FACULTY OF SCIENCE & TECHNOLOGY

BSc (Hons) Computing

May 2025

Classifying Outbreak Severity: A Web-Based SEIR Simulation Tool Enhanced Using Binary Classification

By Hill Lam

Faculty of Science & Technology
Department of Computing
and Informatics Final Year Project

Abstract

This project has implemented a web-based tool for simulating the spread of infectious diseases and classifying outbreak severity using a binary machine learning model. The system combines a modified SEIRSD (Susceptible–Exposed–Infected–Recovered–Susceptible–Deceased) features model with a dynamic network structure to capture more realistic transmission patterns, including reinfection and mortality. A Gaussian Naive Bayes classifier was trained on historical COVID-19 data using features derived from simulated infection curves, achieving a test accuracy of 65%. Although this failed of the initial 90% target, it reflected a methodologically sound and generalisable performance. The simulation interface, built with FastAPI and Plotly, allows users to explore outbreak dynamics by adjusting transmission parameters and viewing real-time network animations. Through experimental comparison with traditional ODE-based SEIR models and public COVID-19 case data, the network-based simulation demonstrated greater alignment with realistic epidemic curves. While simplified for performance and accessibility, the tool serves as a proof of concept for deploying interpretable epidemiological models in public or educational contexts.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

1. Introduction

Global pandemics pose significant threats to health, economies, and societies worldwide. According to van de Wijgert and Bonten (2025), Traditional methods often fail to recognise risks early due to low awareness in public health services. This leads to delays in reporting, incomplete data, and resource limitations, resulting in delayed responses and ineffective intervention strategies. Dyvik (2024) reports that during COVID-19 in 2020, the global GDP fell by 3.4%. However, the global economy was able to recover from it, reaching positive growth levels again in 2021. In addition to these economic challenges, the unemployment and labour market have major disruptions, The International Labour Organisation states that 25 million jobs were lost globally in 2020, and an additional 255 million full-time equivalents were lost due to reduced hours, which have not been fully recovered to date. To address and prevent future pandemics, Liu (2023) agreed that an early detection system is needed, and they had successfully implemented a simulation system that has been able to detect COVID-19 about 0.4 weeks earlier than it was officially found, potentially reducing cases and earlier response to control losses.

Despite significant advances in early detection systems Liu (2023), many existing solutions are still complicated, lacking interactivity and requiring specialist knowledge. This makes it difficult for the general public to understand disease dynamics or to understand meaningful insights. making it difficult to explore disease dynamics or to educate the public. This project aims to address these gaps by developing a web-based simulation tool based on the modified version of the SEIR (Susceptible-Exposed-Infectious-Recovered) model with network-based modelling. The tool allows users to input custom social network structures and adjust parameters such as infection rates, recovery times, and death rates, designed to be accessible to the general public to explore disease behaviours, help raise greater awareness and potentially reveal new insights into pandemic dynamics.

1.1 Objectives

The primary objectives will be listed in the table below:

ID	Objective	Success Criteria
1	Develop a web-based simulation tool that models disease spread using the SEIR model and network-based approaches.	The tool should accurately simulate disease spread dynamics, validated with historical data within a 10% margin of error.
2	Implement a binary classifier to classify simulated outbreaks as pandemic or localised, ensuring accuracy based on predefined criteria.	Achieve at least 90% accuracy in classifying outbreaks using synthetic datasets mimicking COVID-19 and seasonal flu scenarios.
3	allow users to input their own social network structures, allowing customisation of the simulation.	The tool successfully processes user-uploaded network data in standard formats without errors.
4	Provide real-time visualisations of simulation results to enhance user understanding and engagement.	Display dynamic charts or graphs that update in real-time as the simulation progresses.

Table 1: Project Objectives and Success Criteria

1.2 Risk Assessment

A careful risk assessment is needed before the project starts to identify every possible risk that might be encountered during the development phase, and methods to mitigate them.

ID	Risk	Likelihood	Impact	Mitigation Strategy
1	Limitation of the availability of data	Low	High	Use publicly available data like WHO or CDC to ensure sufficient data for simulation validation.
2	Difficulty in defining classification criteria	Low	High	Use established epidemiological thresholds from a well-known organisation or research to define classification criteria.
3	Technical Complexity	Moderate	High	Prioritize core features (e.g., SEIR simulation and classification) and defer non-essential functionalities.
4	Inaccurate simulation results	Moderate	High	Validate simulation outputs against historical data and adjust model parameters accordingly.
5	Time constraints	High	High	Create a detailed project timeline with Gantt chart and prioritize must-have features.
6	Poor design, UI leading to unclear instructions	Moderate	Moderate	Design a user-friendly interface with clear instructions and tooltips, tested with a sample user group for feedback.

7	Computationally heavy, leading to slow development and an inefficient website	Low	Moderate	During development, focus on using lightweight packages and tools for development.
8	Browser compatibility issues	Moderate	Moderate	Test the web-based tool on multiple browsers during development to ensure compatibility.

Table 2: Risk Assessment

2. Literature review

Global pandemics like COVID-19 have long-lasting effects on health, economic, and social systems and have shown the need for practical tools to model, predict, and manage disease dynamics. This literature review examines existing research on epidemiological modelling, such as the SEIR model, network-based simulation, outbreak classification, and the accessibility of simulation tools for the general public, to better clarify the research gap and improve understanding.

2.1 Epidemiological Modelling with SEIR

The SEIR model (Susceptible–Exposed–Infected–Recovered) was first introduced by Kermack and McKendrick (Hethcote, 2000) as an evolution of the SIR model (Susceptible–Infected–Recovered). SEIR is a widely used framework in infectious disease modelling. It segments the total population (N) into four components: S (Susceptible), E (Exposed), I (Infected), and R (Recovered). This structure allows researchers to capture key dynamics of disease progression and simulate how infections spread over time.

While the SEIR model has become popular because of its simplicity, researchers such as Yingze Hou (2024) and He et al. (2020) have discovered its limitations. The basic SEIR formulation assumes a homogeneous population and fixed transition parameters, which can oversimplify real-world dynamics.

Hou (2024) addressed these issues by introducing a multi-feature SEIR model that accounts for health condition variability and different levels of social activity. The model has improved prediction accuracy, however, it has a downside: the optimisation process required over 12 hours to run even for relatively small scenarios. Without clarification on the hardware used, it shows it has low computational efficiency or optimisation because of its detailed simulation, however, it is not suitable for this project, especially since it is important to require real-time or interactive use.

Similarly, He et al. (2020) have modified the SEIR model by introducing additional elements, which added quarantined and hospitalised groups. Their implementation used a Particle Swarm Optimisation (PSO) algorithm to dynamically tune parameters, this method has led to higher predictive accuracy during the COVID-19 pandemic. However, this approach have introduced significant technical complexity.

This project has taken inspiration from these studies while aiming to improve simplicity and usability. The proposed model adopts a lightweight extension of SEIR, termed SEIRSD, which introduces a "Dead" state to better reflect real-world outcomes without introducing excessive complexity. This decision was influenced by prior models that focused heavily on optimisation while neglecting end-state representation.

Rather than relying on parameter-heavy techniques such as PSO, this project integrates the SEIRSD logic into a network-based simulation, where individuals are represented as nodes and infection spreads through their connections. Next will explore how network-based approaches can help realism and usability in epidemic simulations without heavy computation cost.

2.2 Network-Based Simulations

Network-based simulations were initially developed for domains such as computer science and communication systems. However, they have proven it is valuable in epidemiology, particularly for modelling how individual-level social structures influence disease transmission. These simulations represent populations as nodes (individuals) and edges (social connections), allowing researchers to study the role of network density, clustering, and interaction patterns in the spread of infectious diseases.

