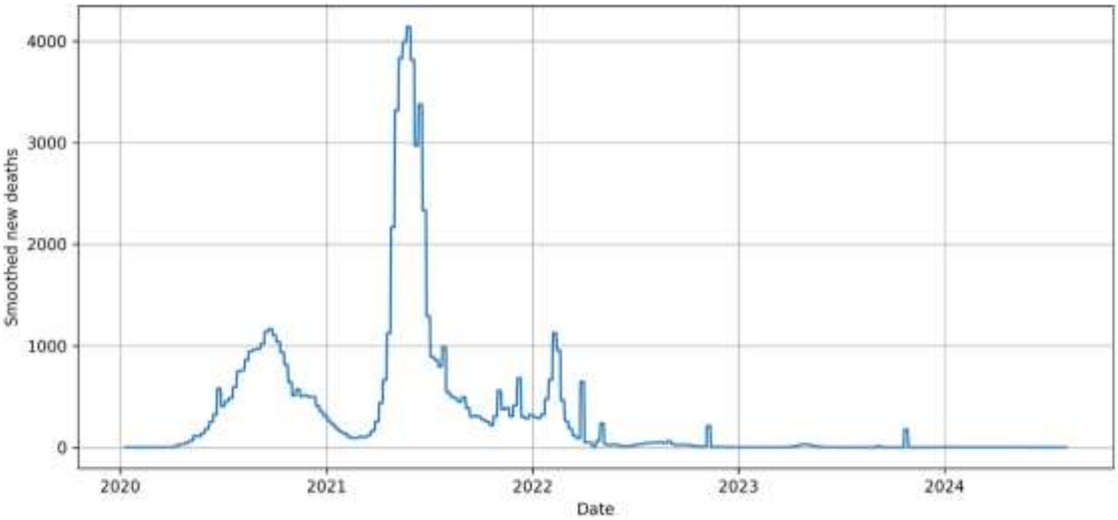

Figure 3. Seven-day smoothed new COVID-19 deaths in India

Figure 3 illustrates that the smoothed number of new deaths fluctuated, with the greatest number of deaths following the peak number of new cases. This suggests a lag between the case numbers and deaths. The peak values are in Table 8.

Table 8. Peak values observed in the India COVID-19 dataset

Indicator	Peak value	Date of peak
Smoothed new cases	391,279.57	9 May 2021
Raw new cases	2,738,957.00	9 May 2021
Smoothed new deaths	4,140.29	23 May 2021
Total confirmed cases	45,041,748.00	End of study period
Total confirmed deaths	533,623.00	End of study period

As shown in Table 8, smoothed new cases peaked on 9 May 2021, while smoothed new deaths peaked on 23 May 2021, approximately two weeks later. The peak of raw daily cases is much higher than the peak of smoothed cases, suggesting reporting corrections or backlog, so new_cases_smoothed is a good predictor.

4.3 Vaccination and Public Health Indicators

Vaccination and public health indicators were included as auxiliary predictors because they may influence transmission, disease severity, and future case trends. The peak number of people who received at least one dose of the vaccine was 1,027,438,924 and 951,990,552 were fully vaccinated. The vaccination trend is shown in Figure 4.

