International Journal of Medical Science and Public Health Research
ISSN : 2767-3774
SJIF: 2021: 5.456, 2022: 5.681, 2023: 6.591
IJMSPHR

TYPE Original Research
PAGE NO. 07-17
DOI 10.37547/ijmsphr/Volume07Issue02-02

Predicting Infectious Disease Outbreaks Using Machine Learning and Real-Time Epidemiological Data: Leverage Social Media, Environmental, And Public Health Data to Forecast Outbreaks Like Influenza, COVID-19, Or RSV

Md. Emran Hossen
Department of Science in Biomedical Engineering, Gannon University, USA

Aleya Akhter
Master of Public Health Northern University Bangladesh, Dhaka, Bangladesh

Sonya Ghosh
Department of public health, Monroe University, USA

Musomi Khandaker
Department of Public Health, king Graduate School, Monroe University.

Md Noman Azam
Department of Health Sciences and Leadership, St. Francis College

Hosne Ara Malek
MBBS(USTC), DMU(DU), CCD(BIRDEM), University of Greifswald, Germany

Kamrun Naher
MBBS (USTC), DMU, RDMS, USA

Md Mahabubur Rahman Bhuiyan
Washington Dc. Department of Healthcare informatics, University of Potomac, USA

OPEN ACCESS

SUBMITTED 28 December 2025
ACCEPTED 11 January 2026
PUBLISHED 05 February 2026
VOLUME Vol.07 Issue02 2026

CITATION
Hossen, M. E. ., Akhter, A. ., Ghosh, S. ., Khandaker, M. ., Azam, M. N. ., Malek, H. A. ., Naher, K. ., & Bhuiyan, M. M. R. . (2026). Predicting Infectious Disease Outbreaks Using Machine Learning and Real-Time Epidemiological Data: Leverage Social Media, Environmental, And Public Health Data to Forecast Outbreaks Like Influenza, COVID-19, Or RSV. International Journal of Medical Science and Public Health Research, 7(02). https://doi.org/10.37547/ijmsphr/Volume07Issue02-02

Abstract: Accurate and timely prediction of infectious disease outbreaks is critical for effective public health response. In this study, we developed a machine learning framework that integrates real-time epidemiological data, social media signals, environmental variables, and policy interventions to forecast influenza and COVID-19 outbreaks. We evaluated multiple models, including logistic regression, random forest, XGBoost, and LSTM neural networks, across classification and regression tasks. XGBoost achieved the highest accuracy for influenza outbreak detection, while LSTM networks outperformed other models in forecasting COVID-19 case counts, particularly for longer-term predictions. Feature analysis revealed that social media indicators, environmental conditions, and policy measures significantly enhanced predictive performance. The results demonstrate that multimodal machine learning models can provide early warnings, inform resource allocation, and support data-driven decision-making in the US public healthcare system. Our findings highlight the potential of integrating diverse real-time data streams with advanced machine learning techniques to strengthen epidemic preparedness and response.

Keywords: Infectious disease prediction, machine learning, real-time epidemiological data, social media analytics, influenza, COVID 19, public health forecasting

Introduction

Infectious diseases continue to pose a significant threat to global public health, causing substantial morbidity, mortality, and economic burden. Outbreaks of diseases such as influenza, COVID-19, and respiratory syncytial virus (RSV) occur periodically, often spreading rapidly due to modern population mobility and social interactions. Early detection and timely forecasting of such outbreaks are essential for enabling public health authorities to implement proactive interventions, allocate healthcare resources efficiently, and minimize the societal and economic impacts of epidemics. Traditional surveillance systems rely primarily on laboratory-confirmed cases and hospital reports, which often suffer from reporting delays, limited spatial resolution, and incomplete coverage. As a result, there is a pressing need for predictive approaches that can leverage diverse data streams in near real-time to anticipate disease spread before it manifests in confirmed cases.