Sony Research (2020) developed a tool called the Pandemic Simulator to visualise disease progression across networked populations. They have used colour-coded nodes to label different health states, such as susceptible, infected, or recovered, and tested various scenarios within the networks. They have found that densely connected networks often lead to faster disease spread than sparse ones, as would be expected in the real situation. This offers clear insights into how structural properties of contact networks can shape outbreak dynamics.

While the simulator provided detailed visualisations and intuitive insights, it remained largely observational and did not integrate classification or tools for evaluating outbreaks. Additionally, it was not designed to use parameter adjustments or user-defined scenarios, limiting its applicability for public education.

These findings informed the design of this project, which extends network-based simulation by incorporating classification outputs and parameter controls. By combining interactive visualisation with phase outcome, this project aims to make network-based modelling more actionable and less computationally costly, especially for non-expert users who can benefit from both visual plots and binary classifications of a pandemic.

2.3 Outbreak Classification

Outbreak classification is a critical task in epidemiology, it is important for researchers to determine whether an event remains localised or is likely to escalate into a pandemic. While SEIR models are useful for simulating disease progression, they often lack mechanisms for formally classifying severity levels. This can make simulation results less understandable, especially for non-experts. To address this, researchers have been increasingly integrating machine learning (ML) techniques into epidemiological studies, help enhance the interpretability and value of simulations.

Donthi et al. (2024) and Ankolekar et al. (2024) explored the use of machine learning classifiers, including Random Forests, Support Vector Machines (SVMs), and Neural Networks, for binary prediction between pandemic-level and non-pandemic outbreaks. These models used features such as infection growth rate, transmission speed, and hospital strain indicators to improve classification accuracy. For example, Donthi et al. reported a model accuracy of 82.7% when classifying simulated outbreaks using data from COVID-19. However, these systems often required significant preprocessing, domain expertise, and computational resources, making them challenging to implement in lightweight, user-friendly tools and proved that the 90% accuracy objectives are highly unachievable.

On the other hand, Liu (2023) focused on early detection by developing a simulation that used historical outbreak data and successfully identified signs of pandemic escalation approximately 0.4 weeks earlier than official COVID-19 alerts. While this shows the predictive potential of ML-enhanced systems, Liu’s framework also required substantial data engineering and lacked direct user interactivity.

Although these studies highlight the benefits of combining machine learning with simulation, they are primarily aimed at researchers or policy analysts. The reviewed models provided classification results in a form that is still not accessible to general users. To help, this project implements a simpler binary classification model, trained on historical pandemic signals, to provide a clear label indicating whether the current simulated scenario resembles a pandemic. If successful, this helps users make sense of simulations without needing to have prior knowledge.

2.4 Accessibility of Simulation Tools for the General Public

Simulation tools are essential for modelling infectious diseases, especially during outbreaks where quick insights can support policy and public health responses. However, as discussed in the previous sections, many existing models are designed for researchers or technical users, relying on specialised environments such as MATLAB or R. This makes them difficult for non-experts to access or understand, limiting the wider understanding and awareness of these tools.

For example, Chen et al. (2020) developed GLEAMviz, a global simulation platform used to model large-scale pandemic scenarios. While the platform offers detailed epidemiological insights using network-based structures, it is highly data-driven and complex, with limited support for interactivity or simplified interpretation. As a result, tools like GLEAMviz are very valuable for professionals as what it is designed for, but remain unavailable and difficult for the general public even if they are interested.

During public health crises, there is value in helping non-experts understand how diseases spread. While advanced tools exist for researchers, more accessible versions can play a role in public education and awareness. This project does not aim to fully bridge that gap but hopes to offer a proof of concept, a simplified, interactive tool that demonstrates how SEIR models and classification can be combined into an intuitive web-based interface.

2.5 Conclusion of the Literature Review

The reviewed studies provide valuable insights into modelling infectious disease dynamics and offer foundational frameworks for understanding pandemics like COVID-19. Researchers such as Yingze Hou (2024) and He et al. (2020) have demonstrated how modified SEIR models can improve predictive and simulation accuracy by incorporating population heterogeneity and dynamic parameters. Sony Research (2020) showed the value of network-based simulations in capturing how social structures influence epidemic spread. Meanwhile, Donthi et al. (2024), Ankolekar et al. (2024), and Liu (2023) explored the potential of machine learning techniques for outbreak classification, revealing the strength of combining data-driven prediction with simulation modelling. Tools like GLEAMviz and Epistemix (Chen et al., 2020) further illustrate how simulation platforms can support both research and public health planning on a global scale.

Despite these advancements, limitations remain. Many tools are designed for experts and require technical knowledge or high-performance computing, which makes them hardly accessible to general users. SEIR models like Hou’s often struggle with computational complexity, reducing their suitability for real-time interaction. Additionally, few of the reviewed systems offer integrated classification features to help users interpret whether an outbreak resembles a pandemic, particularly in a simplified or accessible format.

This project addresses these gaps by proposing a lightweight, local web-based simulation tool that combines a network-driven SEIRSD model with a binary classification feature. While it does not aim to replace large-scale epidemiological platforms, as said, it works as a proof of concept for how simulation and classification can be merged into an interactive, user-friendly system. By allowing users to input custom parameters and visualise outcomes in real time, the tool supports better public understanding of disease dynamics and makes the concept of outbreak risk more accessible, especially for non-expert audiences.

3. Methodology

This section outlines the development methodology and MoSCoW to establish clear artefact requirements before the project deployment phase, by setting a structured approach to ensure the project remains well-organised and achievable.

3.1 Development Methodology

The development methodology will outline how the project will be implemented and documented. As the project contains both data-driven modelling and web-based development, the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework will be employed. CRISP-DM offers a structured, iterative approach for data-driven project development, containing six key phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. However, some adjustments will be applied to better suit the needs of this project. Each phase will be explained and clarified in the coming section.

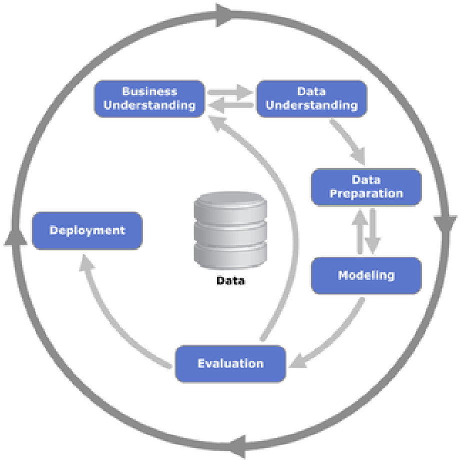


Figure 1 - CRISP-DM (Cross-Industry Standard Process for Data Mining) framework

3.1.1 Business Understanding

As the first phase of CRISP-DM, it involves defining the problem, project objectives and success criteria. The main goal is to discover important factors that could influence the outcome of the project, and set the project in the right direction. As previous sessions have already established a clear understanding of the business understanding, this phase will be simply in a later section for the purpose of not repeating.

3.1.2 Data Understanding and Data Preparation

To ensure that both the classification models accurately represent real-world disease dynamics, this project relies on reliable real-world epidemiological datasets to establish a baseline simulation model, which involves collecting and any necessary cleaning of the data. For this purpose, a well-documented and widely recognised dataset will be used, such as data from the WHO, CDC, Johns Hopkins University COVID-19 repository or other relevant studies, therefore, data understanding and data preparation will be taken together to ensure an organised approach to dataset selection and preprocessing.

3.1.3 Modeling

This phase will be focused on modeling two core models: a network–based SEIRSD simulation model for disease spread analysis and a binary classifier for outbreak classification, both models will rely on the data that will be selected in the later Data Preparation section, to ensure the system are data driven and able to generate accurate predictions and meet the project objectives.

