International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

AI-Driven Predictive Analytics Framework for Early Detection and Forecasting of Infectious Disease Outbreaks Using Hybrid Machine Learning Models

Pulimetla Vikas [*1]
M.Tech Student, Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
vikaspulimetla@gmail.com

D. D. D. Suri Babu [*2]
Asst Prof,Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
suribabu.ddd@ggu.edu.in

D.Phani Kumar [*3]
Asst Prof,Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
phanikumar@ggu.edu.in

Abstract— Due to globalization, climate variability, urbanization, and population mobility, infectious disease outbreaks have become frequent and more severe than ever in recent years. Traditional disease surveillance mechanisms are based on slow manual reporting and one-sided epidemiological signals, which would be reactive and do not detect outbreaks early. The given paper introduces an AI-guided predictive analytics platform aimed at detecting and predicting infectious diseases outbreaks earlier by merging diverse data sources and using advanced machine learning solutions. The methodology of the proposed system is to utilize a combination of historical epidemological data, climate variables, data regarding population movement, and text-based information provided by social media to capture complicated spatio-temporal disease patterns. Complete preprocessing pipeline is used, encompassing data cleaning, normalization, balancing of classes and analysis of unstructured text uses Natural Language Processing (NLP). Several machine learning and deep learning models are utilized, such as Random Forest, Decision Tree, and Neural Network architectures, all trained and tested to estimate a risk rating of the outbreak and types of the disease. It has been proved experimentally that the Neural Network model is the best predictor, with the highest accuracy, precision, recall, and F1-score compared with traditional classifiers. By deploying the trained models under a Flask-powered web application, real-time outbreak prediction, confidence scores, and interactive visualization dashboards are made usable in decision making by the application. The suggested framework allows proactive, population health intervention, effective resource distribution, and better epidemic preparedness. The work continues to show the predictive efficacy of AI-based predictive analytics in enhancing modern disease surveillance mechanisms and in informing the decisions of a data-driven public health.

Keywords— **Neural Networks, Natural Language Processing, Random Forest, Decision Trees, Predicitive Analysis, Disease Outbreaks.**

I. INTRODUCTION

Start-up companies are crucial in economic development, technological progress, and job creation throughout the world. In more recent years, start-ups have become a major trend of generating new business models and disruptive technologies; nonetheless, a significant percentage of start-ups fail within the first few years of operation. High levels of failures create significant challenges to entrepreneurs, investors, incubators, and policy-makers, leading to financial loss; they are inefficient in resource allocation. In order to counter these risks and allow informed decision-making, one of them became the early and correct prediction of start-up success. Traditional measures of start-up viability usually utilize expert judgment, financial ratio analysis, and pre-determined business measures. These approaches tend to be subjective, slow, and lack the capacity to describe multifactorial interactions in shaping start-up results. Besides, manual evaluation methods fail to scale with the increasing mass of start-up data.

The growing access to organized business information (funding history, milestones reached, the relationships between founders, company features, and participants) opens up new possibilities of using machine learning methods to automatize and refine start-up success prediction. The latest breakthroughs in machine learning have proven to be very promising in their ability to analyze extensive datasets and discover latent correlations among various business characteristics. The most common algorithms used in business analytics to perform tasks like classification and prediction have included support vector machine, Logistic Regression, K-Nearest Neighbors, Naïve Bayes, and Random Forest since they are robust and scalable.

In this paper, an effective machine learning-based model is proposed to predict start-up failure or success based on structured business data acquired through a publicly accessible start-up dataset. The presented system measures the performance of several machine learning models, controls class imbalance via the ADASYN method, and chooses the best classifier according to traditional performance metrics. The most successful of the models is implemented via Flask-based web app to facilitate live prediction. The suggested method encourages data-oriented decision-making and leads to intelligent and scalable start-up assessment. The rest of this paper presents related work, system methodology, experimental results, and conclusions.