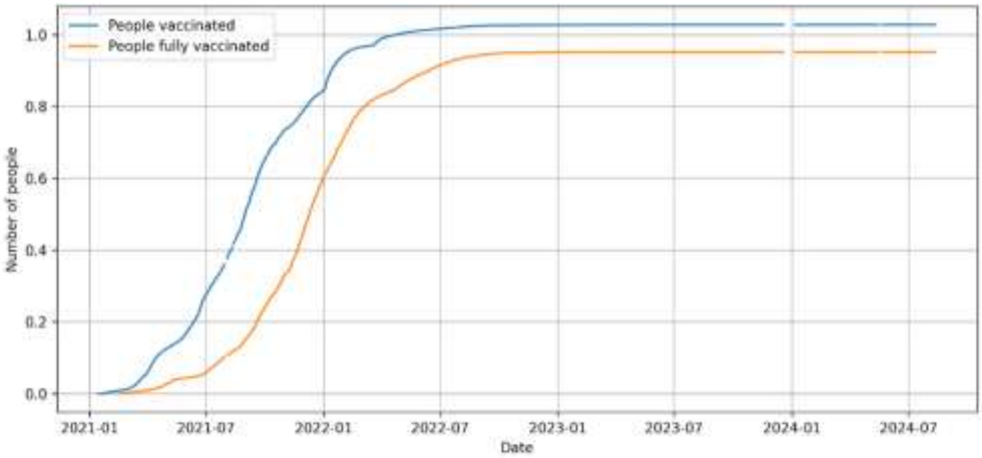

Figure 4. COVID-19 vaccination indicators in India

As shown in Figure 4, vaccination coverage increased rapidly after rollout and later reached a plateau. Vaccination variables were used as auxiliary variables in the AI and hybrid models due to unavailability in the early stages of the pandemic. The stringency_index was also used after pre-processing to account for public health measures and other policies.

4.4 SEIR Model Fitting Results

The SEIR model was used to fit the smoothed daily confirmed case series to describe COVID-19 spread with susceptible, exposed, infectious and removed/recovered compartments. SEIR parameters were calibrated using constrained parameter ranges.

The time frame for the model was 1 March 2020 to 4 August 2024. Following feature engineering and exclusion of incomplete lagged observations, 1,583 observations were used to model the data. The training data set had 1,266 observations from 29 March 2020 to 15 September 2023, and the testing data set had 317 observations from 16 September 2023 to 28 July 2024. SEIR parameters are shown in Table 9.

Table 9. Estimated SEIR model parameters

Parameter	Estimated value	Interpretation
β	0.1673	Transmission rate
σ	0.2222	The rate at which exposed individuals become infectious
γ	0.1429	Recovery or removal rate
R_0	1.1709	Basic reproduction number
$1/\sigma$	4.50 days	Average exposed-to-infectious transition period
$1/\gamma$	7.00 days	Average infectious period
Reporting scale factor	0.0356	Scaling factor linking modeled infections to reported cases
Initial exposed population	1,000	Initial exposed count used in calibration
Initial infectious population	200	Initial infectious count used in calibration

As shown in Table 9, the estimated R_0 was 1.1709, indicating epidemic growth potential. The exposed-to-infectious transition period was 4.50 days, and the infectious period was 7.00 days. The fitted SEIR trajectory is shown in Figure 5.

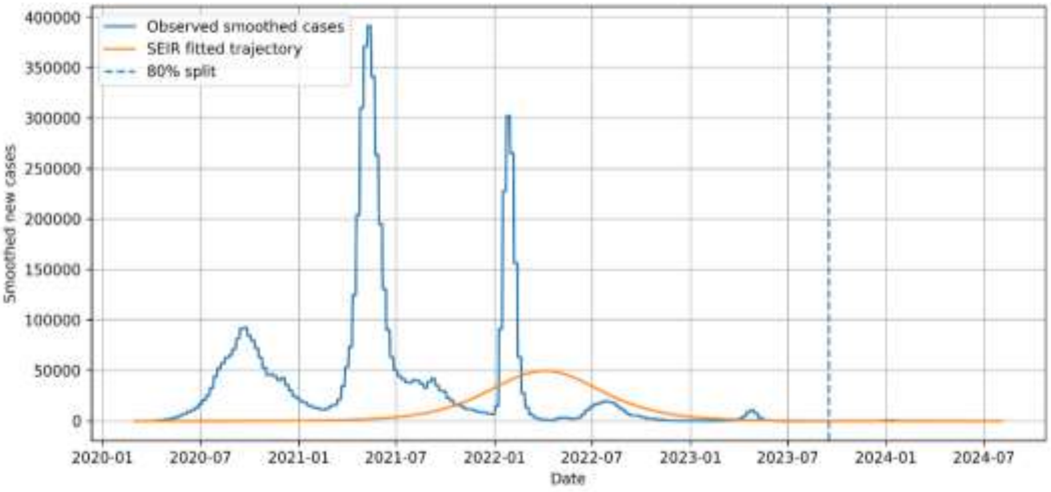

Figure 5. Observed and SEIR-predicted smoothed new COVID-19 cases

In Figure 5, the SEIR model reproduced the general pattern of transmission, but not all waves of the epidemic. This reflects the limitations of a basic SEIR model under changing real-world conditions such as policy shifts, vaccination, behavioral responses, reporting variation, and viral variants.