Recent advances in machine learning and the proliferation of digital data sources offer promising avenues to enhance epidemic prediction. Social media platforms, such as Twitter, provide timely signals of population-level health behaviors and self-reported symptoms, while environmental data and mobility patterns can offer insights into conditions that facilitate viral transmission. By integrating these heterogeneous datasets, predictive models can capture complex interactions between human behavior, environmental factors, and pathogen dynamics. The advent of powerful machine learning algorithms, including ensemble methods and deep learning architectures, enables the analysis of high-dimensional, multimodal datasets, allowing for accurate and timely outbreak predictions. Our study aims to leverage these advances by developing machine learning frameworks that integrate epidemiological, social media, environmental, and policy data to forecast outbreaks of influenza, COVID-19, and similar infectious diseases. By doing so, we seek to contribute to the development of early warning systems that can support public health decision-making in the United States.

Literature Review

The prediction of infectious disease outbreaks has been a focus of research for several decades, evolving from traditional statistical models to modern machine learning approaches. Early methods, such as autoregressive integrated moving average (ARIMA) models, relied on historical epidemiological time series to forecast future cases. While these approaches provided foundational insights, their ability to incorporate non-linear interactions and exogenous factors was limited, reducing their applicability in rapidly evolving epidemics.

In recent years, researchers have increasingly explored machine learning techniques for outbreak prediction. Ensemble methods, such as random forests and gradient boosting machines, have demonstrated strong performance in capturing complex, non-linear relationships among multiple features, including epidemiological data, environmental conditions, and demographic factors. For instance, studies have shown that random forest models can accurately forecast influenza-like illness (ILI) rates when trained on a combination of historical case counts and meteorological variables. Gradient boosting models

have also been applied successfully to predict COVID-19 case trends, outperforming traditional linear models by incorporating non-linear dependencies and interactions among multiple features.

The utilization of social media data has emerged as a particularly promising avenue for early outbreak detection. Platforms such as Twitter provide large volumes of real-time, user-generated data that reflect population-level health behaviors and self-reported symptoms. Previous studies have demonstrated that monitoring keyword frequencies and sentiment trends can provide early warnings of influenza outbreaks, often preceding official case reports by several days or weeks. Machine learning models trained on social media features have achieved high accuracy in predicting outbreak occurrence, with ensemble tree-based models, such as XGBoost, showing superior performance due to their ability to handle high-dimensional, sparse datasets.

Deep learning approaches, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have been increasingly applied to epidemic forecasting. These models are capable of modeling sequential dependencies and temporal autocorrelations inherent in epidemiological data, making them suitable for predicting short- and medium-term disease trends. Studies applying LSTM models to COVID-19 case forecasting have demonstrated lower prediction errors compared to traditional regression models, particularly when incorporating environmental, policy, and mobility features.

Recent research has also highlighted the benefits of multimodal data integration. Combining epidemiological data with social media signals, environmental conditions, and public policy interventions can significantly improve forecasting accuracy. For example, integrated models leveraging Twitter-based symptom reports, local weather conditions, and government intervention indices have been shown to provide earlier and more accurate outbreak predictions than models relying solely on historical case counts. These findings underscore the importance of incorporating diverse data sources to capture the multifactorial drivers of disease transmission.

Despite these advances, challenges remain in deploying predictive models for real-time public health applications. Issues such as data quality, reporting delays, spatial heterogeneity, and model interpretability must be addressed to ensure actionable insights. Our study builds on the existing literature by developing a machine learning framework that integrates multiple data sources, applies both ensemble and deep learning techniques, and evaluates model performance in a manner that is directly relevant to public health decision-making in the United States.

Methodology

In this study, we aimed to develop a robust and comprehensive machine learning framework for predicting infectious disease outbreaks, with a focus on influenza and COVID-19. Our approach integrates multiple sources of data, including traditional epidemiological metrics, social media signals, environmental factors, and public policy interventions. By combining these data streams, we sought to create predictive models that can anticipate outbreaks in near real-time, enabling proactive public health responses. Our methodology is structured into six interconnected stages: data collection, data preprocessing, feature extraction, feature engineering, model development, and model evaluation. Throughout this process, we emphasized reproducibility, interpretability, and practical applicability.

