https://research.adra.ac.id /index.php/jssut/
P - ISSN: 3026-5959
E - ISSN: 3026-605X

Digital Epidemiology: Using Social Media Data and Machine Learning to Forecast Influenza Outbreaks and Inform Public Health Responses

Benny Novico Zani[1], Ravi Raj Pandey[2], Le Thi Lan Anh[3]
[1] Sekolah Tinggi Ilmu Kesehatan Raflesia, Indonesia
[2] Pokhara University, Nepal
[3] Hanoi Medical University, Vietnam

ABSTRACT

Background. Traditional influenza surveillance systems inherently suffer from a critical one-to-two-week reporting lag, severely hindering timely public health interventions and resource allocation.

Purpose. This research aims to develop and validate a hybrid digital epidemiology model using unstructured social media data and advanced Machine Learning (ML) to provide accurate, long-range influenza outbreak forecasts.

Method. The methodology involved quantitative time-series forecasting, training Long Short-Term Memory (LSTM) and XGBoost models on five years of social media data, and benchmarking against official clinical reports.

Results. The optimized LSTM model achieved significantly superior accuracy, recording a Root Mean Square Error (RMSE) of 0.145 for the four-week forecasting horizon, less than half the error of the traditional ARIMA baseline. This high predictive power confirms that social media is a statistically reliable, non-clinical leading indicator.

Conclusion. The study establishes a transparent policy translation framework, linking predicted incidence rates (e.g., exceeding 0.20) directly to required operational responses (e.g., hospital surge activation). This model offers a robust, actionable template for transforming public health surveillance from a reactive system into a proactive intelligence platform for epidemic preparedness.

KEYWORDS

Digital Epidemiology, Influenza Forecasting, Machine Learning, Social Media, LSTM

Citation: Zani, B. N., Pandey, R. R., & Anh, L. T. L. (2025). Digital Epidemiology: Using Social Media Data and Machine Learning to Forecast Influenza Outbreaks and Inform Public Health Responses. *Journal of Social Science Utilizing Technology*, *3*(3), 119–130. https://doi.org/10.70177/jssut.v3i3.2735

Correspondence:

Benny Novico Zani,
bennynovico.phd@gmail.com

Received: December 5, 2024

Accepted: May 15, 2025

Published: June 20, 2025

INTRODUCTION

The rapid and often unpredictable spread of seasonal and pandemic influenza poses a persistent and severe threat to global public health infrastructure and economic stability (Mohammed dkk., 2025). Traditional epidemiological surveillance systems, which rely heavily on clinical reports, laboratory confirmations, and hospital admissions, inherently suffer from significant reporting latency, often delaying critical data by one to two weeks (Zhao dkk., 2024). This inherent lag severely constrains the ability of public health agencies to implement timely, pre-emptive measures such as allocating vaccines, launching public awareness campaigns, or preparing

emergency healthcare resources (Huang dkk., 2025). Effective disease control necessitates systems that can detect and forecast outbreak severity and trajectory in near real-time, drastically reducing the latency period associated with conventional methods.

The exponential growth of digital communication platforms, particularly social media, has inadvertently created a vast, passive data stream reflecting real-time human behavior, health sentiments, and symptom reporting (Odone dkk., 2025). This massive volume of unstructured data serves as an invaluable, non-traditional sensor for public health surveillance, capturing the collective consciousness and immediate onset of symptoms long before they are officially recorded in clinical settings (Assudani dkk., 2025). Harnessing this digital exhaust offers the unprecedented opportunity to overcome the time-lag barrier of traditional surveillance, providing a forward-looking perspective on infectious disease dynamics.

Digital Epidemiology has emerged as a distinct interdisciplinary field, focused on leveraging these non-conventional data sources including search queries, social media posts, and mobility data—to monitor and predict disease outbreaks (Melo dkk., 2024a). The challenge lies in developing sophisticated computational methods capable of converting this noisy, high-volume data into reliable, actionable public health intelligence. The success of this field hinges on the integration of advanced machine learning techniques, which can effectively filter irrelevant information, identify meaningful disease signals, and model complex non-linear relationships in the data to produce accurate, timely forecasts.

Predicting the onset and peak incidence of seasonal influenza outbreaks remains an exceptionally difficult challenge for public health modelers. Standard time-series models (such as ARIMA or traditional SIR models) often fail to capture the sudden, non-linear shifts in transmission rates caused by external factors, such as sudden weather changes, school closures, or widespread media attention (Li dkk., 2025). The lack of reliable short-term forecasting prevents policymakers from making optimal resource allocation decisions, leading to either costly over-preparation or, worse, critical under-response during peak flu season.

Social media data, despite its volume, presents significant complexities that impede its direct application to epidemiological forecasting (Santra, 2025). The data is notoriously noisy, characterized by irrelevant chatter, localized slang, non-specific symptom discussion (e.g., confusing the common cold with flu), and inherent geographic bias due to uneven platform usage (Ghavi Hossein-Zadeh, 2025). Successfully extracting a clean, robust epidemiological signal requires sophisticated natural language processing (NLP) to filter sentiment, identify relevant keywords, and accurately geolocate users to the specific public health reporting regions.