4.5 Artificial Intelligence Model Results

The AI-only model was trained with engineered time-series predictors, such as smoothed cases with lag, rolling averages, reproduction rate, vaccination variables, stringency index, time index, and SEIR model predictions. Random Forest regression was used as it can model non-linearities and interactions without strong assumptions. The predictors are presented in Table 10.

Table 10. Predictor variables used in the AI-only and hybrid models

Predictor	Description	Analytical role
1-day lag	Previous day’s smoothed case count	Captured immediate temporal dependency
7-day lag	Smoothed case count seven days earlier	Captured weekly pattern
14-day lag	Smoothed case count fourteen days earlier	Captured the delayed epidemic trend
21-day lag	Smoothed case count twenty-one days earlier	Captured longer temporal dependency

28-day lag	Smoothed case count twenty-eight days earlier	Captured the monthly trend behavior
7-day rolling mean	Short-term average of smoothed cases	Represented the short-term trend
14-day rolling mean	Medium-term average of smoothed cases	Represented broader temporal trend
28-day rolling mean	Longer moving average of smoothed cases	Represented slower changes in disease burden
Reproduction rate	Estimated transmission intensity	Represented epidemic growth potential
Vaccination variables	Vaccination coverage indicators	Represented intervention and immunity effects
Stringency index	Public health restriction indicator	Represented policy response
Time index	Sequential day count	Represented long-term temporal progression
SEIR prediction	Mechanistic model output	Added biological information to the predictive model

The AI-only model was trained on the chronological training set and tested with a seven-day horizon. The observed and AI-predicted cases are in Figure 6.

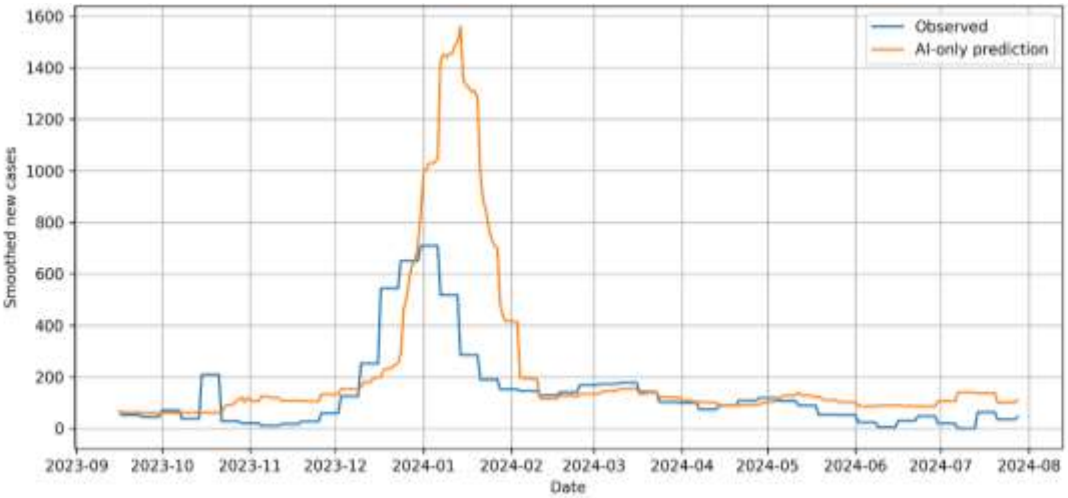

Figure 6. Observed and AI-predicted smoothed new COVID-19 cases

Figure 6 shows that the AI-only model captured part of the temporal movement in the testing period but did not outperform the simpler reference forecast. Its predictions were flexible but less biologically interpretable because they relied on engineered statistical predictors rather than explicit transmission compartments.