3.1.4 Evaluation

The evaluation phase focuses on assessing the performance and accuracy of both the SEIRSD simulation model and the outbreak classification system to ensure they meet the project's objectives.

To evaluate the SEIRSD simulation, this will involve comparing the simulation outputs with real-world pandemics and how different environments influence disease transmission dynamics.

This will ensure that the model has effectively captured the variability in real-world outbreak scenarios while maintaining simplicity.

Similarly, to evaluate the effectiveness of the binary classifier, a performance calculation matrix will be used, for example, Recall (True Positive Rate), which calculates all actual positives that were correctly classified as positives, while the False Positive Rate will measure the negatives that were classified incorrectly as positives, and Preciaion will be used for the probability of a positive prediction being correct. These metrics will provide insights into the classifier's accuracy and help evaluate if there is a bias in the model.

3.1.5 Deployment

The deployment phase will focus on assembling the network–based SEIRSD simulation and outbreak classification on a website and ensuring accessibility by providing a simple yet clear UI for non-technical users to interact with the system and understand the outcome.

3.2 MoSCoW

The MoSCoW method is a well-known prioritisation technique used for managing project objectives by ranking them from high to low. It consists of four main priorities: M (Must-Have), S (Should-Have), C (Could-Have), and W (Won't-Have). This ensures the project remains manageable and not overscaled, and it helps prioritise the essential features.

ID	Priority	Feature	Descriotion
1	Must-Have	SEIR-Based Simulation	The core model for simulating disease spread uses real-world epidemiological parameters.
2	Must-Have	Outbreak Classification	A binary classifier to determine whether an outbreak is localised or pandemic-level.
3	Must-Have	Web-Based Interface	A functional UI for users to interact with the simulation and view results.

4	Must-Have	User-Defined Parameters	It allows users to set infection rates, recovery times, and population structures.
5	Must-Have	Network-Based Simulation	Implement disease transmission within a structured social network model.
6	Should-Have	Visualisations (Graphs/Charts)	Graphs to display infection trends and classification outcomes dynamically.
7	Should-Have	User-Uploaded Network Structures	Allow users to upload their contact networks for customised simulations.
8	Should-Have	Interactive Visualizations	Includes charts for daily infections, recoveries, or classification confidence scores and that update in real time when users tweak parameters.
9	Should-Have	Help/Info Section in UI	Simple tooltips or a “?” button explaining input parameters and results.
10	Should-Have	Pre-Defined Scenarios	Let users switch between sample outbreak scenarios, for example, COVID-19 in 2020 and the Flu season baseline.
11	Could-Have	Tutorial Mode or Walkthrough	A guide that introduces the interface step-by-step for users who have never used such a system before.
12	Could-Have	Localised Language Support	Implement different languages for better usability.
13	Won’t-Have	Server deployment	A functional website for users to run the simulation on the internet.
14	Won’t-Have	User Accounts	Optional login to save users scenarios or simulations for later

15	Won’t-Have	Extra Classification Categories	Other than pandemic or not pandemic, add on different stages such as local outbreak, epidemic and pandemic.
16	Won’t-Have	High-Level System	Contingency plan suggestion and economic simulation

Table 3 - MoSCoW Table

4. Implementation

This section follows the modified CRISP-DM framework as discussed in the earlier methodology section. It details the practical steps to build and validate the simulation system in each phase.

4.1 Business Understanding

The goal of this project is to develop a web-based simulation tool that simulates the spread of infectious diseases using a modified SEIR (SEIRSD) model and applies a binary classifier to identify outbreak severity to help better understand the simulated result. The system is designed to be interpretable and accessible to non-technical users, allowing for real-time simulation and customisation of disease dynamics.

4.2 Data Understanding and Preparation

A suitable dataset is essential for developing a binary classification model that supports outbreak severity identification. For this purpose, the Johns Hopkins University (JHU) COVID-19 global dataset was selected. This dataset is widely adopted and reputable in epidemiological research, providing time-series records of confirmed cases, deaths, recoveries and geographic information such as lat (north or south) and long (east or west), across countries and regions between 22 January 2020 to 9 March 2023.

This dataset provided a solid basis for the classification model to predict whether a simulated outbreak resembles a pandemic. However, for this project network-based simulation component does not directly rely on real-world epidemiological data. Instead, the simulation network will be generated using synthetic structures designed to approximate real-world contact patterns, specifically employing the Watts–Strogatz small-world model. This allows flexible, dynamic network creation suitable for visualisation and interactive simulation.

4.2.1 Data Cleaning and Aggregation.

For the classification model, the confirmed, deaths, and recovered datasets were loaded and merged into a single DataFrame for efficiency. The combined dataset initially contained 306,324 entries, with some missing values in the Lat and Long columns. Since the geographic information was not needed for the simulation, these fields were dropped without further impact.

4.2.2 Feature Engineering and Labelling Strategy

To ensure a reliable outbreak classification model, a set of engineered features was derived from the preprocessed dataset. These features were designed to capture different aspects of epidemic dynamics, such as case growth volatility, death-to-case ratios, and overall progression trends. The feature design was informed by public health frameworks such as the Pandemic Severity Assessment Framework (PSAF) introduced by the U.S. Centres for Disease Control and Prevention (CDC). For more features information see the table below.

ID	Features	Descriotion
1	Confirmed	Cumulative total of confirmed cases.
2	NewCases	Daily difference in confirmed cases.
3	StdDev_NewCases	7-day rolling standard deviation of new cases.
4	Smoothed_NewCases	7-day rolling average of new cases.

5	Cumulative_Smoothed_NewCases	Cumulative sum of smoothed new cases.
6	Deaths	Cumulative total of deaths.
7	Daily_Deaths	Estimated daily deaths, based on a fixed ratio.
8	Smoothed_Daily_Deaths	7-day rolling average of estimated daily deaths.
9	Smoothed_DeathRate	Ratio of smoothed deaths to smoothed new cases.
10	Cumulative_Smoothed_Deaths	Cumulative sum of smoothed estimated deaths.
11	GrowthRate	Percentage increase in smoothed new cases.
12	GrowthRate_Volatility	Standard deviation of growth rate over a 7-day window.

Table 4 - Features Engineering Table

Rather than assigning pandemic labels based on official announcements, for example, the WHO's declaration on 11 March 2020, this project defined outbreak severity using a data-driven threshold. By using a Growth Rate Threshold was computed as the 75th percentile of the smoothed growth rate within the training period. Each day in the dataset was labelled as either pandemic (1) or non-pandemic (0) depending on whether it exceeded this threshold.

This method expanded the number of class 1 samples without applying synthetic methods like SMOTE, preserving the historical data's real-world temporal structure. It also ensured that labels reflected generalised outbreak dynamics rather than a single fixed date, improving the model's robustness in unseen simulations.

Following feature engineering and labelling, the dataset was ready for model training. Additional preprocessing steps, such as scaling and imputation, were performed during the modelling stage and are detailed in Section 4.3.

4.3 Modeling

Following the Data Understanding and Preparation phase, this section outlines the process used to build the final binary classification model and the modification to the Watts–Strogatz small-world network-based SEIRSD simulation system. Details regarding evaluation and results, and unsuccessful experiments will be discussed separately in the Evaluation section.

4.3.1 Binary Classification Model

To enable binary classification between pandemic and non-pandemic scenarios, the preprocessed dataset was first split using a time-based split method, as 80% of the earlier data was used for training, and the remaining 20% of the later data was reserved for testing. The benefit of this approach preserves the temporal sequence of events, ensuring that the model learns from past patterns without leaking future information into training. Compared to a random split, this method provides a more realistic evaluation for time series data, to truly reflect how the model would perform in predicting future outbreaks.

