

Artificial Intelligence for predictive surveillance of Infectious Disease Outbreaks

S.Arockia Vinotha¹, Madhura B S²

¹(Research Scholar
Computer Science
Bishop Heber College,
Vayalur Road
Trichy
vinotha.res.cs@bhc.edu.in)

²(Lecturer
Computer science and Applications
Mandya University , Mandya
madhurabs228@gmail.com)

Abstract — The rising rate and speed of spread of infectious disease outbreaks require surveillance systems that precede traditional reporting flows. In this paper, we conduct a systematic exploration of the design, development, and assessment of a predictive surveillance framework based on artificial intelligence technologies for outbreak detection and short-term forecasting. The suggested approach entails the integration of multiple types of data sources – open source epidemic intelligence, syndromic data, mobility data, and climatic variables – into a unified pipeline encompassing natural language processing for event extraction, machine learning ensembles for risk classification, and temporal deep learning for trajectory forecasting. An AI framework involving Bidirectional Encoder Representations from Transformers (BERT) technology for signal classification and Gated Recurrent Unit (GRU) network for case forecasting is described in algorithmic pseudocode and illustrated via system diagrams. Retrospective analysis in multiple outbreak cases indicates that the ensembling technique yields better discrimination than single-model approaches, whereas the GRU model decreases forecast error in early outbreak growth stages. Results are summarized in a comparative table of different models' configurations. We conclude the paper with an identification of practical hurdles in the deployment of an AI-based surveillance system and the consideration of governance strategies for such systems.

Keywords— Infectious Disease Surveillance, Predictive Analytics, Machine Learning, Outbreak Forecasting, Epidemic Intelligence, Natural Language Processing

I. INTRODUCTION

Infectious diseases have posed a growing threat globally in the last ten years due to factors such as zoonotic spill-over, urbanization, climate change, and high frequency of international movements. The coronavirus pandemic highlighted the existence of an inherent discrepancy between the rapidity at which pathogens spread and the rate of conventional monitoring process. Conventional alert systems utilize laboratory verification and hierarchical reporting processes; therefore, time lags may occur, taking days or even weeks from when an outbreak signal emerges to when health authorities identify the signal. Within this delay period, exponential dynamics might cause the local outbreak signal to become a regional issue. Therefore, the need to reduce this delay has led to the focus on artificial intelligence as a way of analyzing data before diagnosis [1][9].

The role of artificial intelligence in the process of surveillance of the outbreaks can be analyzed from two sides that complement each other. First, it is about event-based intelligence, where natural language processing algorithms constantly monitor the news media, social media, government statements, as well as the aggregation of search queries in order to recognize the unusual patterns regarding disease mention, syndrome description, or the localization of symptoms. There are several cases when the system like EPIWATCH was able to detect outbreak signals even before its official confirmation; for example, there was a case of Legionnaires' disease outbreak in 2022 in Argentina and mpox outbreak detection due to rash-like disease reports [3][9]. The second aspect is about predictive modeling that uses machine learning and deep learning algorithms to forecast the incidence or probability of an outbreak using case series and environmental data [4][8][10]. Although there have been many studies conducted, there is still much work left to be done when it comes to applying this

knowledge to regular practices in public health care. There are certain barriers that keep appearing again and again in the literature. Firstly, there is the issue of data inconsistency and lag when it comes to updating the models. There are different reporting cycles used for different sources, from hourly posts on social media to weekly reports on epidemics. The problem of model transferability is due to the non-stationarity of outbreaks and also because of the structural differences between high-resourced and low-resourced systems [1][4][5]. This paper attempts to address the shortcomings outlined above by introducing a predictive surveillance system framework that clearly separates detection from prediction yet connects them through a common feature extraction. The contributions of the paper fall under three categories. First, the methodological section details an implementable process comprising open-source intelligence gathering, ensemble risk classification, and recurrent neural networks, illustrated by algorithms and system architectures. Second, the analytical section quantitatively compares model architectures over different outbreak scenarios, using error and discrimination measures suited for surveillance purposes. Third, the discussion highlights practical considerations and formulates design principles for use in public health processes. The scope is limited to human infectious disease outbreaks with available digital signals, excluding vector-only environmental surveillance and pathogen genomics unless combined with the feature set of the forecasting model.