4.6 Hybrid SEIR–AI Model Results

The hybrid SEIR-AI model was implemented using residual learning. The SEIR model was used to make mechanistic predictions, and the residual error between the observations and SEIR predictions was determined. The residual pattern was learned by a Random Forest model, and the hybrid prediction was produced by adding the residual prediction to the SEIR prediction. The hybrid model components are presented in Table 11.

Table 11. Hybrid SEIR–AI model components

Component	Function	Output
SEIR model	Represents biological disease transmission	SEIR-predicted case trend
Residual calculation	Measures the unexplained SEIR prediction error	Residual error series
AI residual model	Learns nonlinear residual patterns	AI-predicted residual correction
Hybrid prediction	Adds residual correction to SEIR prediction	Final hybrid forecast

The observed, SEIR-only, and hybrid SEIR–AI predictions are compared in Figure 7.

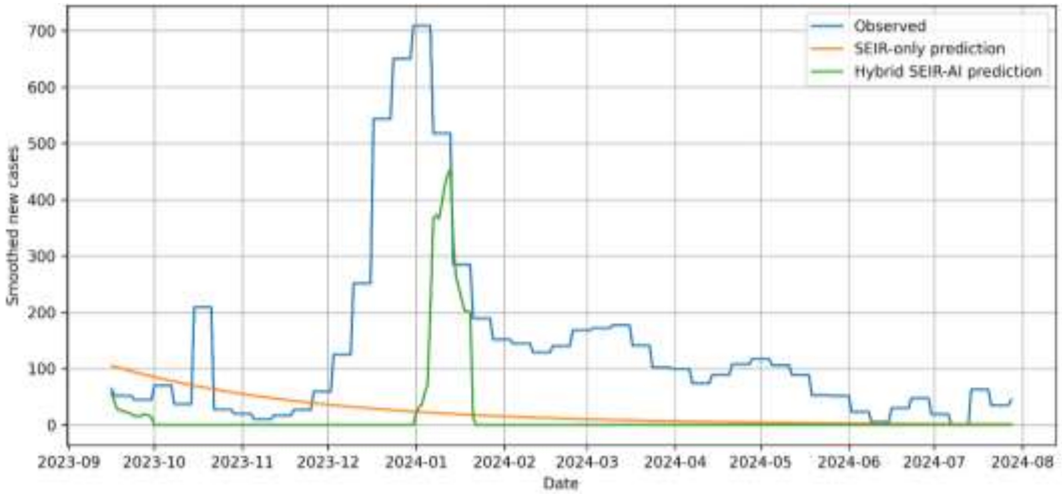

Figure 7. Observed, SEIR-only, and hybrid SEIR–AI-predicted smoothed new COVID-19 cases

Figure 7 illustrates that the hybrid approach could reduce some of the prediction error of the SEIR-only model using AI-based residual correction. While there was still some prediction error, the hybrid model improved the mechanistic model by adding data-driven correction.

4.7 Comparative Predictive Performance

The naive baseline, SEIR-only model, AI-only Random Forest model, and hybrid SEIR–AI model were evaluated using MAE, RMSE, MAPE, and R^2. The results are presented in Table 12.

Table 12. Comparative predictive performance of evaluated models

Model	MAE	RMSE	MAPE	R^2
Naive baseline	48.99	81.75	61.91	0.7436
SEIR-only model	126.56	200.00	98.75	-0.5347
AI-only random forest	120.75	250.93	168.33	-1.4160
Hybrid SEIR–AI model	120.49	188.86	94.00	-0.3686

As shown in Table 12, the naive baseline produced the lowest MAE and RMSE because the testing period contained low and relatively stable case counts. However, it did not provide a biological interpretation. Among the mechanistic and AI-assisted models, the hybrid SEIR–AI model performed best, with lower RMSE and MAPE than both the SEIR-only and AI-only models. The graphical comparison of prediction errors is shown in Figure 8.