To further improve the model, missing value imputation and feature scaling were applied. These preprocessing steps were only applied to the training set and then applied to the test set to prevent any risk of data leakage. After testing different approaches, a K-Nearest Neighbours (KNN) imputer was chosen for its highest performance and to fill up missing values, while a StandardScaler was used to normalise the feature ranges and reduce the impact of outliers.

After preprocessing, several well-known classification models were trained and compared. The top three best-performing models were Random Forest, K-Nearest Neighbours (KNN), and Gaussian Naive Bayes. Among them, Gaussian Naive Bayes achieved the highest test accuracy of 65%, even though it failed to achieve 90%, it was chosen as the final classification model.

The final trained model is later integrated into the simulation pipeline to provide real-time classification of outbreak severity as the network evolves.

4.3.2 SEIRSD Simulation Engine

To better simulate the dynamics of disease spread more realistically within the network, the standard SEIR model was further extended and formed a SEIRSD (Susceptible → Exposed → Infected → Recovered → Susceptible → Deceased) framework. This includes both waning immunity (R → S) and mortality (I → D). allowing simulation of fatal outcomes, and reintroducing recovered individuals back into the susceptible population. These changes have created a more dynamic life mechanics of infectious disease dynamics compared to traditional SEIR structures.

The disease progression is modelled using a system of differential equations. These equations manage the rates of change between states over time. Infection spreads from susceptible to exposed individuals through contact with infectious individuals, and that individuals will transition to the infected state. Infected individuals either recover or die by chance, set by parameters, and recovered individuals can eventually lose immunity and return to the susceptible state. The system of equations is as shown below:

$$\frac{dS}{dt} = -\beta\frac{S\,I}{N} + \rho R$$

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

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

$$\frac{dR}{dt} = \gamma I - \rho R$$

$$\frac{dD}{dt} = \mu I$$

Symbol	Description
S	Number of susceptible individuals
E	Number of exposed individuals
I	Number of infectious individuals
R	Number of recovered individuals
D	Number of deceased individuals
N	Total population
β (beta)	Transmission rate
σ (sigma)	Transition rate from exposed to infected (1 / incubation period)
γ (gamma)	Recovery rate
μ (mu)	Mortality rate (death rate for infected)
ρ (rho)	Rate of loss of immunity (recovered to susceptible)

Table 5 - Equation Explanation Table

4.3.3 Implementation in Network Structure

To move further from traditional simulation, the SEIRSD model was implemented within a network-based structure using a Watts–Strogatz small-world graph. Each node in the graph represents an individual, and edges define their potential social interactions, through which infections may spread.

The infection logic is implemented using custom functions and applied iteratively across the network. The simulation supports control of parameter values before the simulation starts and time-step progression through a web interface. This allows users to understand how changes to transmission or recovery rates affect the dynamics of the virus over time. In addition to SEIRSD logic, the network supports: Node reintroduction, simulating new individuals joining the population, Edge rewiring, to model changes in social structure and timeline synchronisation, allowing SEIRSD states and network dynamics to animate in parallel.

Together, this architecture has successfully transformed the SEIRSD equations into a visually driven, dynamic system that recreates real-world dynamics while maintaining user interactivity and control.

5. Evaluation

5.1 Binary Classifier Evaluation

The final binary classification model was the Gaussian Naive Bayes (GNB). After time-based data splitting, preprocessing with KNN imputation, and standard scaling, GNB achieved a test accuracy of 65% on the unseen test set. However, a set of experiences has been taken to achieve this accuracy, even though it is not ideal.

5.1.1 Process Evaluation

The dataset was processed naively in the early stages of classifier modelling, and the initial feature engineering was overly simplistic. As a result, substantial data leakage happened, leading to extremely high and misleading accuracy scores across many models and dataset variations. Models were consistently achieving close to 100% accuracy, which immediately indicates their reliability and generalisability are not correct.

After further examination, three issues were discovered. First, although the dataset was split correctly, it used a random split during the process rather than a time-based approach. This allowed information from future events to leak into the training set, enabling the models to "cheat" by learning patterns if that they would not realistically have access to in real-world conditions. Second, the initial features are not able to represent pandemic stage dynamics meaningfully, which has limited the model's ability to generalise to new scenarios. Lastly, the labelling Strategy was overly simple by differentiating the class only on the World Health

Organisation (WHO) Announcement, where it is reasonable; however, for a time split, it caused a limited number of non-pandemic data, therefore, Growth Rate Thresholds were later applied to create more non-pandemic data while keeping it data-driven.

To correct these issues, the dataset was reprocessed by applying a time-based split, careful feature engineering and a labelling Strategy that captured more realistic pandemic dynamics, as discussed in Section 4.3.1. These changes reduced data leakage, forced the model to learn only from past patterns, and increased the reliability of performance. Even though the final accuracy of 65% was lower than the overfitted models, it reflects a much more realistic and trustworthy model suitable for real-world generalisation.

5.1.2 Labelling Strategy Reflection

Labelling the dataset accurately was an important step to build a reliable classifier in any related task, as discussed in the earlier section. After exploration, this project has used the 75th percentile threshold of the smoothed growth rate that was created during feature engineering. It has some benefits; however, it could be the root cause of the 65% accuracy.

First, this approach has avoided using oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE), preserving the real-world temporal structure of the outbreak data, meanwhile to minimise the class imbalance issue under the binary classification approach. It allowed for a more detailed separation of pandemic and non-pandemic phases based on the outbreak behaviour, and it improved dynamism in the labelling process, so that days showing unusually high growth rates before formal pandemic recognition were labelled appropriately.

However, this approach also has important limitations. The fixed 75th percentile threshold, even though it is useful for COVID-19 data, it may not generalise to future outbreaks that have different growth dynamics, and the threshold was totally derived from the historical COVID-19 data, meaning the model is most likely overfit to that specific outbreak profile and may struggle with different epidemiological patterns.

5.1.3 Model Evaluation

The final chosen model was Gaussian Naive Bayes (GNB), which was evaluated on the unseen hold-out test set. After time-based splitting, KNN imputation, and standard scaling, GNB achieved a test accuracy of 65%, which has failed to achieve the target of 90% as noted unrealistic, however, it has been able to represent a more realistic generalisation ability compared to the earlier overfitted models.

Detailed test set performance is summarised below:

Model	Matrix	Performance
Gaussian Naive Bayes (GNB)	Accuracy	65%
	Precision (Class 0: Non-Pandemic)	87%
	Recall (Class 0)	67%
	Precision (Class 1: Pandemic)	31%
	Recall (Class 1)	59%
	Macro-Averaged F1 Score	58%
	Weighted F1 Score	68%

Table 6 - Gaussian Naive Bayes (GNB) Result Table

While precision for detecting the pandemic class (1) was relatively low (31%), the recall (59%) indicated that the model was able to detect a majority of true pandemic periods.

The binary classifier, although methodologically sound it still does not perform. This is likely due to a combination of factors, including the noisy labelling process, the limited range of available features, and the inherent difficulty of distinguishing pandemic escalation purely from growth trends in the early stages.

5.2 SEIRSD Simulation Evaluation

5.2.1 Core Simulation Behaviour

The SEIRSD simulation model is an extension of the classical SEIR framework, introducing two additional transitions: waning immunity (Recovered → Susceptible) and mortality (Infected → Deceased). These additional elements allow the simulation to capture more realistic long-term dynamics, such as reinfection cycles and death, although with some limitations, particularly regarding recovery rate (γ) calibration due to population size constraints.

To verify the plausibility of the SEIRSD structure, the simulation was run under standard COVID-19-like parameter settings, and the resulting epidemic curves were compared against historical pandemic patterns. As shown in the plot below.