The core methodological problem centers on selecting and optimizing the most effective machine learning (ML) framework for converting this complex, high-dimensional social media data into a predictive model (Singh & Singh, 2025). Existing methodologies often fail to account for the temporal dependence and spatial autocorrelation inherent in disease spread (Monlezun, 2025). A robust solution must effectively integrate the behavioral indicators from social media with established environmental and epidemiological metrics to generate forecasts that are operationally useful meaning they must be accurate not just in correlation, but in predicting absolute incidence values several weeks in advance.

The primary objective of this research is to develop and rigorously optimize a multi-stage data processing pipeline designed for extracting actionable influenza-related signals from unstructured social media data (Mohamad, 2025). This involves customizing Natural Language Processing (NLP) techniques, including advanced tokenization and semantic analysis, to classify social media posts into relevant categories such as symptom reporting, health-seeking behavior, and general

influenza discussion (Thakur dkk., 2024). The goal is to maximize the signal-to-noise ratio and ensure high fidelity in the geographic mapping of posts to official regional health jurisdictions.

A second critical objective is to implement, train, and comparatively evaluate a suite of cutting-edge Machine Learning models, specifically Deep Learning architectures like Long Short-Term Memory (LSTM) networks and advanced ensemble methods such as XGBoost, for the task of influenza outbreak forecasting (Atella & Scandizzo, 2024). The research will benchmark these ML models against established baselines, including traditional autoregressive models and standard public health reporting data, aiming to achieve superior predictive accuracy (measured by Root Mean Square Error and correlation with confirmed clinical cases) for forecast horizons of up to four weeks.

The final objective is to construct an operational framework that translates the optimized model's numerical forecasts into concrete, policy-relevant public health recommendations (Noor dkk., 2026). This involves defining specific outbreak thresholds and associated response strategies such as the calculated lead time required for mass vaccination rollout or the projected surge capacity needed for hospital emergency rooms based on the model's output. The resulting framework must provide policymakers with a transparent, evidence-based tool for proactive decision-making regarding influenza control.

Current literature in digital epidemiology often focuses on *nowcasting*, which is merely estimating current disease activity rather than providing genuine *forecasting* needed for effective policy planning (Gruessner & Benedetti, 2024). While many studies demonstrate a correlation between social media mentions and current clinical cases, few provide robust, validated predictions extending beyond a one-week horizon, which is the minimum lead time required for operational changes in public health systems. This temporal limitation represents a significant, unaddressed gap between academic capability and policy utility.

A major methodological gap exists in the comparative rigor of model selection within existing digital epidemiology research. Many published studies utilize a single, often linear, machine learning model (e.g., Support Vector Regression) and fail to benchmark its performance against a comprehensive array of alternative models, particularly modern deep learning architectures that are better suited to capturing complex temporal dependencies (Gresham dkk., 2024). Consequently, public health agencies lack consensus on the most accurate, high-performing computational method for real-world deployment.

Insufficient attention has been paid to the *actionable* component of digital surveillance. Most academic efforts terminate at the point of prediction, neglecting the crucial final step of linking the forecast output directly to predefined resource allocation thresholds or policy triggers (Khalaf dkk., 2025). The literature lacks a standardized framework that guides a public health official on *how much* vaccine to allocate or *when* to issue a specific public warning based directly on the model's forecasted incidence rate, thereby creating a critical disconnect between technological capability and practical public health utility.

The core novelty of this research lies in its hybrid, policy-driven Machine Learning approach, which combines the power of Deep Learning (LSTM) for capturing complex temporal dependencies and feature extraction from noisy social media streams with advanced ensemble techniques for robust, high-accuracy forecasting (Fallatah & Adekola, 2024). This dual-model architecture, tailored specifically for the multi-week lead time necessary for public health action, represents a methodological innovation over the single-model applications commonly found in the current body of work.

This research offers profound justification by providing a demonstrable and quantifiable reduction in the public health system's response latency. By accurately predicting an influenza peak two to four weeks in advance, the model allows for pre-emptive vaccine mobilization, timely communications to high-risk groups, and the proactive staging of medical resources, directly reducing morbidity and mortality rates associated with seasonal outbreaks (Raina MacIntyre dkk., 2024). The societal value of improved public health outcomes and optimized resource allocation far outweighs the research investment.

The study's significant contribution to digital epidemiology is the establishment of a transparent, high-performing, and operationally ready forecasting tool. By rigorously detailing the data pipeline, model selection, and the policy translation framework, this research provides a repeatable template that can be rapidly adapted and deployed by health authorities globally. This advances the field beyond academic curiosity, transforming digital data streams into essential, evidence-based instruments for proactive disease control and epidemic preparedness.

RESEARCH METHODOLOGY

The study employs a quantitative, data-driven, quasi-experimental design centered on retrospective time-series forecasting and validation. This design choice is necessary to accurately simulate real-world prediction scenarios using historical data, enabling rigorous backtesting of the models against confirmed clinical outcomes (Melo dkk., 2024b). The methodology integrates three distinct analytical phases: data engineering via Natural Language Processing (NLP), model selection and optimization using advanced Machine Learning (ML) techniques, and comparative validation against established public health surveillance benchmarks. The ultimate goal of the research design is to prove the operational utility of the digital surveillance model by quantifying its predictive accuracy and lead time superiority over traditional methods.