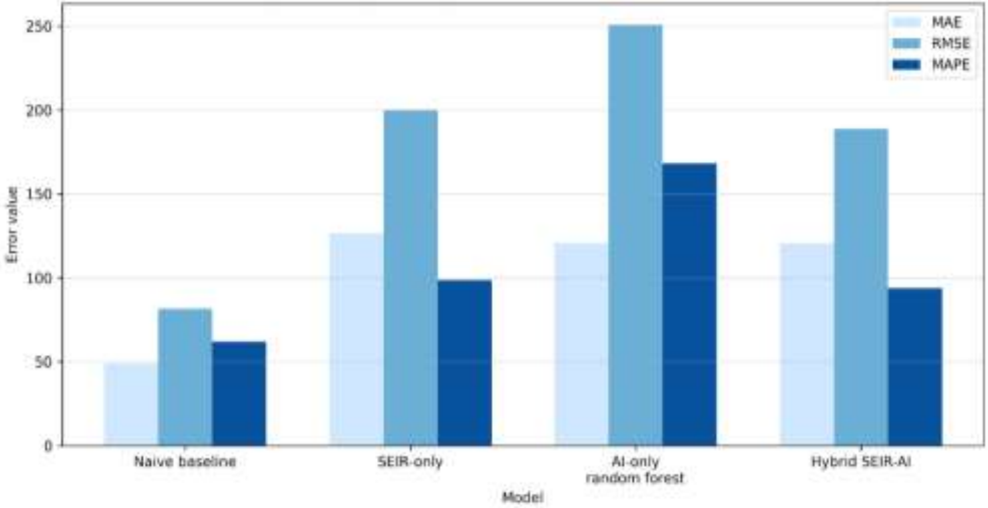

Figure 8. Comparative prediction error across evaluated models

Figure 8 shows that the hybrid SEIR–AI model reduced error compared with the SEIR-only and AI-only models, although the naive baseline remained strongest during the low-transmission testing period. The comparative usefulness of each model is presented in Table 13.

Table 13. Integrated interpretation of model performance and usefulness

Model	Predictive role	Main strength	Main limitation	Interpretation
Naive baseline	Simple reference forecast	Produced the lowest test error during the low-transmission testing period	Does not represent disease dynamics or biological mechanisms	Useful as a benchmark, especially when cases are stable, but limited for mechanistic understanding
SEIR-only model	Mechanistic disease transmission model	Provides biological interpretability through susceptible, exposed, infectious, and recovered compartments	Limited ability to capture multiple waves, reporting changes, vaccination effects, and policy shifts	Useful for explaining transmission structure, but insufficient as a standalone forecasting model
AI-only random forest	Data-driven prediction model	Captures nonlinear relationships among lagged, rolling, epidemiological, and intervention-related predictors	Less biologically interpretable and sensitive to training-test distribution differences	Useful for flexible prediction but weaker for biological explanation
Hybrid SEIR–AI model	Mechanistic model with AI-based residual correction	Improved RMSE and MAPE compared with SEIR-only and AI-only models	Did not outperform the naive baseline in the low-transmission testing period	Most useful when both biological interpretability and improved mechanistic prediction are required

As shown in Table 13, the hybrid SEIR–AI model provided the most balanced approach by improving SEIR-based prediction while retaining biological structure.

5. Discussion

In this study, a hybrid SEIR-AI framework for predictive analysis in complex biological and computational systems using COVID-19 dynamics in India as a case study was developed. The results showed that the SEIR model gave a biological interpretation of disease transmission. The Random Forest part improved prediction by learning residual patterns that were not captured by the mathematical model. The descriptive results showed the existence of several epidemic waves, with smoothed new cases peaking on 9 May 2021 and smoothed new deaths peaking on 23 May 2021. This time lag between cases and deaths reflects the expected delay between infection, disease progression, and mortality.

The SEIR model predicted $R_0 = 1.1709$, indicating that the epidemic could grow when the susceptible population was sufficiently large. The period from exposure to infectiousness was 4.50 days with an infectious period of 7.00 days. These parameters seemed to be biologically meaningful. But all the epidemic waves were not fully reproduced in the SEIR model. The modelling of transitions in basic models is simplified and does not directly include behavioural change, vaccination, restrictions, variants, or reporting. This is consistent with the previous comparative studies on COVID-19 modelling, which have shown that mathematical pandemic models can be helpful to model the structure of an epidemic but still depend on model assumptions, parameter estimates, and evolving public health conditions (Adiga et al., 2020).