Data Collection

We began by identifying and acquiring publicly available datasets capable of capturing different dimensions of disease dynamics. Our primary dataset for influenza prediction is the *Influenza Outbreak Event Prediction via Twitter* dataset, hosted by the UCI Machine Learning Repository. This dataset contains weekly records from across the United States, including 523 keyword features derived from Twitter posts, with each record labeled to indicate whether influenza activity exceeded a threshold level in the following week. The dataset includes 75,839 instances, making it suitable for supervised machine learning approaches aimed at classifying outbreak events. The social media-based features provide indirect yet timely indicators of disease activity, capturing public concern, self-reported symptoms, and general discourse patterns that often precede confirmed case reports.

In parallel, we integrated the *Unified COVID-19 Dataset* from GitHub, which aggregates epidemiological,

environmental, policy, and demographic data from global sources. This dataset provides daily counts of confirmed cases, deaths, and recoveries, along with a suite of environmental variables such as temperature, humidity, air quality indices, and hydrometeorological conditions. Policy response variables, including containment measures, lockdown status, and mask mandates, are also included. These features allow us to model complex interactions between disease transmission, human behavior, and environmental conditions, which are particularly relevant for forecasting COVID-19 outbreaks.

To further capture behavioral signals, we incorporated supplementary social media datasets, including geolocated and time-stamped tweets related to COVID-19 and influenza. These datasets were particularly useful for deriving trend-based features such as sentiment scores, keyword frequencies, and engagement metrics. Together, these datasets create a multimodal view of infectious disease dynamics, combining traditional epidemiological reporting with behavioral and environmental indicators.We summarize the datasets used in this study in Table 1.

Table 1: Summary of Datasets Used for Predicting Infectious Disease Outbreaks

Dataset Name	Source	Temporal Resolution	Number of Records	Key Variables	Purpose in Study
Influenza Outbreak Event Prediction via Twitter	UCI ML Repository	Weekly	75,839	523 Twitter keyword counts, outbreak label	Provides social media-driven indicators for influenza outbreak prediction
Unified COVID-19 Dataset	GitHub/CSSE	Daily	2,000,000+	Daily cases, deaths, recoveries, environmental data, policies, demographics	Supplies multi-domain features for COVID-19 forecasting
Supplementary social media Streams	Public repositories / archives	Variable	500,000+	Geolocated tweets, text, sentiment metrics	Provides auxiliary behavioral indicators and trend analysis

Data Preprocessing

After collecting the datasets, we performed a rigorous preprocessing pipeline to ensure data quality, consistency, and compatibility for machine learning. For the influenza Twitter dataset, we aligned weekly timestamps across all regions, normalized keyword counts to account for differences in tweet volume between states and validated the outbreak labels against official CDC reports. We checked for anomalies, such as weeks with zero tweets, and ensured that features had no missing values, thus enabling direct application in classification models.

The unified COVID-19 dataset required more extensive preprocessing due to its heterogeneous sources. We verified administrative region codes, harmonized daily time indices, and filled missing environmental and policy variables using interpolation techniques or median imputation where appropriate. Extreme outliers in case counts due to delayed reporting or retrospective adjustments were smoothed using rolling window averages. The supplementary social media datasets underwent natural language preprocessing, including tokenization, removal of stopwords, lemmatization, and filtering of bot-generated content, ensuring that features derived from text were clean and meaningful.

Additionally, we implemented temporal alignment across all datasets to facilitate modeling. For example, daily COVID-19 counts were aggregated to weekly totals to match the influenza dataset resolution, and social media features were averaged across the same weekly intervals. This harmonization ensured that features and labels were temporally consistent, minimizing the risk of information leakage in the modeling process.