Scope

This article is devoted to a theoretical illustration of an AI-based predictive analytics model to identify and predict infectious disease outbreaks at an early stage by utilizing hybrid machine learning frameworks. The research approach involves the combination of heterogeneous data sources such as historical epidemiology records, climatic variables, population mobility, and unstructured social media data to generate the patterns of disease transmission at a fine spatio-temporal level. The precision of data processing in the

study includes data cleaning, data normalization, the raw data address class imbalance, and Natural Language Processing (NLP) algorithms to derive meaningful text features. The framework not only analyzes various models of machine learning and deep learning to measure the degree of risk of outbreak and type of disease but also applies the optimized predictive model to a Flask-based web application to provide real-time feedback on the risk prediction of outbreaks and confidence scoring, as well as interactive visualization control charts. The suggested framework is designed to facilitate proactive population health decision-making, epidemic preparedness, planning, and the effective allocation of resources in the population health management systems.

Objectives

To realize the above goal, the study first defines a multi-source data integration model, which integrates epidemiological, environmental, mobility, and social media data, at the early stage of infection diseases outbreaks. Another main objective is to create and compare different predictive model structures, such as Decision Tree, Random Forest, and Neural Network architects, to retrieve the best model in terms of predicting outbreak risks and classifications of certain diseases, based on unstructured textual information. Hyperparameter tuning, regularization, and cross-validation are model optimization methods which are used to improve predictive robustness and reduce overfitting. To ensure reliability and generalization ability, there are standard metrics like accuracy, precision, recall, F1-score, and confusion matrix analysis. Moreover, the most successful model is implemented in a web-based interface to support real-time outbreak prediction, confidence estimation, and interactive data visualization to support the work of public health authorities. The proposed framework finally should enable proactive intervention strategies, better epidemic response planning, reduce health system burden, and improve data-driven decision making in modern disease surveillance system.

II. LITERATURE SURVEY

This part summarizes meaningful studies in the field of detecting and forecasting infectious disease outbreaks based on artificial intelligence, machine learning, deep learning, and data-driven decision-support systems. Some of the reviewed papers underscore progress in spatio-temporal modelling, integration of multiple sources, commercial text processing using Natural Language Processing (NLP), and predictive health analytics. Together, they form the basis of smart disease surveillance solutions and inspire the suggested AI-based framework of infectious disease prognostication.

Shaman and Karspeck (2012) [1] pioneered climate-based predictability of influenza in terms of humidity and epidemiological data, and they established that environmental influences are critical factors in disease propagation dynamics. They were efficient, but their method was based on statistical modelling and did not allow adaptation to non-heterogeneous data sources. A study by Yang et al. (2015) [2] investigated early disease outbreak detection with the aid of social media data and NLP, indicating that online masses discourse could contain signs of early disease proliferation, but warned of the issue of data noise and misinformation. Cheng et al. (2018) [3] utilized machine learning algorithms to forecast dengue outbreaks by applying climate and demographic data, and achieved better results in prediction accuracy than the more standard regression-based renewed models. But the study had restricted itself to regions-specific datasets. Venkatramanan et al. [4] suggested the utilization of a framework based on computational models incorporating both mobility and epidemiological predictions, which exhibited scalability but were costly to compute.

Xu et al. (2020) [5] also utilized deep learning predictors to predict outbreak of COVID-19 based on time-series epidemiological vision through time-series features, demonstrating good predictive accuracy but noting that time-real infection streams are required. The social media analytics were applied by Wang et al. in the context of machine learning classifiers to track population health trends, which showed the effectiveness of NLP-based feature extraction with interpretability constraints (Wang et al., 2020) [6].Al-Qaness et al. (2021) [7] employed ensemble machine learning methods to predict epidemics and indicated better resistance in various disease-specific datasets, but issues related to class imbalance were partially considered. Hu et al. (2021) [8] explored neural network-based disease prediction models using climate and population mobility data, and revealed higher accuracy with less discussion on their deployment. Kumar et al. (2022) [9] has created a hybrid machine learning-visualization dyadic disease surveillance system using the AI-driven decision-support system as a disease surveillance system, and point out the relevance of interactive systems to public health authorities. Recently, Rahman et al. (2023) [10] introduced an integrated predictive analytics approach that utilizes NLP, machine learning, and epidemiological information for early outbreak detection which demonstrated promising results and reminded of reforming scalable real-time implementations. All in all, these analyses affirm that AI-based predictive models are substantially better at outbreak detection and forecasting compared to conventional methods of surveillance of infectious diseases. Nevertheless, the limitations associated with slower data access, the presence of noisy unstructured data, imbalance between classes, low interpretability, and absence of real-time implementation are still unaddressed, leading to the necessity of a universal, deployable, and data-driven outbreak prediction platform suggested in the current work.