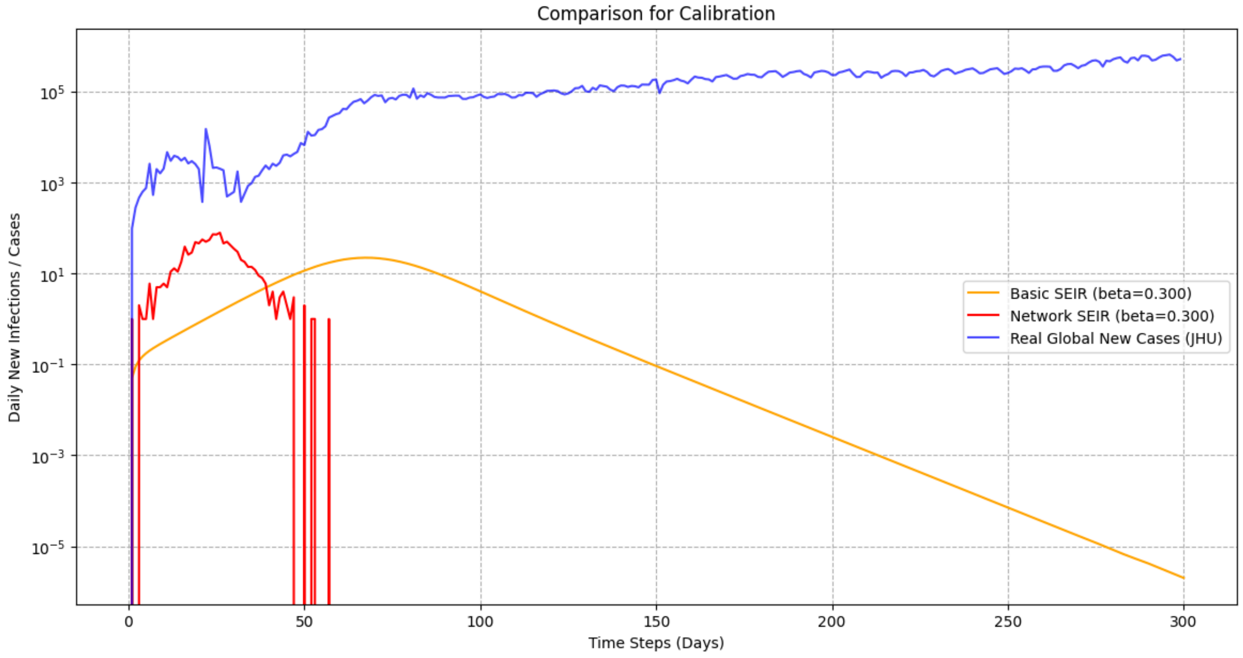


Figure 2 - Basic SEIR vs Network SEIR

As shown in Figure 2, the Basic SEIR without network was smooth and unrealistic compared to the real global data, while the Network SEIR were able to show a more realistic pattern; however, as the structure of SEIR, the infected population once recovered, will be the end of its cycle, which causes no new infected as shown on the drop of the plot, therefore a new state that allow recovered population turn to susceptible is needed for more realistic result.

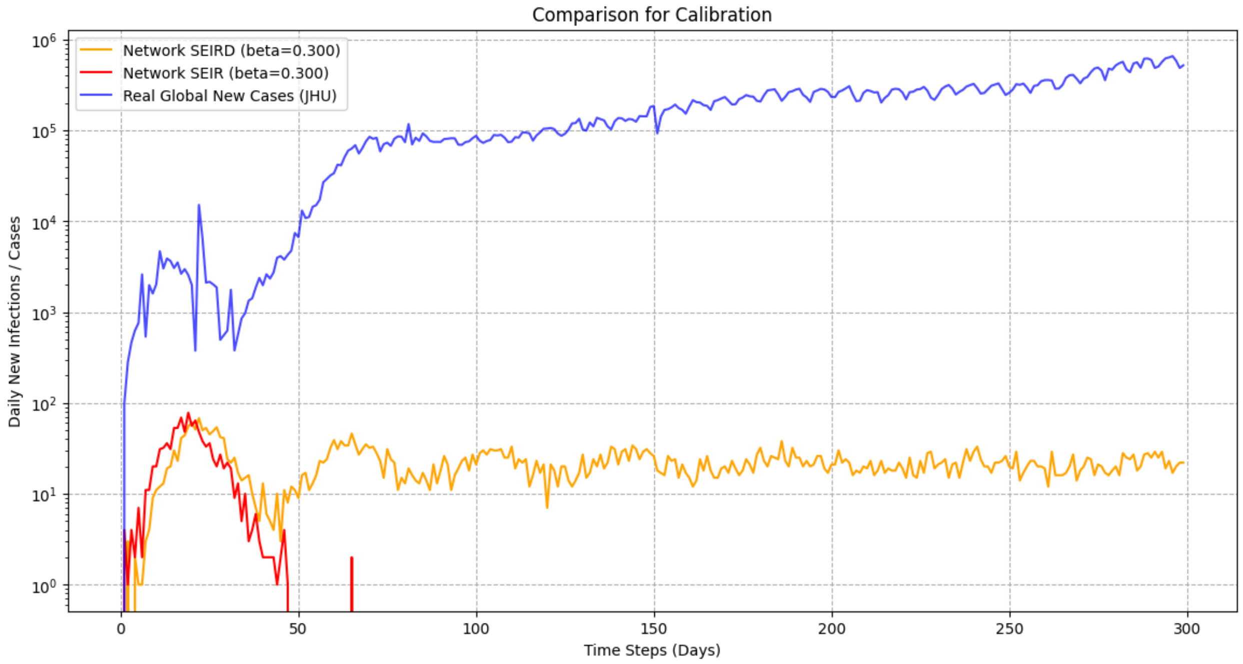


Figure 3 - Network SEIR vs Network SEIRSD

The state for recovered turn to susceptible is implemented as shown in Figure 3, now it can show a consistent infection over time; however, for this result, the immunity duration is set to 25, which means the infected population will take 25 days to turn recovered to susceptible, where for covid it is 150 days, however because of the size of the population it is not suitable as it will die out before populaion turn back to susceptible again.