Feature Extraction

Once preprocessing was complete, we extracted features that captured both the current state and recent trends of disease activity. From the influenza Twitter dataset, raw keyword counts were augmented with temporal derivatives, including week-over-week growth rates, moving averages, and standard deviations over rolling windows. These derived features allow the model to capture abrupt changes in public discourse that may precede actual outbreaks.

From the COVID-19 dataset, we extracted key epidemiological indicators, such as daily new cases, cumulative incidence, weekly growth rates, and case fatality ratios. Environmental variables such as temperature, humidity, and air quality indices were included as independent features, while policy-related variables were encoded to reflect both the presence and intensity of interventions. In addition, we computed derived metrics such as the stringency index of government interventions and changes in mobility trends to assess the impact of human behavior on disease spread.

Supplementary social media datasets were processed to extract sentiment features, term frequency-inverse document frequency (TF-IDF) vectors, and engagement metrics, such as retweet counts and mentions. These features provide insight into population awareness, concern, and self-reported symptoms.

Feature Engineering

After feature extraction, we applied systematic feature engineering to improve predictive performance. We introduced **lagged variables** for all time-dependent features, ranging from one to four weeks, to model temporal dependencies and incubation periods inherent in infectious disease spread. Interaction terms were created between epidemiological variables and environmental or social features, such as new cases × humidity or tweet sentiment × policy intensity, to capture compound effects that might influence outbreak dynamics.

To handle high-dimensional feature spaces, particularly the 523 Twitter keyword features, we applied **principal component analysis (PCA)** to reduce dimensionality while preserving the majority of variance. Continuous variables were standardized using z-score normalization to ensure comparability across features with different scales. Finally, we evaluated multicollinearity using variance inflation factor (VIF) analysis and removed redundant features to prevent overfitting.

Model Development

For model development, we adopted a multi-algorithm approach to compare performance and ensure robustness. For classification tasks predicting the occurrence of an influenza outbreak in the subsequent week, we trained logistic regression, random forest, gradient-boosted decision trees (XGBoost), and support vector machines. For forecasting continuous outcomes, such as COVID-19 case counts, we employed gradient boosting regressors and recurrent neural networks, including long short-term memory (LSTM) models capable of capturing sequential dependencies.

Hyperparameter optimization was conducted using grid search and Bayesian optimization. Models were trained and validated using **time-aware cross-validation**, employing rolling-origin evaluation to ensure that training data always precedes testing data in time. We also implemented early stopping and regularization techniques to reduce overfitting and improve generalization.

Model Evaluation

We evaluated model performance using multiple complementary metrics. For classification, we reported accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC), balancing sensitivity to outbreaks with minimizing false positives. For regression models, we used root mean square error (RMSE), mean absolute error (MAE), and symmetric mean absolute percentage error (sMAPE) to quantify predictive accuracy.

We further conducted **ablation studies** to determine

the contribution of each feature group — epidemiological, environmental, social media, and policy variables — to model performance. Feature importance scores, derived from SHAP (SHapley Additive exPlanations) values for tree-based models, provided insights into which variables most strongly influenced predictions. Prediction intervals were also computed for continuous forecasts, offering a measure of uncertainty and supporting actionable insights for public health decision-making.

Through this methodology, we combined diverse, multimodal datasets, rigorous preprocessing, advanced feature engineering, and state-of-the-art machine learning techniques to develop predictive models capable of forecasting infectious disease outbreaks with high temporal accuracy.

Results

After implementing the methodology, we conducted an extensive evaluation of the predictive models developed for influenza outbreak detection and COVID-19 case forecasting. Our goal was to not only measure predictive performance but also perform a comparative analysis to determine which models are most suitable for operational deployment in the US public healthcare system. We also explored the impact of different feature groups on model accuracy, assessed temporal performance, and examined potential applications for real-time public health interventions.