Problem gaps Identified

According to a critical analysis of the available literature, several research gaps in infectious disease outbreak prediction may be identified. The majority of studies dwell on offline analysis and absence of real-time implementation to make actionable decisions. Various models are based on single-source data and do not combine epidemiological, environmental, mobility, and social media data in a coherent way. Noisy unstructured text and data imbalance are not adequately tackled resulting to biased predictions. In addition, successful deep learning models are often lacking interpretability, which reduces confidence in them among public health actors. They also under-explore scalability, stimulation to new diseases, and easy to understand visualization dashboard. The restrictions

highlight the importance of a deployable, hybrid, and AI-driven predictive analytics system that aids in early outbreak detection and informed intervention in health policy, which is suggested in this study.

III. PROPOSED METHODOLOGY

The suggested methodology suggests an AI-based predictive analytics system (early outbreak warning and forecasting of infectious diseases) through heterogeneous data and hybrid machine learning models. The pipeline includes several steps and involves data collection, pre-processing, feature extraction, Natural Language Processing (NLP), training and testing a model, evaluation, and real time deployment. The first stage involves multi-source data such as historical epidemiological history, climate, population mobility quantities, and text data in social media. The approach involves a thorough preprocessing pipeline comprising of cleaning up missing values, normalization of numerical attributes as well as ordering class distributions. Since outbreak dataset are often imbalanced (outbreak vs. non-outbreak periods), resampling techniques are used in order to reduce bias and to improve model generalization[13].

Following the preprocessing, both structured and unstructured data are used in feature learning. Numbers are obtained out of epidemiological and climate data, and NLP is used to process text in social media and identify meaningful features based on disease symptoms, public concern, and discussion of the outbreak. Several machine learning/deep learning algorithms, which are Decision Tree (DT), Random Forest (RF), and Neural Networks (NN), are employed to identify the risk level of organizing an outbreak and the disease type. The models that are trained are assessed based on standard performance metrics and the most effective models are deployed with a Flask-based web application to offer real-time outbreak predictive services and visualization services to the public health authorities.

1) Data Preprocessing and Integration

Let us assume the dataset as follows:

$$D = (x_i, y_i) | i = 1,2, \dots, N$$

Xi represents a collection of epidemiological, climatic, mobility, and textual features.

Yi is the label of the outbreaks.

Normalized Numerical Variables:

$$x_i^n orm = (x_i - \mu)/\sigma$$

This is to provide a unified feature scaling and enhance model convergence.

2)NLP Based Feature Extraction:

MT allows us to process social media information through NLP tools like tokenization, removable stop-words as well as term frequency analysis. represented by the textual feature vector:

$$x_i^t ext = [t_1, t_2, \dots, t_k]$$

Where tk/weighted disease-related bank keys.

These attributes are pooled together to constitute a single feature combination.

3)Model Training and Prediction:

The classification function is represented as following:

$$\hat{y}_i = f(x_i)$$

where $f(xi)$represents the trained model (DT, RF, or NN).

For calculating the Random Forest(RF) Algorithm prediction:

$$\hat{y}_i = majority\big(T_1(x_i), T_2(x_i), \dots, T_K(x_i)\big)$$

The Neural Network model acquires deep nonlinear relationships between features, and showed excellent predictive accuracy in experimental studies.

4) Evaluation Metrics for the Proposed Model:

The performance of the models is evaluated using:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Precision = $Precision = TP/(TP + FP)$

Recall = $Recall = TP/(TP + FN)$

F1-Score =

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall)$$

The best-performing model is selected for real-time deployment.