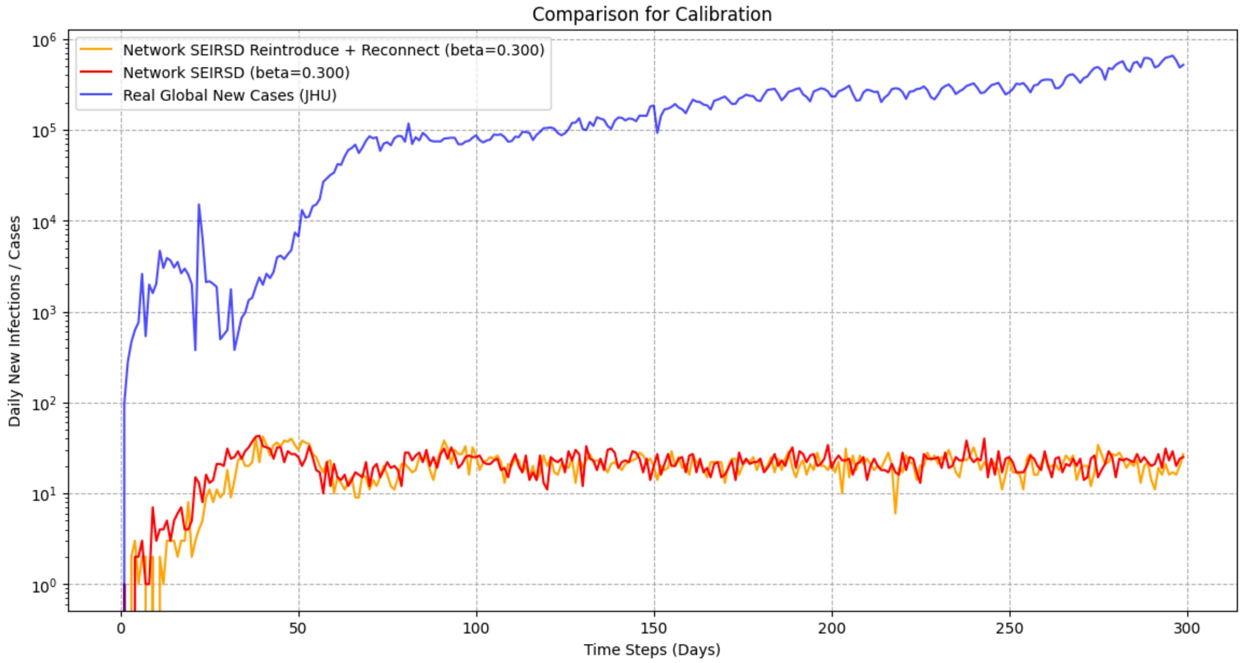


Figure 4 - Network SEIRSD vs Network SEIRSD Reintroduce + Reconnect

The SEIRSD implemented with the Reintroduce and Reconnect nodes model has successfully reproduced key epidemic phases, including exponential growth, peak infection periods, and gradual decline, supporting its validity for this project's objectives. It is very similar to the normal Network SEIRSD; it is still chosen despite the similarity, while adding the computation cost, as it has built a stronger foundation for future implementation.

Now, the network structure evolved dynamically through periodic node reintroduction and edge rewiring, simulating population and social connectivity changes over time. This helped the simulation approximate real-world dynamics. Overall, the SEIRSD were validated to ensure it correctly followed the logic. Transmission, Incubation, Recovery, Mortality, and Immunity loss were simulated randomly across the dynamic network, providing realistic infection dynamics over time.

5.2.2 Limitations

While the SEIRSD simulation model has successfully recreated reasonable epidemic dynamics, several limitations remain. First, as mentioned, the parameter settings were limited, particularly the recovery rate (γ), which was simplified into a limited size throughout the test simulation. In reality, these rates would change over time and much higher for such as 60 to 120 days, because of factors such as medical advancements, public health interventions, or virus mutations, moreover, this project has not taken into account of any policy changes.

Second, the small-world network structure, although appropriate for modelling social contacts, it does not explicitly model spatial heterogeneity, demographic differences, or real-world mobility patterns. As a result, some complexity of real pandemics, such as local outbreaks or region-specific effects, will not be captured under this setting.

Lastly, it has limitations to the network sizes of 1,000 to 2,000 nodes for web deployment, limiting the resolution and granularity of simulated epidemics. While sufficient for proof-of-concept purposes, scaling the model to city or country-level populations would require more optimisation or different architectural choices. Despite these constraints, the SEIRSD engine has met its primary goal of providing an intuitive and flexible simulation foundation for public-facing pandemic exploration.

5.3 Failed or Abandoned Experiments

Throughout development, several alternative approaches were explored but ultimately not adopted due to technical, practical, or performance-related reasons.

5.3.1 Multi-Class Labelling

Early experiments attempted to label outbreak phases using a role-based strategy into 4 categories, such as pre-pandemic, escalation, pandemic and decline. If it has worked, it will largely help in class imbalance and provide more detailed insight than a binary result. However, clear phase boundaries were difficult to define based solely on available features like growth rate, case counts and being visually correct. Models trained on this multi-class structure struggled with class confusion and had unstable results. the complexity and limited time, the final design reverted to binary classification.

5.3.2 Clustering Approach

Another labelling strategy was explored during the project's early stage, which was unsupervised clustering using K-Means. The objective was to automatically identify pandemic phases without relying on fixed thresholds as used in the binary class. The Elbow method has suggested an optimal k of 4 clusters, which means 4 classes. However, the result showed that the clusters did not correspond meaningfully to known pandemic phases. A classifier trained to predict these cluster labels achieved very poor accuracy at 0%. It can improve, however, because of time limitations. No further clustering methods, such as DBSCAN, were tested. It was concluded that manually engineered labels based on growth rate dynamics remained a more appropriate and reliable approach under these circumstances.

5.3.3 Other Networks

Alternative network models, such as scale-free networks and Holme-Kim clustered graphs, were considered to replace the small-world Watts–Strogatz model. However, simulations show that these methods either produced unrealistic epidemic spread patterns or added complexity without significant gains in visual clarity. Because of the project more focus on accessibility and intuitive learning, the Watts–Strogatz structure was kept as the most balanced choice.

5.4 Summary of Evaluation

The evaluation phase revealed important insights into the strengths and limitations of the system developed. The binary classification model, despite having only achieved 65% accuracy, has successfully demonstrated realistic generalisation without relying on overfitting or artificial data generation. The labelling strategy based on outbreak growth dynamics provided a flexible and data-driven foundation, but also introduced challenges when trying to achieve higher accuracy and predict more complex real-world escalation patterns.

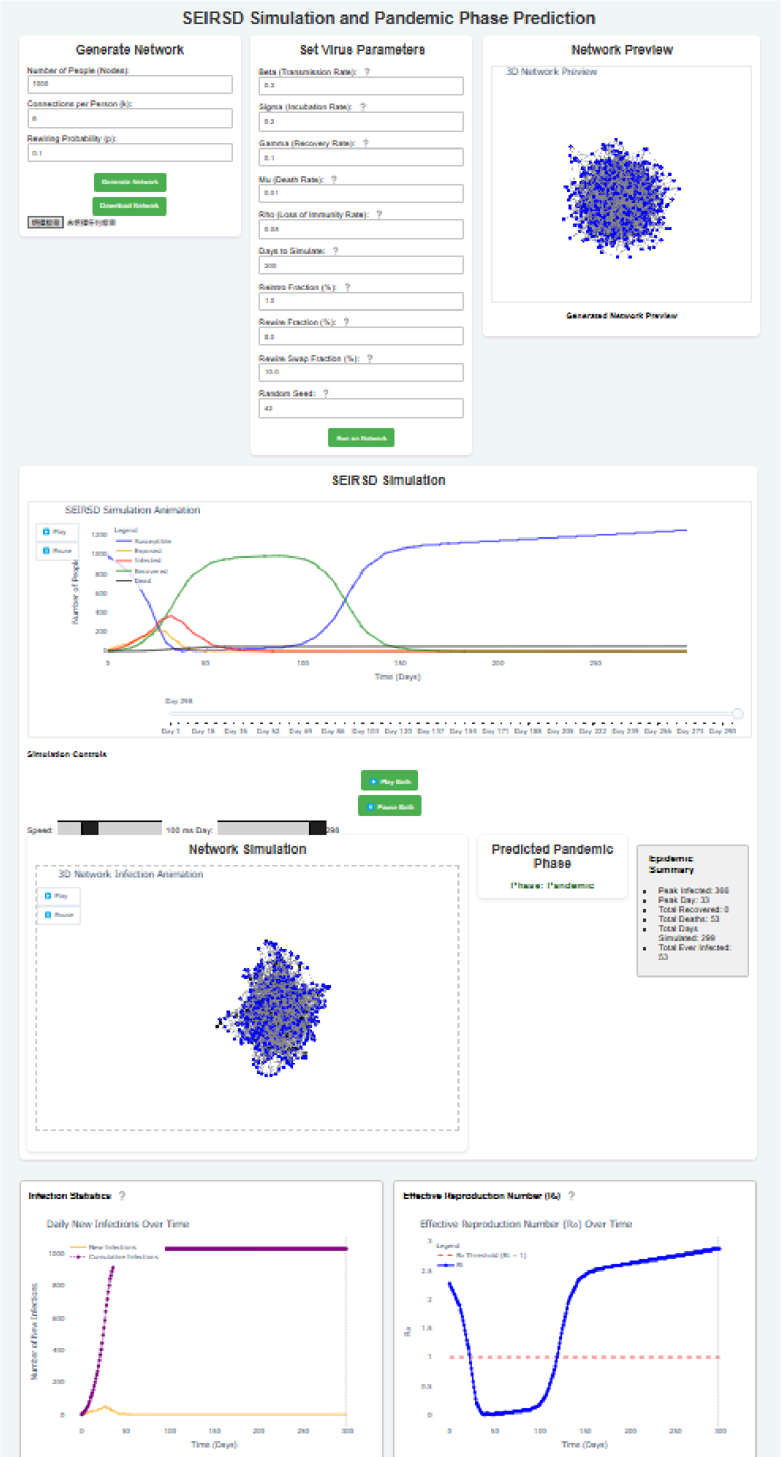
The SEIRSD simulation engine effectively captured realistic epidemic dynamics, including reinfection and mortality cycles, while maintaining computational efficiency suitable for real-time visualisation. However, constraints in parameter tuning and the simplified network infection logic limited its biological realism compared to full agent-based models.