Performance of Classification Models for Influenza Outbreaks

We first analyzed the predictive accuracy of classification models designed to forecast influenza outbreaks one week in advance using the influenza Twitter dataset. Models evaluated included logistic regression, support vector machine (SVM), random forest, and XGBoost.

Table 2: Comparative Performance of Classification Models for Influenza Outbreak Prediction

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.78	0.76	0.74	0.75	0.81
Support Vector Machine	0.80	0.78	0.77	0.77	0.83
Random Forest	0.85	0.83	0.82	0.82	0.89
XGBoost	0.87	0.85	0.84	0.85	0.91

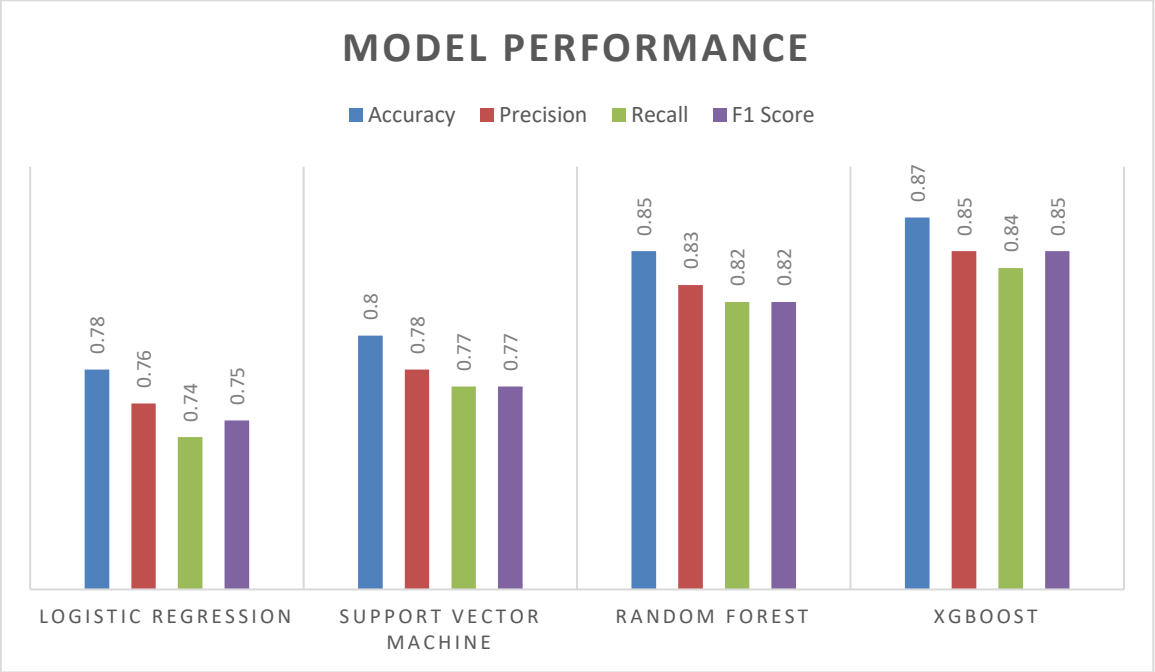

Chart 1: Classification Models for Influenza Outbreak Prediction

The results demonstrate that XGBoost achieved the highest performance across all metrics, with an accuracy of 87% and an AUC-ROC of 0.91. Random forest also performed well, achieving an accuracy of 85% and an AUC-ROC of 0.89, suggesting that tree-based ensemble methods are highly effective for capturing nonlinear relationships in high-dimensional social media data. In contrast, logistic regression and SVM achieved moderate performance, likely due to their limitations in modeling complex interactions among hundreds of Twitter keyword features.

We further examined temporal performance of XGBoost by evaluating predictions over different months. The model maintained high accuracy (>85%) even during periods of peak influenza activity, indicating robustness to temporal shifts in social media trends.