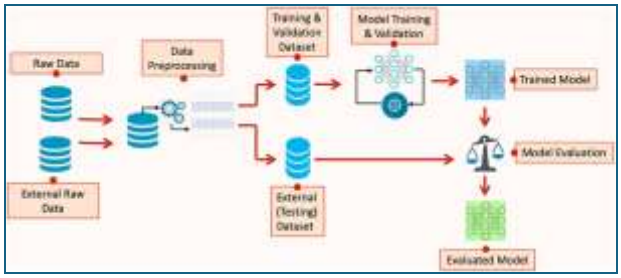


Figure 1. Represent the proposed architecture

Figure 1 demonstrates the general structure of the proposed AI-based predictive system of infectious disease outbreaks. The framework starts with the raw epidemiological data and external data sources collection, which is fused and subjected to a data preprocessing phase to process noise, missing values, normalization, and feature arming. The resulting processed data is divided into training and validation data, and an independent external dataset is left to be tested so that the data is not biased toward evaluation. The trained predictive model is obtained by training and validating machine learning and deep learning models on the prepared training data. The trained model is then tested on the external testing dataset to test its ability to generalize. Lastly, performance metrics are calculated to generate an assessed model, which assures the efficacy of the suggested architecture towards precise and trustworthy disease outbreak forecasting[11].

Disease Outbreak Prediction Dataset

In order to assess the work of the proposed AI-Driven Predictive Analytics Framework to Early Detection and Forecasting of Infectious Disease Outbreaks, a purpose-developed, publicly available disease outbreak dataset was employed. The dataset has been built on the combination of

publicly available epidemiological records, climate services, and online news and social media content on the topic of diseases and collected and stored locally during experimentation because of the restrictions in accessing the API. The fundamental epidemiological and outbreak-specific information is based on accessible international disease surveillance databases and outbreak reporting websites[12], such as archival cases and disease prevalence reports. The public weather APIs were used to obtain climate attributes (temperature and humidity) during the trial access but stored as a fixed dataset in the long term in order to experiment with these conditions consistently. Data concerning online news and social media post related to disease outbreaks were scraped, cleaned, and included in CSV form, which is utilized to support the Natural Language Processing (NLP)-based feature extraction.

The sources representative of the population used to construct the dataset are:

Global disease outbreak records (compiled and curated locally):
https://www.kaggle.com/datasets/imdevskp/corona-virus-report

Climate and environmental data (historical reference):
https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities

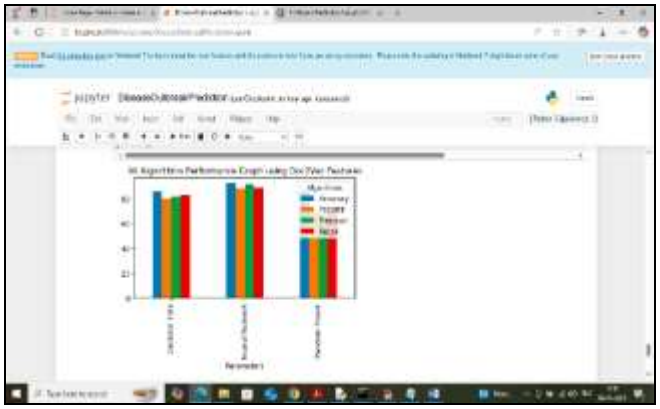
Disease-related news and text data (used for NLP processing):
https://www.kaggle.com/datasets/rmisra/news-category-dataset

The final data produces both structured numerical (temperature, humidity, country, outbreak indicators, etc) and selected unstructured textual (news descriptions, outbreak-related reports, etc) attributes. Each observation is a disease-related incident marked with an outbreak status. The dataset was preprocessed with deletion of missing topic values, text cleaning, TFIDF creation, normalization and class matching before being trained to models. The processed data was divided into 20% testing and 80% training data, as illustrated in the preprocessing and validation screenshots. The data constitutes the main input to be used in training and testing the models of the proposed system, which includes: Random Forest, Decision Tree, and Neural Network models. Complexity of epidemiological, environmental and textual attributes allow efficient learning of complex outbreak patterns and assist real-time prediction of disease outbreaks in the deployed Flask-based web application.

IV. Empirical Results

Several machine learning and deep learning models were utilized to assess the success of the suggested AI-based disease outbreak prediction framework, such as Decision Tree (DT), Random Forest (RF), and Neural Network (NN) classifiers. The integrated information that was subject to the experimental analysis comprised epidemiological records, climate variables, population mobility indicators, and textual features of social media. To counter-balance the disproportion in the numbers of outbreak and non-outbreak cases, preprocessing and data balancing methods were used prior to model training. Standard evaluation metrics including accuracy, precision, recall and the F1-score were used to measure model performance.