Failed experiments, such as multi-class classification, clustering, and alternative network generation, showed the complexity of pandemic dynamics and the difficulty of achieving strong predictive performance without extremely fine-tuned or expert-labelled datasets. These outcomes reinforced the value of keeping the project focused on a lightweight, interpretable proof-of-concept rather than pursuing complexity at the cost of usability and difficulty.

Overall, the project successfully achieved its core target and objective as listed in section 1.1, creating a web-accessible simulation with integrated pandemic classification, while critically engaging with the trade-offs between model simplicity, usability, scientific depth and difficulties.

6. Deployment

Following development and evaluation, the final system was deployed as a local web application using FastAPI as the backend server and Plotly.js for interactive frontend visualisation. The website UI is displayed below:



SEIRSD Simulation and Pandemic Phase Prediction

Generate Network

Number of People (Nodes): 1000

Connections per Person (k): 6

Rewiring Probability (p): 0.1

Set Virus Parameters

Beta (Transmission Rate): 0.3

Sigma (Incubation Rate): 0.3

Gamma (Recovery Rate): 0.1

Mu (Death Rate): 0.01

Rho (Loss of Immunity Rate): 0.05

Days to Simulate: 300

Reintro Fraction (%): 1.0

Rewire Fraction (%): 5.0

Rewire Swap Fraction (%): 10.0

Random Seed: 42

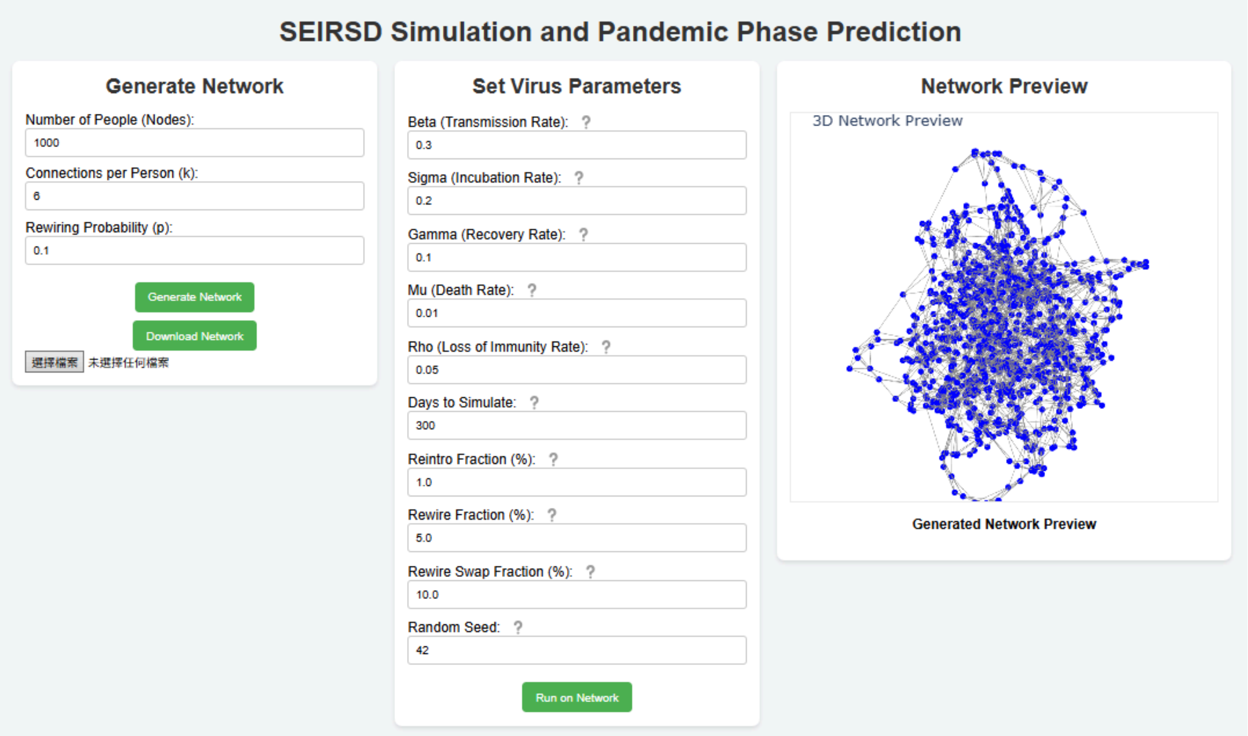
Predicted Pandemic Phase

Phase: Pandemic

Epidemic Summary

- Peak Infected: 366
- Peak Day: 33
- Total Recovered: 0
- Total Deaths: 53
- Total Days Simulated: 299
- Total Ever Infected: 53

Figure 5 - User Interface



SEIRSD Simulation and Pandemic Phase Prediction

Generate Network

Number of People (Nodes): 1000

Connections per Person (k): 6

Rewiring Probability (p): 0.1

Set Virus Parameters

Beta (Transmission Rate): 0.3

Sigma (Incubation Rate): 0.2

Gamma (Recovery Rate): 0.1

Mu (Death Rate): 0.01

Rho (Loss of Immunity Rate): 0.05

Days to Simulate: 300

Reintro Fraction (%): 1.0

Rewire Fraction (%): 5.0

Rewire Swap Fraction (%): 10.0

Random Seed: 42

Network Preview

Figure 6 - Parameter Settings

The simulation engine and classification model have been successfully integrated into the web framework to allow user input and live updating of epidemic curves and network graphs that are provided on the website. The application was designed in a modular section for better understanding and is responsive across standard desktop browsers.

In the top section, Users are allowed to generate or upload a network in JSON format, and a virus with parameters shown table below:

Network Settings	Parameters	Uses example	Explanation
	Number of People (Nodes)	1000	1000 people will be inside the simulation
	Connections per Person (k)	6	Each person will connect to the other 6 people as edges
	Rewiring Probability (p)	0.1	The chance to link to

			other nodes that were far away.

Table 7 - Network Parameters Setting Table

Virus Settings	Parameters	Uses example	Explanation
	Beta (Transmission Rate)	0.3	Each infected person has a 30% chance per day of infecting other people who are linked.
	Sigma (Incubation Rate)	0.2	The exposure period is around 2 days before turning infected.
	Gamma (Recovery Rate)	0.1	The infectious period is about 10 days.
	Mu (Death Rate)	0.03	3% chance per day that an infected person dies.
	Rho (Loss of Immunity Rate)	0.05	5% chance per day to lose immunity and return to susceptible.
	Days to Simulate	300	The simulation will run for 300 days
	Advanced Parameters	**Uses example**	**Explanation**
	Reintroduced Fraction (%)	1.0	1% of the individuals will be reintroduced per week
	Rewire Fraction (%)	5.0	5% of the alive nodes will be selected for *rewiring*.
	Rewire Swap Fraction (%)	10.0	10% of the old connection will be dropped, and a new connection will be made for the selected node

Table 8 - Virus Parameters Setting Table

Simulation output is visualised both through traditional SEIRD element plots and through dynamic 3D network animations that reflect the current state of each node.