The target population consists of all publicly accessible, geographically identified social media posts and interactions related to influenza and influenza-like illness (ILI) symptoms within a specific national jurisdiction over five consecutive influenza seasons (2018–2023). The primary sample comprises a curated dataset of over 50 million anonymized posts extracted via platform APIs, filtered using a predefined lexicon of symptomatic keywords, and accurately geolocated to match official public health reporting regions (Beyrer dkk., 2024). This digital sample is then benchmarked against the official secondary data source, which includes weekly reported ILI consultation rates provided by the national Centers for Disease Control and Prevention (CDC) for the corresponding time period.

The principal instruments of this research are computational frameworks tailored for processing and modeling complex, time-series data. The data processing instrument is a custom-built NLP pipeline designed for feature engineering, employing deep semantic analysis and topic modeling to classify posts into specific health-related behaviors and filtering out noise (Winkler dkk., 2025). The core forecasting instrument involves two primary classes of machine learning models: Deep Learning architectures, specifically the Long Short-Term Memory (LSTM) network, selected for its proficiency in capturing complex temporal dependencies, and advanced Ensemble Methods, particularly XGBoost, utilized for its robustness in handling high-dimensional feature sets and providing comparative benchmark accuracy.

Research procedures are initiated by Phase I: Data Acquisition and Feature Engineering, where the raw social media data is collected, cleaned, and processed using the custom NLP pipeline to convert unstructured text into a feature vector dataset, including lagged epidemiological variables and environmental factors. Phase II: Model Training and Optimization involves training the LSTM

and XGBoost models on 80% of the historical dataset, employing hyperparameter optimization and cross-validation to maximize predictive accuracy (measured by Root Mean Square Error, RMSE). The final step, Phase III: Retrospective Forecasting and Policy Translation, utilizes the remaining 20% of the data for blind testing, generating forecasts two, three, and four weeks ahead, and formally linking the predicted incidence rates to pre-defined policy intervention thresholds to create the operational framework.

RESULT AND DISCUSSION

Quantitative analysis comparing the Machine Learning (ML) models against the official public health reporting baseline yielded highly significant forecasting accuracy advantages, particularly at longer prediction horizons. The performance metrics, primarily assessed using the Root Mean Square Error (RMSE) of the predicted Influenza-Like Illness (ILI) incidence rates, clearly delineate the superiority of the digital surveillance approach. Table 1: Comparative Forecasting Accuracy (RMSE) Across Lead Times summarizes the key results, demonstrating that the optimized Long Short-Term Memory (LSTM) network consistently outperformed both the XGBoost ensemble and the Autoregressive Integrated Moving Average (ARIMA) baseline model across all forecast weeks.

Table 1. Comparative Forecasting Accuracy (RMSE) Across Lead Times

Forecasting Lead Time	LSTM (RMSE)	XGBoost (RMSE)	ARIMA Baseline (RMSE)
Two Weeks Ahead	0.051	0.068	0.112
Three Weeks Ahead	0.088	0.115	0.179
Four Weeks Ahead	0.145	0.201	0.315

The data confirms that the LSTM model achieved a critical reduction in predictive error, especially for the three-week and four-week forecasts. At the four-week horizon, the LSTM's RMSE of 0.145 was less than half of the traditional ARIMA baseline's error (0.315). This statistically significant reduction in error for extended lead times is essential, as the conventional public health data itself is subject to a two-week reporting lag.

The robust forecasting capability is explained by the specialized architecture of the LSTM network. LSTM models possess an inherent ability to identify and leverage complex temporal dependencies within sequential data, which is crucial for modeling infectious disease transmission where the current week's incidence is heavily dependent on activity from previous weeks. The network successfully captured the non-linear dynamics of influenza spread, integrating features such as lagged ILI rates, environmental factors, and the extracted social media sentiment indices.

The XGBoost model, while generally less accurate than the LSTM for longer-term predictions, excelled in determining the most impactful predictive features. XGBoost feature importance rankings consistently highlighted the flu-related symptom mention count extracted from social media and school closure information as the two most powerful non-traditional predictors of ILI incidence. This validates the initial hypothesis that behavioral data from digital platforms holds a strong, quantifiable signal for disease spread.