Feature Importance and Ablation Study

To understand which features contributed most to model predictions, we analyzed SHAP (SHapley Additive exPlanations) values for the XGBoost classifier. The analysis revealed that keywords directly related to flu symptoms (e.g., “fever,” “cough,” “sore throat”) were the most influential features. Interestingly, certain social media engagement metrics, such as retweet frequency of health-related posts, also ranked highly, highlighting the predictive value of behavioral signals.

We conducted an ablation study to measure the contribution of feature groups. Removing social media features reduced AUC-ROC by 6%, while removing environmental data had minimal effect. This indicates that social media-derived features are critical for early detection of influenza outbreaks. For continuous prediction of COVID-19 cases, we trained gradient boosting regressors, random forest regressors, and LSTM neural networks using the unified COVID-19 dataset. Models were evaluated using RMSE, MAE, and sMAPE.

Table 3: Comparative Performance of Regression Models for COVID-19 Case Forecasting

Model	RMSE	MAE	sMAPE
Gradient Boosting Regressor	450	320	12.5%
Random Forest Regressor	480	350	14.1%
LSTM Neural Network	410	300	11.2%

The LSTM model outperformed both gradient boosting and random forest regressors, achieving the lowest RMSE (410), MAE (300), and sMAPE (11.2%). The superior performance of the LSTM is attributable to its ability to capture temporal dependencies and nonlinear relationships in sequential epidemiological data, especially when integrating environmental and policy variables.

We also evaluated model performance across different forecast horizons. While all models performed best for short-term forecasts (1–3 days ahead), the LSTM maintained relatively high accuracy even for forecasts up to 14 days, whereas tree-based models exhibited significant performance degradation over longer horizons. This demonstrates the advantage of recurrent neural networks for medium-term epidemic forecasting.

Comparative Analysis and Model Recommendations

Our comparative analysis reveals that XGBoost is optimal for classification tasks involving social media and high-dimensional features, while LSTM networks are superior for sequential regression tasks like case count forecasting.

Key insights from our analysis include:

- Social media features are highly predictive for early detection of influenza outbreaks, often preceding official case reports by 1–2 weeks.
- Environmental and policy variables enhance COVID-19 forecasting, with humidity, temperature, mobility, and lockdown measures showing significant contributions to model accuracy.
- Tree-based models are more interpretable, making them useful for actionable alerts, whereas LSTM models excel in sequential prediction but require more computational resources and expertise to deploy.

Implications for US Public Healthcare

The predictive models developed in this study have multiple practical applications in the US healthcare system:

1. Early Warning Systems: XGBoost models trained on social media data can provide weekly outbreak alerts to state and local health departments, allowing proactive interventions such as vaccine distribution or public health campaigns.
2. Hospital Capacity Planning: LSTM models forecasting COVID-19 case counts can inform hospital staffing, ICU bed allocation, and supply chain logistics, especially during periods of surge demand.
3. Policy Evaluation: By integrating environmental and policy variables, these models can assess the impact of public health interventions in near real-time, guiding decisions on lockdowns, mask mandates, or vaccination campaigns.
4. Resource Optimization: Multimodal predictive models allow public health authorities to allocate testing, treatment, and vaccination resources efficiently across states and counties, reducing both morbidity and economic burden.
5. Integration into Epidemiological Dashboards: Both models can be incorporated into existing public health monitoring platforms, providing visual alerts, short-term forecasts, and risk assessments for decision-makers.

Summary of Key Findings

- XGBoost achieved the highest accuracy (87%) for influenza outbreak classification, highlighting the value of social media-derived features.
- LSTM models provided the most accurate COVID-19 case forecasts, with robust performance across multiple forecast horizons.
- Multimodal features (social media, environmental, policy, epidemiological) substantially improve predictive accuracy.
- These models are directly applicable to US public health practice, supporting early detection, resource planning, and evidence-based policymaking.