AI-only Random Forest learned some temporal movement in the testing period, but was not able to beat the naive baseline. This was impacted by the low and comparatively stable case counts over this time period, during which persistence-based forecasting is likely to be more successful. However, the AI-only model was biologically less interpretable by being based on statistical patterns among engineered predictors, instead of explicit disease-transmission compartments. This indicates a broader challenge in modeling for healthcare. Machine learning can help with prediction, but often needs to be combined with interpretable mathematical constructs for scientific and clinical understanding (Hassan et al., 2024).

The hybrid SEIR AI model provided the most balanced approach based on modeling. It lowered the RMSE from 200.00 in the SEIR-only model to 188.86, and it reduced the MAPE from 98.75% to 94.00%. It has also performed better than the AI-only Random Forest model, which yielded an RMSE of 250.93 and an MAPE of 168.33%. The naive baseline was the least accurate of all the baselines, but lacked biological interpretation and did not clarify transmission mechanisms. Therefore, the hybrid model proved to be successful in combining mechanistic knowledge with data-driven correction. This reflects the broader use of mathematical modeling and learning-based prediction techniques in nonlinear health-related systems (Sayed-Ahmed et al., 2024). There are some limitations in the study. First, there is a possibility that the data on a national level may not capture the entire regional variations in Indian States and Districts. Second, the number of confirmed cases is subject to changes in testing capacity, reporting practices, and administrative corrections. Third, there were some simplifications in the SEIR model, including fixed transition parameters and homogeneous population mixing. Fourth, the absence of values for the reproduction rate, vaccination variables, and stringency index could have influenced the AI. Finally, the testing period consisted of low, stable cases, favoring the naïve baseline, and reducing the apparent benefit of more sophisticated models. The results indicate that mathematical models and Artificial Intelligence can be used together and not against each other. Mechanism models add biological structure; AI models add flexibility, learning from observation. This work should be extended to regional COVID-19 data, directly model vaccination by using an SEIRV model, compare other AI models such as XGBoost or LSTM, and bring in some uncertainty quantification. The same hybrid strategy could be used for the progression of cancer, the immune response, the microbiome, antimicrobial resistance, and ecological forecasting.

Conclusion

The results of this study prove that artificial intelligence and mathematical modeling can successfully be combined for predictive analysis in complex computational and biological systems. The study used COVID-19 dynamics in India as a representative case study. The study showed that the SEIR model gave a biologically interpretable structure to understand the disease transmission, and the Random Forest component increased the flexibility by learning non-linear residual patterns that are not captured fully by the mechanistic model. The results revealed that the dynamics of COVID-19 cases were nonlinear and wave-like with significant differences between raw and smoothed case counts, advocating for the use of smoothed data for stable predictions. Although the naive baseline had the lowest overall error during the low-transmission testing period, it was biologically meaningless and did not provide any explanation for the mechanisms of transmission. Among the model-based approaches, the hybrid SEIR–AI model achieved better RMSE and MAPE than the SEIR-only and AI-only models, which indicates that residual learning improved mechanistic prediction. The results show that hybrid modeling provides a good trade-off between interpretability and predictive performance. This approach is not only applicable to infectious disease forecasting, but also to other biological systems with nonlinear dynamics, uncertainty, and complex interactions. Future applications may include extending the framework to regional disease modeling, SEIRV structures, uncertainty quantification, and other biomedical or ecological prediction problems.