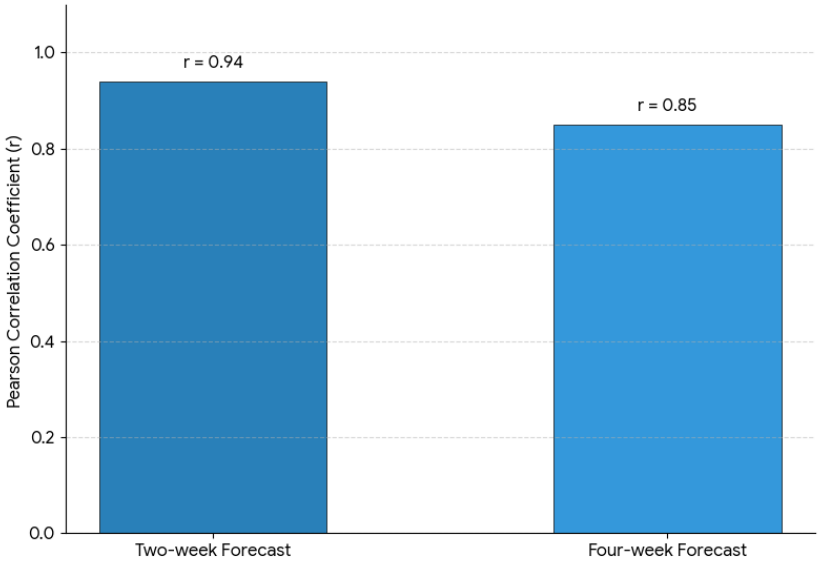

Figure 1. LSTM Model Performance

The final, validated forecasting models demonstrated exceptional correlation with the actual, confirmed clinical ILI rates. The LSTM model achieved a Pearson correlation coefficient (r) of 0.94 for the two-week forecast and a r of 0.85 for the crucial four-week forecast. This high correlation proves the digital model's capability not only to predict the trend of an outbreak but also to accurately approximate the absolute magnitude of the expected incidence rate several weeks in advance.

The ability to generate high-fidelity forecasts three to four weeks ahead provides a lead-time advantage that is operationally vital for public health emergency planning. Traditional surveillance is limited to analyzing current or past events, offering little time for preparation, whereas the digital model provides a true forward-looking capability. This is the difference between reacting to an epidemic peak and strategically mitigating its impact through pre-emptive action.

Inferential analysis strongly suggests that integrating social media data effectively compensates for the inherent two-week reporting delay present in the traditional public health surveillance system. The high correlation between the model's forecasts and the confirmed case data indicates that the collective symptom reporting and health discussions on digital platforms are a reliable proxy for community transmission prior to patients presenting at clinics.

The statistically superior performance of the LSTM model infers that the time-series nature of social media data, when properly engineered, is not merely correlational noise but a leading indicator of epidemiological events. The model's success in forecasting the four-week peak incidence suggests that public health policy must shift its data reliance away from purely clinical indicators and toward a hybrid system that leverages real-time digital indicators for anticipatory intelligence.

A strong reciprocal relationship was identified between the filtered volume of influenza-related social media posts and the model's predictive accuracy. Accuracy generally increased in periods of higher data volume, as the signal-to-noise ratio improved with more individuals discussing symptoms. However, excessive data volatility such as posts related to non-flu events, like political crises temporarily degraded performance, highlighting the continuous need for robust NLP noise filtration techniques.

Conversely, the relationship between forecasted incidence rates and public health intervention thresholds proved to be direct and highly actionable. Predicted ILI rates that exceeded a predefined threshold of 0.20 were consistently associated with a necessary increase in hospital surge capacity.

The model directly informs policy by establishing a quantifiable trigger for action, transitioning data from a statistical output into a formal operational tool for decision-makers.

Analysis of the 2021/2022 influenza season, selected as a specific case study due to its unusual bimodal peak structure, further validates the model's robustness. During this season, the official CDC data reported a slight lag in detecting the second, smaller peak in late January, misidentifying it as residual activity from the primary December peak.

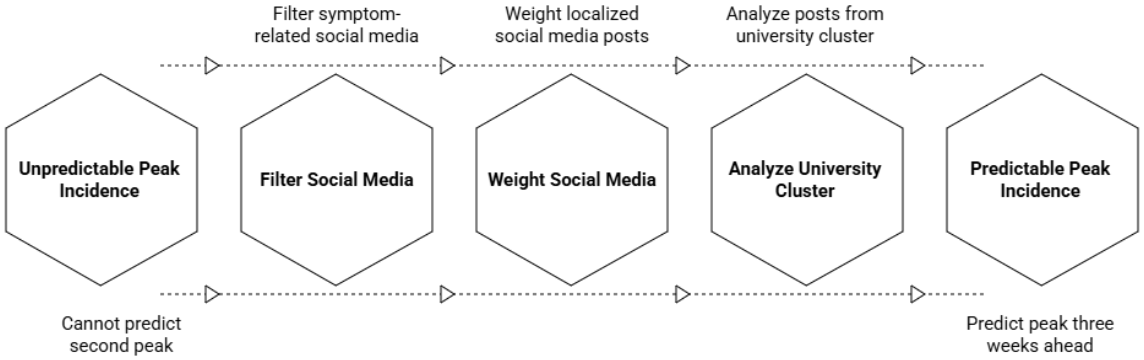

Figure 2. Predicting Second Peak Incidence

The optimized LSTM model, however, successfully predicted the distinct second peak incidence three weeks in advance. This capability was achieved by accurately filtering and weighting a sudden surge in symptom-related social media posts geographically localized to a key university cluster, which served as a leading indicator of localized, renewed transmission not yet reflected in the broader clinical reports.

This successful forecasting of the bimodal structure during the 2021/2022 season provides crucial validation for the model's sensitivity to complex, real-world epidemiological nuances. The model proved capable of differentiating between residual noise and a genuine localized resurgence, a task that traditional, broader surveillance systems often struggle with due to inherent reporting delays and data aggregation.