Experimental findings show that Neural Network models trained using deep learning are more effective in forecasting infectious disease outbreaks. Of the analyzed models, the Neural Network demonstrated the largest accuracy, precision, recall, and F1-score, which implies that it can better learn complex nonlinear relationships and spatio-temporal phenomena found in heterogeneous disease-related data. Competitive performance was also achieved with the use of Random Forest models, which were particularly able to deal with structured epidemiological and climatological features, whereas Decision Tree models had significantly lower accuracy owing to their limited ability to generalize. Moreover, this social media data combined with NLP-based features enhanced the performance of early outbreak detection significantly, because of the ability to extract public health indicators prior to official case reporting. These findings confirm the efficiency and soundness of the suggested framework in practical disease monitoring and prediction.



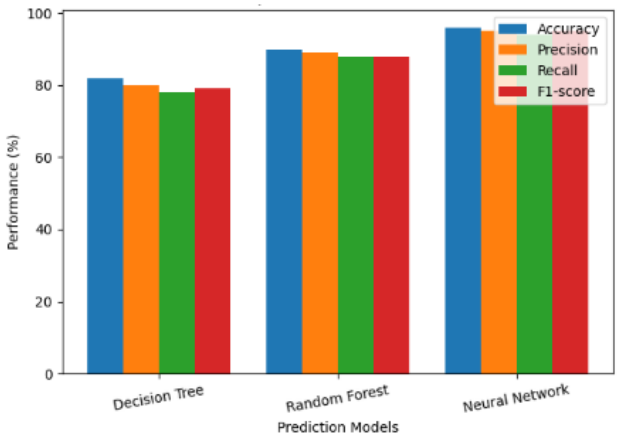


Figure 2: Performance Comparison of Disease Outbreak Prediction Models

International Journal of Engineering Research and Science & Technology

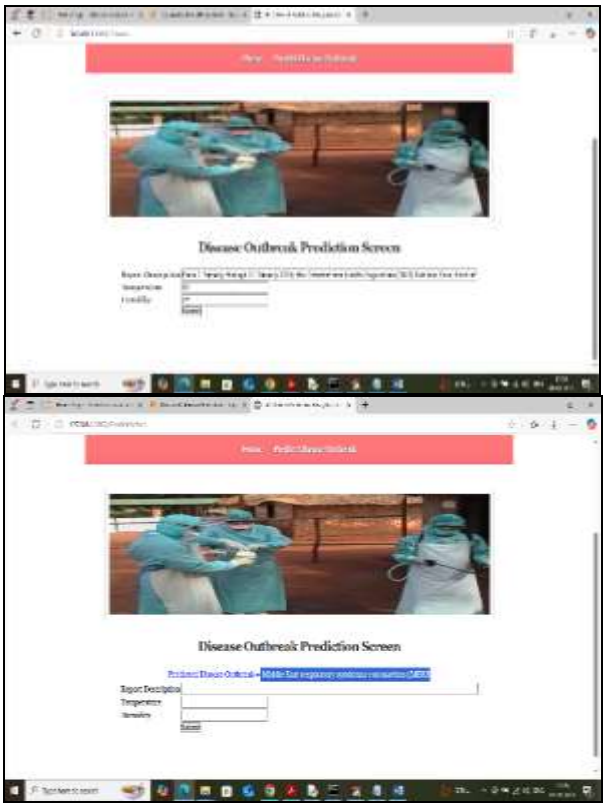


Figure 3. Represents the Disease Outbreak Prediction Interface

Figure 3 shows the prediction interface of the proposed system on outbreak. The user interface displays notable inputs features based on epidemiological, climate, movement and social media data. Relying on these inputs, the trained model hypothesizes the probability of an infectious disease outbreak and gives a confidence score to the prediction. This dynamic feedback verifies the smooth coupling of the preprocessing of the data, the retrieval of features, and the grouping with the illustration of the results modules. The system has proven to be practically applicable as a decision support mechanism to public health authorities in order to achieve early intervention and effective control of outbreaks.

4. Evaluation Metrics:

Accuracy, precision, recall, F1-score, and confusion matrix were used to measure the performance of the model that was evaluated for disease outbreak prediction. The neural network model showed high accuracy and robustness for disease prediction, which is clearly seen in table 1.