References

1. Abosaooda, H. N., Ariffin, S. B., Alyasiri, O. M., & Noor, A. A. (2025). Evaluating the Effectiveness of AI Tools in Mathematical Modelling of Various Life Phenomena: A Proposed Approach. InfoTech Spectrum: Iraqi Journal of Data Science, 2(1), 16-25.
2. Adiga, A., Dubhashi, D., Lewis, B., Marathe, M., Venkatramanan, S., & Vullikanti, A. (2020). Mathematical models for the COVID-19 pandemic: a comparative analysis. Journal of the Indian Institute of Science, 100(4), 793.
3. Benzekry, S. (2020). Artificial intelligence and mechanistic modeling for clinical decision making in oncology. Clinical Pharmacology & Therapeutics, 108(3), 471-486.
4. Biswas, N., & Chakrabarti, S. (2020). Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. Frontiers in Oncology, 10, 588221.
5. Harkos, C., Hadjigeorgiou, A. G., Voutouri, C., Kumar, A. S., Stylianopoulos, T., & Jain, R. K. (2025). Using mathematical modelling and AI to improve delivery and efficacy of therapies in cancer. Nature Reviews Cancer, 25(5), 324-340.
6. Hasan, M. R., Alsaiari, A. A., Fakhurji, B. Z., Molla, M. H. R., Asseri, A. H., Sumon, M. A. A., ... & Kim, B. (2022). Application of mathematical modeling and computational tools in the modern drug design and development process. Molecules, 27(13), 4169.
7. Hassan, J., Saeed, S. M., Deka, L., Uddin, M. J., & Das, D. B. (2024). Applications of machine learning (ML) and mathematical modeling (MM) in healthcare with special focus on cancer prognosis and anticancer therapy: current status and challenges. Pharmaceutics, 16(2), 260.
8. Hassoun, S., Jefferson, F., Shi, X., Stucky, B., Wang, J., & Rosa Jr, E. (2021). Artificial intelligence for biology. Integrative and Comparative Biology, 61(6), 2267-2275.
9. Ji, Z., Yan, K., Li, W., Hu, H., & Zhu, X. (2017). Mathematical and computational modeling in complex biological systems. BioMed Research International, 2017(1), 5958321.
10. Karaca, Y., & Baleanu, D. (2022). Evolutionary mathematical science, fractional modeling, and artificial intelligence of nonlinear dynamics in complex systems. Chaos Theory and Applications, 4(3), 111-118.
11. Krzywanski, J., Sosnowski, M., Grabowska, K., Zylka, A., Lasek, L., & Kijo-Kleczkowska, A. (2024). Advanced computational methods for modeling, prediction, and optimization—a review. Materials, 17(14), 3521.
12. Liu, Y., Wu, R., & Yang, A. (2023). Research on medical problems based on mathematical models. Mathematics, 11(13), 2842.
13. Lopatkin, A. J., & Collins, J. J. (2020). Predictive biology: modelling, understanding and harnessing microbial complexity. Nature Reviews Microbiology, 18(9), 507-520.
14. Our World in Data. (2024). Data on COVID-19 (coronavirus): Cases, deaths, hospitalizations, tests, and vaccinations. Our World in Data. https://github.com/owid/covid-19-data/tree/master/public/data
15. Sayed-Ahmed, N. M. Z., Hossain, M. A., Limkar, S., El-Bahkiry, H. S., Alam, N., & Amins, S. T. (2024). Mathematical modelling and deep learning techniques for predicting. Math. Model, 34(4), 230-244.
16. Shanmugam, L., Lim, C. P., Rihan, F. A., & Kurths, J. (2025). Dynamical analysis of nonlinear physical systems: mathematical modelling, artificial intelligence, and applications. The European Physical Journal Special Topics, 234(8), 1331-1339.

17. Sosnowski, M., Krzywanski, J., & Ščurek, R. (2021). Artificial intelligence and computational methods in the modeling of complex systems. Entropy, 23(5), 586.
18. Tedeschi, L. O. (2023). The prevailing mathematical modeling classifications and paradigms support the advancement of sustainable animal production. Animal, 17, 100813.
19. Yu, C., & Wang, J. (2022). Data mining and mathematical models in cancer prognosis and prediction. Medical Review, 2(3), 285-307.
20. Yue, R., & Dutta, A. (2022). Computational systems biology in disease modeling and control: Review and perspectives. npj Systems Biology and Applications, 8(1), 37.
21. Zinovyev, A. (2015). Overcoming complexity of biological systems: from data analysis to mathematical modeling. Mathematical Modelling of Natural Phenomena, 10(3), 186-205.