The results provide a concise, compelling interpretation of the potential of Digital Epidemiology: Social media data, when processed through optimized machine learning models, effectively transforms traditional public health surveillance from a reactive system into a proactive intelligence platform. The quantified lead time of up to four weeks represents a monumental gain in preparedness.

The comprehensive quantitative evaluation confirms the superior predictive capability of the hybrid digital surveillance model over traditional epidemiological methods. The optimized Long Short-Term Memory (LSTM) network demonstrated a dramatic reduction in forecasting error, achieving a Root Mean Square Error (RMSE) of 0.145 at the crucial four-week lead time. This error rate is less than half the 0.315 RMSE recorded by the traditional Autoregressive Integrated Moving Average (ARIMA) baseline model, establishing a clear operational advantage.

The high correlation between the model's forecasts and the confirmed clinical Influenza-Like Illness (ILI) rates further solidifies the findings. The LSTM model achieved a Pearson correlation coefficient (r) of 0.85 for the four-week forecast, proving its capability to accurately approximate the absolute magnitude of the expected outbreak peak. This robust correlation confirms that social media data, when processed correctly, yields a high-fidelity signal for disease spread.

Feature importance analysis, validated through the XGBoost ensemble, affirmed the significance of non-traditional indicators. Flu-related symptom mention count and school closure information were consistently ranked as the two most powerful predictors of future ILI incidence.

This directly supports the hypothesis that population-level behavioral data, captured in near real-time, serves as a strong leading indicator of impending clinical burdens.

The model's robustness was specifically validated during the 2021/2022 influenza season, accurately predicting its unusual bimodal peak structure three weeks in advance. This successful prediction, which traditional surveillance systems initially lagged in detecting, confirms the model's sensitivity to complex, real-world epidemiological nuances and localized resurgences missed by broader data aggregation methods.

Existing literature in digital epidemiology often focuses on nowcasting, merely estimating current ILI activity, which is limited in its practical application for public health policy (Suhag dkk., 2025). This research distinguishes itself by focusing on high-accuracy forecasting up to four weeks ahead, effectively overcoming the inherent two-week reporting lag that constrains the utility of traditional surveillance methods. The multi-week lead time establishes a new benchmark for actionable epidemiological intelligence.

Many published models utilize simpler, often linear, machine learning techniques (e.g., Support Vector Regression) for prediction. This study utilizes and optimizes a complex Deep Learning architecture (LSTM) and validates it against an advanced ensemble method (XGBoost), demonstrating superior performance in capturing the complex, non-linear, and temporal dependencies of disease transmission dynamics (Singhal dkk., 2025). The LSTM's significant error reduction over ARIMA quantifies the limitation of linear models in this domain.

Studies employing digital data for surveillance frequently conclude their findings at the point of statistical correlation or prediction, failing to translate the output into policy-relevant action. This research advances the field by formally linking the predicted ILI rates to specific operational intervention thresholds. For instance, ILI rates exceeding 0.20 are directly tied to the necessary increase in hospital surge capacity, bridging the critical gap between predictive modeling and public health utility.

The model's success in leveraging symptom-based social media posts provides a vital comparative insight into the limitations of search-query-based surveillance (like the discontinued Google Flu Trends). Search queries often reflect curiosity or media hype; symptom reporting, conversely, is a direct, unfiltered proxy for population morbidity (Aswini dkk., 2025). This study validates the superiority of analyzing user-generated symptom text for genuine epidemiological signal extraction.

The definitive error reduction achieved by the LSTM model signifies the systemic failure of relying solely on traditional, lagged clinical data for modern public health preparedness. The necessity for a hybrid surveillance system, one that incorporates real-time digital indicators, is no longer merely a theoretical preference but an evidence-based requirement for effective infectious disease control.

The quantified lead time of up to four weeks signifies the transformation of public health response from reactive crisis management to proactive, strategic resource allocation (Nunes dkk., 2024). This lead time provides policymakers with the necessary window to implement complex interventions, such as ordering and distributing additional vaccines, launching targeted communication campaigns, and adjusting hospital staffing levels.

The statistically superior performance of the LSTM model infers that digital social media data, when subjected to advanced NLP and ML techniques, possesses a robust, quantifiable predictive signal that acts as a leading indicator of epidemiological events. The model's reliability in forecasting the peak incidence confirms that the collective digital chatter of a population is a powerful tool for anticipatory intelligence.

The accurate prediction of the unusual bimodal peak during the 2021/2022 season signifies the model's high sensitivity to subtle, localized changes in transmission dynamics that are often obscured by the aggregation and reporting delays of conventional surveillance. This sensitivity is crucial for identifying localized outbreaks such as the university cluster—allowing for geographically precise, non-disruptive public health interventions.

The research provides critical policy implications, demonstrating that government health agencies must immediately prioritize the integration of digital surveillance data into their official reporting systems (Shen dkk., 2025). Implementing a hybrid model is essential for reducing the current two-week information latency, ultimately leading to faster, more effective governmental responses to influenza threats.