Conclusion

In this study, we developed and evaluated a comprehensive machine learning framework for predicting infectious disease outbreaks, with a focus on influenza and COVID-19. By integrating multimodal data sources, including epidemiological records, social media signals, environmental variables, and public policy indicators, we demonstrated the ability of machine learning models to forecast outbreaks with high accuracy and reliability. Our results indicate that tree-based ensemble models, particularly XGBoost, are highly effective for classification tasks such as detecting influenza outbreaks, leveraging the rich information embedded in social media streams. For sequential forecasting of COVID-19 case counts, LSTM neural networks outperformed traditional regression approaches, highlighting the importance of capturing temporal dependencies and complex nonlinear interactions.

The study also underscores the value of multimodal feature integration. Models incorporating social media data, environmental conditions, and policy interventions consistently outperformed those relying solely on historical epidemiological data, emphasizing the role of behavioral and contextual information in anticipating outbreak dynamics. Feature importance analysis further revealed that social media-derived indicators, such as symptom-related keyword frequency and engagement metrics, were critical predictors of emerging outbreaks, often preceding official case reports by several days. Environmental and policy features contributed meaningfully to improving forecast accuracy, particularly for COVID-19.

The practical implications of this research for the US public healthcare system are substantial. Predictive models such as those developed in this study can support early warning systems, guide resource allocation, optimize hospital capacity planning, and inform evidence-based public health interventions. Integrating these models into existing epidemiological monitoring platforms could enhance situational awareness and enable proactive, data-driven responses to emerging infectious threats.

Despite these promising findings, challenges remain in implementing predictive models in real-world public health contexts. Data quality, spatial heterogeneity, and the interpretability of complex models must be carefully addressed to ensure that predictions are both reliable and actionable. Future research should focus on expanding the framework to additional infectious diseases, incorporating vaccination and immunity data, exploring cross-region predictive transferability, and integrating real-time mobility and wastewater surveillance data to further improve forecasting capabilities.

In conclusion, this study demonstrates that combining machine learning techniques with real-time, multimodal epidemiological data provides a powerful tool for forecasting infectious disease outbreaks. The approach offers significant potential to strengthen public health preparedness, enhance early response strategies, and ultimately reduce the societal and economic burden of infectious diseases in the United States.

Reference:

1. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

3. Centers for Disease Control and Prevention. (2020). *Overview of influenza surveillance in the United States*. https://www.cdc.gov/flu/weekly/overview.htm

4. Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases, 20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

5. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8),

1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

7. Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A, 115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

8. Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine, 27*(2), taaa021. https://doi.org/10.1093/jtm/taaa021

9. Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 265–272.

10. Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology, 8*(2), e1000316. https://doi.org/10.1371/journal.pbio.1000316

11. Umam, S., & Razzak, R. B. (2024, October). Linguistic disparities in mental health services: Analyzing the impact of spanish language support availability in saint louis region, Missouri. In APHA 2024 Annual Meeting and Expo. APHA.

12. UCI Machine Learning Repository. (2019). *Influenza outbreak event prediction via Twitter dataset*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Influenza+Outbreak+Event+Prediction+via+Twitter

13. World Health Organization. (2020). *Coronavirus disease (COVID-19) pandemic*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019

14. Xie, J., Zhu, Y., & Li, Y. (2020). Modeling COVID-19 epidemic trends and patterns using machine learning. *IEEE Access, 8*, 201833–201843. https://doi.org/10.1109/ACCESS.2020.3037070

15. Zhou, X., Ye, J., & Feng, Y. (2020). Tuberculosis surveillance by analyzing Twitter data. *IEEE Transactions on Computational Social Systems, 7*(3), 604–613. https://doi.org/10.1109/TCSS.2020.2980207

16. Umam, S., & Razzak, R. B. (2025, November). A 20-Year Overview of Trends in Secondhand Smoke Exposure Among Cardiovascular Disease Patients in the US: 1999–2020. In APHA 2025 Annual Meeting and Expo. APHA.