Deployment was performed only locally for testing and demonstration purposes. The architecture was kept simple using only standard open-source libraries without external cloud services. Limitations of the deployment include the absence of user accounts, lack of server-side data storage for sessions, and performance constraints when scaling to very large network sizes over 2000 nodes, primarily due to the cost of computing performance. See the table below:

Number of Nodes (under the same network setting) 6 Connections per Person (k) and 0.1 (Rewiring Probability)	Major Steps	Time spent (seconds)
500 Nodes	SEIRSD Plot	1.94
	Network Animation	45.51
	Network Structure	2.47
	Total Spend	**63.03 Seconds**
1000 Nodes	SEIRSD Plot	1.88
	Network Animation	95.21
	Network Structure	6.7
	Total Spend	**123.12 Seconds**
2000 Nodes	SEIRSD Plot	1.80
	Network Animation	180.71
	Network Structure	20.96
	Total Spend	**251.12 seconds**

Table 9 - Network Performance Table

The performance results followed an expected trend: doubling the number of nodes roughly doubled the overall simulation and plotting time. While the runtime remains acceptable for a small-scale experiment and proof of concept, it may limit the system's scalability for very large networks.

Nonetheless, the deployed system successfully demonstrates all intended functionalities: allowing users to simulate infectious disease spread, adjust parameters interactively, and receive automated binary classification outputs regarding outbreak severity.

It serves as proof of concept for how a more complex, public-facing epidemic modelling platform could be built in the future.

7. Discussion and Conclusion

7.1 Conclusion

This project successfully developed a web-based simulation tool that integrates a network-based SEIRSD disease model with a binary classification system to assess the outbreak phase. Although the final classifier only achieved an accuracy of 65% and failed to reach the target 90%, it was at least methodologically robust, with careful attention to data leakage, feature engineering, and realistic time-based validation.

The SEIRSD simulation engine is simplified but still able to capture key dynamics of infectious disease spread, including reinfection and death, and also allows users to explore different pandemic scenarios interactively. By integrating these components into an accessible web interface, the project shows that real-time epidemic simulation and outbreak classification can be made more understandable for non-expert audiences without relying on external APIs or high-performance cloud infrastructure.

7.2 Potential Improvements

While the project achieved its core objectives, multiple areas for future improvement were found. First, the binary classification system could be enhanced through better feature engineering or by adopting a more dynamic labelling strategy that adjusts based on contextual outbreak behaviour, rather than relying on a fixed growth rate threshold derived from historical COVID-19 patterns.

Second, the SEIRSD model, although effective as a proof of concept, remains a simplified version of epidemic processes. Future work could explore incorporating additional factors such as vaccination rates, variable contact rates by age group, healthcare system capacity, and the evolution of the virus will be able to better reflect real-world dynamics.

Third, network generation in the current system is based solely on a synthetic Watts–Strogatz small-world structure. While this choice helped approximate realistic contact patterns, allowing users to select between multiple network types for scale-free networks, clustered random networks could offer richer scenario exploration; however, it is out of scope for the project at this stage.

Fourth, more features can be implemented to help better understand the result. As the classification is now able to provide a binary result, however, that is a very limited understanding. To improve a smell-scale, LLMs (Language Models) can be integrated into the system for better interpretation.

Lastly, while the application runs acceptably for small networks, performance limitations become more apparent with very large node network counts above 2000. Future deployment could consider optimising the animation pipeline or simplifying visualisation for larger graphs to support scalability.

Word count: 5871

8. References

Ankolekar, A., Eppings, L., Bottari, F., Pinho, I.F., Howard, K., Baker, R., Nan, Y., Xing, X., Walsh, S.L.F., Vos, W., Yang, G. and Lambin, P., 2024. Using artificial intelligence and predictive modelling to enable learning healthcare systems (LHS) for pandemic preparedness. *Computational and Structural Biotechnology Journal*, 24, pp.412–419. Available from: https://doi.org/10.1016/j.csbj.2024.05.014 [Accessed 26 February 2025].

Bonten, M.J.M., van de Wijgert, J.H.H.M., et al., 2020. Modeling the effects of interventions on COVID-19 transmission. *PLoS Computational Biology*, 16(7), e1008034. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7365652/ [Accessed 19 February 2025].

Chen, T., Li, X., Wang, Y., et al., 2020. GLEAMviz: A global epidemic simulator for modeling infectious diseases. *Journal of Computational Science*, 45, 101215. Available from: https://doi.org/10.1016/j.jocs.2020.101215 [Accessed 26 February 2025].

Donthi, S., Kumar, R., and Patel, A., 2024. Machine learning for outbreak classification: Predicting pandemic potential using epidemiological data. *Epidemiology and Infection*, 152, e45. Available from: https://doi.org/10.1017/S0950268824000321 [Accessed 26 February 2025].

Dyvik, E.H., 2024. Impact of the coronavirus pandemic on the global economy - Statistics & Facts. Available from: https://www.statista.com/topics/6139/covid-19-impact-on-the-global-economy/ [Accessed 26 February 2025].

He, S., Peng, Y. and Sun, K., 2020. SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics*, 101(3), pp.1667–1680. Available from: https://doi.org/10.1007/s11071-020-05743-y [Accessed 26 February 2025].

Hethcote, H.W., 2000. The mathematics of infectious diseases. *SIAM Review*, 42(4), pp.599–653. Available from: https://doi.org/10.1137/S0036144500371907 [Accessed 26 February 2025].

Holme, P. and Kim, B.J., 2002. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2), 026107. Available from: https://doi.org/10.1103/PhysRevE.65.026107 [Accessed 26 February 2025].

Hou, Y., 2024. Multi-feature SEIR model for epidemic analysis and vaccine prioritization. *Journal of Medical Virology*, 96(3), e29452. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10906911/ [Accessed 5 March 2025].

International Labour Organization (ILO), 2020. ILO Monitor: COVID-19 and the world of work. Available from: https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@dcomm/documents/briefingnote/wcms_740877.pdf [Accessed 26 February 2025].

Johns Hopkins University (JHU), 2025. Novel Coronavirus (COVID-19) Cases Data. Available from: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases [Accessed 14 March 2025].

Kermack, W.O. and McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772), pp.700–721. Available from: https://doi.org/10.1098/rspa.1927.0118 [Accessed 26 February 2025].

OpenAI, 2024. San Francisco: OpenAI. Available from: https://chat.openai.com/ [Accessed 26 February 2025].

Google, 2024. Mountain View: Google. Available from: https://gemini.google.com/ [Accessed 26 February 2025].

Liu, A.B., Lee, D., Jalihal, A.P., Hanage, W.P. and Springer, M., 2023. Quantitatively assessing early detection strategies for mitigating COVID-19 and future pandemics. *Nature Communications*, 14(1), pp.1–10. Available from: https://www.nature.com/articles/s41467-023-44199-7 [Accessed 26 February 2025].

Sony Research, 2020. Pandemic Simulator. Available from: https://github.com/SonyResearch/PandemicSimulator [Accessed 26 February 2025].

Wellcome, 2021. The Covid-19 effects on societies and economies. Available from: https://wellcome.org/news/equality-global-poverty-how-covid-19-affecting-societies-and-economies [Accessed 26 February 2025].