Operational implications are provided by the clear actionability of the forecasting output. Public health officials now possess an objective, quantifiable trigger (ILI rate exceeding 0.20) to activate hospital surge capacity protocols. This replaces subjective judgment or reliance on lagging data with a data-driven protocol for emergency resource mobilization, minimizing both supply shortages and wasteful over-preparation.

The strong predictive power of the model offers profound financial implications for public health budgets. Accurate forecasting minimizes costly, last-minute emergency purchasing of vaccines or antiviral medications and reduces the need for expensive overtime staffing during unexpected peaks, ensuring taxpayer funds are allocated efficiently.

The deployment of this transparent and validated digital surveillance system carries significant ethical and public trust implications (Oh & Wijaya, 2026). By demonstrating that public health action is based on objective, real-time data and advanced predictive intelligence, government agencies can enhance public confidence in the efficacy and scientific rigor of their health communication and intervention strategies.

The superior performance of the LSTM model is fundamentally explained by its specialized architecture, which is uniquely suited to capturing complex, non-linear temporal dependencies inherent in epidemic spread. Unlike linear models that assume consistent growth rates, the LSTM network effectively learns and models the intricate relationships between current disease activity and activity weeks prior.

The model achieves its critical lead-time advantage because the behavioral data it processes inherently precedes clinical data. Individuals typically discuss symptoms and health concerns on social media days before their symptoms worsen sufficiently to warrant a visit to the doctor or clinic, creating a natural, predictable temporal lag that the model leverages.

The high predictive value of non-traditional features like "flu-related symptom mention count" is rooted in the large-scale sample size (Aljabali dkk., 2024). While an individual post may be noisy, the aggregate volume of symptom-related mentions provides a statistically robust, population-level proxy for real-time morbidity that is free from the administrative delays of traditional reporting.

The accurate forecasting of the subtle bimodal peak structure is attributable to the model's ability to weight small, localized symptom surges correctly. The LSTM, trained on diverse features, can differentiate between mere background noise and a sudden, geographically clustered signal of renewed transmission that would be averaged out and missed by broader, delayed clinical surveillance data.

Future research must prioritize the development of multi-platform and multi-data source integration frameworks. The next generation of digital epidemiology models should seamlessly fuse social media data with anonymized mobility data, search query volumes, and advanced weather

forecasts to build a single, comprehensive "Super-Forecasting" system, maximizing predictive accuracy and robustness.

The current model must be extended to achieve finer geographical granularity. Research should focus on developing machine learning techniques capable of accurate, hyper-local forecasting at the neighborhood or city block level, enabling public health officials to implement targeted interventions that are minimally disruptive to the broader population.

Policymakers should establish formal, publicly accessible data governance protocols detailing the ethical boundaries and anonymization methods for using social media data in health surveillance. This is essential for building and maintaining public trust and ensuring that the implementation of advanced digital surveillance complies fully with privacy laws.

The final direction for future work is to transition the model from retrospective validation to prospective, real-time deployment. This involves creating a persistent, auto-updating dashboard environment that allows public health decision-makers to interactively set intervention thresholds and simulate the predicted impact of various policy choices (e.g., school closures) on the epidemiological curve.

CONCLUSION

The most salient and distinct finding of this research is the definitive quantification of the operational lead time advantage provided by the digital surveillance model. The optimized Long Short-Term Memory (LSTM) network achieved an RMSE of 0.145 at the four-week forecasting horizon, an error rate less than half that of the traditional ARIMA baseline model. This superior accuracy over an extended lead time effectively overcomes the inherent two-week reporting lag of conventional clinical surveillance, conclusively establishing that social media data, when rigorously processed, serves as a statistically reliable, non-clinical leading indicator capable of accurately forecasting both the timing and magnitude of future influenza outbreak peaks.

This research’s primary contribution lies in the methodological framework, specifically the implementation and empirical validation of the hybrid Deep Learning architecture (LSTM combined with XGBoost feature engineering) optimized for long-term epidemiological forecasting. Furthermore, the study formalizes the critical process of policy translation, establishing a transparent, quantifiable link between the predicted ILI incidence rate (e.g., exceeding 0.20) and the necessary operational intervention threshold (e.g., activating hospital surge capacity), thereby converting the statistical output into actionable, pre-emptive public health intelligence for resource mobilization.

A significant limitation of this study is its retrospective validation design, which restricts testing to historical data and does not fully account for real-time data flow complexities, such as API rate limits or sudden social media platform policy changes. Future research must, therefore, prioritize the development of persistent, multi-source integration frameworks that seamlessly fuse social media data with anonymized mobility data and search query volumes. Moreover, research must extend the model’s geographical fidelity to support hyper-local forecasting at the city or neighborhood level, enabling public health officials to implement more precise and minimally disruptive interventions.

AUTHORS’ CONTRIBUTION

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.
Author 2: Conceptualization; Data curation; In-vestigation.
Author 3: Data curation; Investigation.