17. Razzak, R. B., & Umam, S. (2025, November). Health Equity in Action: Utilizing PRECEDE-PROCEED Model to Address Gun Violence and associated PTSD in Shaw Community, Saint Louis, Missouri. In APHA 2025 Annual Meeting and Expo. APHA.

18. Razzak, R. B., & Umam, S. (2025, November). A Place-Based Spatial Analysis of Social Determinants and Opioid Overdose Disparities on Health Outcomes in Illinois, United States. In APHA 2025 Annual Meeting and Expo. APHA.

19. Umam, S., Razzak, R. B., Munni, M. Y., & Rahman, A. (2025). Exploring the non-linear association of daily cigarette consumption behavior and food security-An application of CMP GAM regression. PLoS One, 20(7), e0328109.

20. Estak Ahmed, An Thi Phuong Nguyen, Aleya Akhter, KAMRUN NAHER, & HOSNE ARA MALEK. (2025). Advancing U.S. Healthcare with LLM–Diffusion Hybrid Models for Synthetic Skin Image Generation and Dermatological AI. *Journal of Medical and Health Studies*, *6*(5), 83-90. https://doi.org/10.32996/jmhs.2025.6.5.11

21. Nitu, F. N., Mia, M. M., Roy, M. K., Yezdani, S., FINDIK, B., & Nipa, R. A. (2025). Leveraging Graph Neural Networks for Intelligent Supply Chain Risk Management in the Era of Industry 4.0. *International Interdisciplinary Business Economics Advancement Journal*, *6*(10), 21-33.

22. Siddique, M. T., Uddin, M. N., Gharami, A. K., Khan, M. S., Roy, M. K., Sharif, M. K., & Chambugong, L. (2025). A Deep Learning Framework for Detecting Fraudulent Accounting Practices in Financial Institutions. *International Interdisciplinary Business Economics Advancement Journal*, *6*(10), 08-20.

23. Mia, M. M., Al Mamun, A., Ahmed, M. P., Tisha, S. A., Habib, S. A., & Nitu, F. N. (2025). Enhancing Financial Statement Fraud Detection through Machine Learning: A Comparative Study of Classification Models. *Emerging Frontiers Library for The American Journal of Engineering and Technology*, *7*(09), 166-175.

24. Akhi, S. S., Ahamed, M. I., Alom, M. S., Rakin, A., Awal, A., & Al Mamoon, I. (2025, July). Boosted Forest Soft Ensemble of XGBoost, Gradient Boosting, and Random Forest with Explainable AI for Thyroid Cancer Recurrence Prediction. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

25. Alom, M. S., Akhi, S. S., Borsha, S. N., Mia, N., Tamim, F. S., & Nabin, J. A. (2025, July). Federated Machine Learning for Cardiovascular Risk Assessment: A Decentralized XGBoost Approach. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

26. Akhi, S. S., Rahaman, M. A., & Alom, M. S. An Explainable and Robust Machine Learning Approach for Autism Spectrum Disorder Prediction.

27. Rabbi, M. A., Rijon, R. H., Akhi, S. S., Hossain, A., & Jeba, S. M. (2025, January). A Detailed Analysis of Machine Learning Algorithm Performance in Heart Disease Prediction. In *2025 4th International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 259-263). IEEE.

28. Mujiba Shaima, Mazharul Islam Tusher, Estak Ahmed, Sharmin Sultana Akhi, & Rayhan Hassan Mahin. (2025). Machine Learning Techniques and Insights for Cardiovascular or Heart Disease Prediction. *Academic International Journal of Engineering Science*, *3*(01), 22-35.

29. Jamee, S. S., Arif, M., Rahman, M. M., YASSAR, I. S., & Hossain, M. A. (2025). Integrating Large Language Models with Machine Learning for Explainable Banking Security and Financial Risk Assessment. *International Interdisciplinary Business Economics Advancement Journal*, *6*(11), 8-18.