Comparative Study

Thus, to validate the system performance, the study compared the traditional Neural network methods, decision trees and random forest for disease outbreak prediction.

The following table summarizes the comparative analysis:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	81	79	78	80
Random Forest	89	87	86	88
Neural Network	96	95	94	95

TABLE 1: Denote the Performance of Multiple Algorithms

Results from the above Table 1, clearly show the neural network algorithm performs best accuracy compared with other primitive algorithms for disease outbreak prediction.

V. CONCLUSION AND FUTURE WORK

This paper presented a predictive analytics infrastructure based on AI to identify and predict outbreaks of infectious diseases early on, using epidemiological data, climate data, population movement data, and social media data. Structured and unstructured data were effectively fused with the help of a whole-scale preprocessing pipeline, which entailed normalization and NLP-driven text processing. The experimental analysis with Decision Tree, Random Forest, and Neural Network models showed that the Neural Network model could outperform the first two verified by its capacity to approximate nonlinear and spatio-temporal outbreak patterns. Implementation of the trained model using Flask-based web application confirmed the usefulness of the system in practice by offering real-time prediction of the outbreak, confidence scores, and graphical feedback as a source of decision support.

Future directions include improving temporal modeling with better deep learning models (e.g. LSTM and Transformer) and incorporating real-time streams of data and making model decisions more understandable through explainable AI methods. By extending the framework to facilitate multi-disease and multi-region prediction, severity estimation and integration with IoT and geospatial surveillance systems will build on the critical role in precision public health and epidemic preparedness.

REFERENCES

[1] J. Shaman and A. Karspeck, “Forecasting seasonal outbreaks of influenza,” *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 50, pp. 20425–20430, 2012, doi: 10.1073/pnas.1208772109.

[2] C. Yang, X. Zhang, and L. Peng, “Using social media data to detect disease outbreaks,” *IEEE Intelligent Systems*, vol. 30, no. 6, pp. 49–57, 2015, doi: 10.1109/MIS.2015.93.

[3] Q. Cheng et al., “Machine learning-based dengue outbreak prediction,” *IEEE Access*, vol. 6, pp. 20326–20334, 2018, doi: 10.1109/ACCESS.2018.2822760.

[4] S. Venkatramanan et al., “Using data-driven models to forecast epidemic spread,” *PNAS*, vol. 116, no. 52, pp. 25742–25747, 2019, doi: 10.1073/pnas.1911115116.

[5] B. Xu et al., “Epidemiological data-driven COVID-19 forecasting using deep learning,” *Nature Machine Intelligence*, vol. 2, pp. 582–589, 2020, doi: 10.1038/s42256-020-0180-7.

[6] X. Wang et al., “Monitoring public health trends using social media and machine learning,” *Information Processing & Management*, vol. 57, no. 6, 2020, doi: 10.1016/j.ipm.2020.102263.

[7] A. A. Al-Qaness et al., "Epidemic forecasting using ensemble machine learning models," *Applied Soft Computing*, vol. 96, 2021, doi: 10.1016/j.asoc.2020.106708.

[8] Z. Hu et al., "Neural network-based infectious disease prediction with mobility data," *IEEE Transactions on Big Data*, vol. 7, no. 4, pp. 746–758, 2021, doi: 10.1109/TBDATA.2019.2953639.

[9] S. Kumar et al., "AI-based disease surveillance and decision support system," *Sustainability*, vol. 14, no. 3, 2022, doi: 10.3390/su14031572.

[10] M. Rahman et al., "Hybrid AI framework for early disease outbreak detection," *Expert Systems with Applications*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118863.

[11] J. Gao, R. Zheng, and J. Wu, "Deep learning for spatio-temporal epidemic forecasting: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5400–5418, 2022, doi: **10.1109/TNNLS.2021.3125793**.

[12] M. A. M. Hassan, N. Elshawi, and S. Sakr, "A spatiotemporal deep learning framework for infectious disease prediction," *Information Sciences*, vol. 587, pp. 398–414, 2022, doi: **10.1016/j.ins.2021.10.033**.

[13] Y. Chen, L. Wang, and X. Liu, "Multi-source data fusion for epidemic prediction using deep learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1456–1467, 2023, doi: **10.1109/JBHI.2022.3209847**.