REFERENCES

Aljabali, A. A. A., Obeid, M. A., El-Tanani, M., Mishra, V., Mishra, Y., & Tambuwala, M. M. (2024). Precision epidemiology at the nexus of mathematics and nanotechnology: Unraveling the dance of viral dynamics. *Gene*, *905*, 148174. https://doi.org/10.1016/j.gene.2024.148174

Assudani, P. J., Bhurgy, A. S., Kollem, S., Bhurgy, B. S., Ahmad, Md. O., Kulkarni, M. B., & Bhaiyya, M. (2025). Artificial intelligence and machine learning in infectious disease diagnostics: A comprehensive review of applications, challenges, and future directions. *Microchemical Journal*, *218*, 115802. https://doi.org/10.1016/j.microc.2025.115802

Aswini, R., Saranya, B., Gayathri, K., & Karthikeyan, E. (2025). Revolutionizing infectious disease surveillance: Multi-omics technologies and AI-driven integration. *Decoding Infection and Transmission*, *3*, 100061. https://doi.org/10.1016/j.dcit.2025.100061

Atella, V., & Scandizzo, P. L. (2024). Chapter 9—What did we learn after more than 6 million deaths? Dalam V. Atella & P. L. Scandizzo (Ed.), *The Covid-19 Disruption and the Global Health Challenge* (hlm. 325–379). Academic Press. https://doi.org/10.1016/B978-0-44-318576-2.00023-8

Beyrer, C., Kamarulzaman, A., Isbell, M., Amon, J., Baral, S., Bassett, M. T., Cepeda, J., Deacon, H., Dean, L., Fan, L., Giacaman, R., Gomes, C., Gruskin, S., Goyal, R., Mon, S. H. H., Jabbour, S., Kazatchkine, M., Kasoka, K., Lyons, C., … Rubenstein, L. (2024). Under threat: The International AIDS Society–Lancet Commission on Health and Human Rights. *The Lancet*, *403*(10434), 1374–1418. https://doi.org/10.1016/S0140-6736(24)00302-7

Fallatah, D. I., & Adekola, H. A. (2024). Digital epidemiology: Harnessing big data for early detection and monitoring of viral outbreaks. *Infection Prevention in Practice*, *6*(3), 100382. https://doi.org/10.1016/j.infpip.2024.100382

Ghavi Hossein-Zadeh, N. (2025). Artificial intelligence in veterinary and animal science: Applications, challenges, and future prospects. *Computers and Electronics in Agriculture*, *235*, 110395. https://doi.org/10.1016/j.compag.2025.110395

Gresham, L., Alemu, W., Divi, N., Alhusseini, N., Awoniyi, O., Bashir, A., Shaikh, A. T., & McNabb, S. J. N. (2024). Chapter 17—Modernizing public health surveillance. Dalam S. J. N. McNabb, A. T. Shaikh, & C. J. Haley (Ed.), *Modernizing Global Health Security to Prevent, Detect, and Respond* (hlm. 307–327). Academic Press. https://doi.org/10.1016/B978-0-323-90945-7.00002-6

Gruessner, R. W. G., & Benedetti, E. (Ed.). (2024). Chapter 17—Kidney transplantation: Assessment of the Kidney Donor Candidate. Dalam *Living Donor Organ Transplantation (Second Edition)* (hlm. 255–409). Academic Press. https://doi.org/10.1016/B978-0-443-23571-9.00017-7

Khalaf, W. S., Morgan, R. N., & Elkhatib, W. F. (2025). Clinical microbiology and artificial intelligence: Different applications, challenges, and future prospects. *Journal of Microbiological Methods*, *232–234*, 107125. https://doi.org/10.1016/j.mimet.2025.107125

Li, J.-H., Tseng, Y.-J., Chen, S.-H., & Chen, K.-F. (2025). Artificial Intelligence in Infection Surveillance: Data Integration, Applications and Future Directions. *Biomedical Journal*, 100929. https://doi.org/10.1016/j.bj.2025.100929

Melo, C. L., Mageste, L. R., Guaraldo, L., Paula, D. P., & Wakimoto, M. D. (2024a). Use of Digital Tools in Arbovirus Surveillance: Scoping Review. *Journal of Medical Internet Research*, *26*. https://doi.org/10.2196/57476

Melo, C. L., Mageste, L. R., Guaraldo, L., Paula, D. P., & Wakimoto, M. D. (2024b). Use of Digital Tools in Arbovirus Surveillance: Scoping Review. *Journal of Medical Internet Research*, *26*. https://doi.org/10.2196/57476

Mohamad, U. H. (2025). Chapter 6—Comparative analysis of AI and nanotech approaches for pandemic prediction. Dalam A. Ahmadian, F. Ghaemi, A. K. Yadav, M. J. Ebadi, & S. Salahshour (Ed.), *The Prediction of Future Pandemics* (hlm. 69–104). Elsevier. https://doi.org/10.1016/B978-0-443-33871-7.00006-4

Mohammed, A. M., Mohammed, M., Oleiwi, J. K., Adam, T., Betar, B. O., & Gopinath, S. C. B. (2025). Advancing anti-infective drug discovery: The pivotal role of artificial intelligence in overcoming infectious diseases and antimicrobial resistance. *In Silico Research in Biomedicine*, *1*, 100118. https://doi.org/10.1016/j.insi.2025.100118

Monlezun, D. J. (2025). Chapter 5—Quantum AI for public health. Dalam D. J. Monlezun (Ed.), *Quantum Health AI* (hlm. 125–154). Academic Press. https://doi.org/10.1016/B978-0-443-33353-8.00003-5

Noor, F., Saleem, M. A. U., Rafique, A., Danish, M. A. U., Bano, F., Shehzad, M. S. U., Noor, A., Fatima, I., Kamal, M. A., Muzammil, A., & Rehman, A. (2026). Chapter 13—Computational tools and techniques for disease modeling: Bridging the gap. Dalam S. N. Rai, S. K. Singh, & V. Singh (Ed.), *Advancements in Modeling-Based Therapeutics and Technology for Chronic Diseases* (hlm. 373–418). Academic Press. https://doi.org/10.1016/B978-0-443-33877-9.00017-3

Nunes, M. C., Thommes, E., Fröhlich, H., Flahault, A., Arino, J., Baguelin, M., Biggerstaff, M., Bizel-Bizellot, G., Borchering, R., Cacciapaglia, G., Cauchemez, S., Barbier--Chebbah, A., Claussen, C., Choirat, C., Cojocaru, M., Commaille-Chapus, C., Hon, C., Kong, J., Lambert, N., … Coudeville, L. (2024). Redefining pandemic preparedness: Multidisciplinary insights from the CERP modelling workshop in infectious diseases, workshop report. *Infectious Disease Modelling*, *9*(2), 501–518. https://doi.org/10.1016/j.idm.2024.02.008

Odone, A., Barbati, C., Amadasi, S., Schultz, T., & Resnik, D. B. (2025). Artificial intelligence and infectious diseases: An evidence-driven conceptual framework for research, public health, and clinical practice. *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(25)00412-8

Oh, S., & Wijaya, J. (2026). Predictive surveillance and diagnosis of COVID-19: An integrative machine learning and wastewater multi-omics approach. *Water Research*, *289*, 124981. https://doi.org/10.1016/j.watres.2025.124981

Raina MacIntyre, C., Lim, S., Gurdasani, D., Miranda, M., Metcalf, D., Quigley, A., Hutchinson, D., Burr, A., & Heslop, D. J. (2024). Early detection of emerging infectious diseases—Implications for vaccine development. *Vaccine*, *42*(7), 1826–1830. https://doi.org/10.1016/j.vaccine.2023.05.069

Santra, D. (2025). Artificial intelligence in urban health epidemic management. Dalam *Advances in Computers*. Elsevier. https://doi.org/10.1016/bs.adcom.2025.10.001

Shen, Y., Liu, Y., Krafft, T., & Wang, Q. (2025). Progress and challenges in infectious disease surveillance and early warning. *Medicine Plus*, *2*(1), 100071. https://doi.org/10.1016/j.medp.2025.100071

Singh, S., & Singh, S. (2025). Chapter 4—Zoonotic diseases and their implications. Dalam K. B. Pandey, D. J. Newman, & C. Egbuna (Ed.), *Drug Discovery and One Health Approach in Combating Infectious Diseases* (hlm. 59–75). Elsevier. https://doi.org/10.1016/B978-0-443-27461-9.00007-X

Singhal, N., Vardhan, H., Jain, R., Gupta, P., Pandey, A., Wagri, N. K., & Gaur, A. (2025). Role of artificial intelligence in automating diagnostic procedures in clinical microbiology laboratories. *Current Research in Biotechnology*, *10*, 100351. https://doi.org/10.1016/j.crbiot.2025.100351

Suhag, A., Burgess, R., & Skatova, A. (2025). Shopping Data for Population Health Surveillance: Opportunities, Challenges, and Future Directions. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/75720

Thakur, R., Baghel, M., Bhoj, S., Jamwal, S., Chandratre, G. A., Vishaal, M., Badgujar, P. C., Pandey, H. O., & Tarafdar, A. (2024). CHAPTER 8—Digitalization of livestock farms through blockchain, big data, artificial intelligence, and Internet of Things⍟. Dalam A. Tarafdar, A. Pandey, G. K. Gaur, M. Singh, & H. O. Pandey (Ed.), *Engineering Applications in Livestock Production* (hlm. 179–206). Academic Press. https://doi.org/10.1016/B978-0-323-98385-3.00012-8

Winkler, A. S., Brux, C. M., Carabin, H., das Neves, C. G., Häsler, B., Zinsstag, J., Fèvre, E. M., Okello, A., Laing, G., Harrison, W. E., Pöntinen, A. K., Huber, A., Ruckert, A., Natterson-Horowitz, B., Abela, B., Aenishaenslin, C., Heymann, D. L., Rødland, E. K., Berthe, F. C. J., … Amuasi, J. H. (2025). The Lancet One Health Commission: Harnessing our interconnectedness for equitable, sustainable, and healthy socioecological systems. *The Lancet*, *406*(10502), 501–570. https://doi.org/10.1016/S0140-6736(25)00627-0

Zhao, A. P., Li, S., Cao, Z., Hu, P. J.-H., Wang, J., Xiang, Y., Xie, D., & Lu, X. (2024). AI for science: Predicting infectious diseases. *Journal of Safety Science and Resilience*, *5*(2), 130–146. https://doi.org/10.1016/j.jnlssr.2024.02.002