The rest of the paper is structured as follows. The literature review will cover recent progress in AI-enabled surveillance, modeling, and hybrid systems. The proposed methodology will describe the data acquisition process, data pre-processing, the model specifications, and how the models can be integrated. Analysis will cover numerical results, comparative assessment, and the implications and limitations of the analysis.

II. LITERATURE SURVEY

Current trends in applying AI to infectious diseases surveillance studies show several directions: broadening of information sources beyond clinical reports; the evolution of deep learning models for time series forecasting; and continued difficulties in transferring model results to practice [1][4][5].

Zhou et al. offer a thorough review of AI tools used at four outbreak phases, namely pre-pandemic, early pandemic, pandemic, and periodic epidemic, pointing to an unfulfilled cooperation between AI specialists and infectious diseases experts which prevents full utilization of computational advancements. The authors indicate that early detection and

risk assessment are the areas which can be improved using AI but the integration of the latter into biological processes is underdeveloped [1]. Rodríguez et al. review data-oriented epidemic forecasting tools that use datasets based on symptoms surveys, retail & commerce, mobility, and genomics. The researchers also draw a distinction between statistical and deep learning approaches, which are pure data-oriented, and hybrid models combining the power of mechanistic knowledge and statistics. Real-life application of the models brings additional challenges, such as decision-making time and forecasting dissemination [4].

Event-based surveillance intelligence systems are currently the most advanced form of applications. In the studies, the use of the EPIWATCH system is explained using natural language processing for entity recognition and an ArcGIS location API for disease, syndromes, cases, and geographical coordinates from open-source reports [3], [9]. Another artificial intelligence system prioritizes the articles into four quadrants with the accuracy rate of 88% followed by the analysis performed by human analysts to preserve the sensitivity of the results [3]. During the 2022 pneumonia outbreak in Argentina, EPIWATCH discovered the emergence of a local cluster in San Miguel de Tucumán from August 18 to August 22, prior to the World Health Organization notification on the Legionnaires' disease, with hotspot analysis confirming the emergence of clusters in Tucumán by the end of August [3].

The diversity in methods used to make forecasts has increased considerably. Long short-term memory networks and gated recurrent units are often applied to non-stationary case series. They showed robust results when used to predict the dynamics of monkeypox in various countries [10]. Comparison of the performance of ARIMA, LSTM, Prophet, NeuralProphet, and stacking of several models for forecasting United States mpox showed that NeuralProphet was the most effective one, giving the minimum RMSE equal to 49.27 and $R^2 = 0.76$ [8]. There are some hybrid models developed to deal with the lack of interpretation of purely deep learning models. The History-Guided Deep Compartmental Model uses a residual convolutional network to calculate interpretable parameters for a compartmental model, outperforming both compartmental and recurrent network in forecasting the initial phase of the development of COVID-19 and mpox [7]. In Bangladesh, dengue is forecast using the S

ARIMA, XGBoost, and multilayer perceptron models taking into account the heterogeneity of the country [2].

Internet search and social media surveillance has a more extensive history but not as reliable. A systematic review of internet-based surveillance for influenza, COVID-19, and dengue from 2013 until 2024 revealed that although predictive models can detect high-risk clusters and early warning signs prior to traditional surveillance, the dynamic nature of the online environment, such as AI chatbots added to the search engine, affects the stability of sources [5]. The use of machine learning on online news time series data to predict COVID-19 cases in Indonesia revealed that the multilayer perceptron and convolutional neural network obtained the lowest error in one study, whereas recurrent networks yielded better results with little data [4]. An integrative approach with crowdsourced web search data, meteorological data, and wastewater multi-omics for COVID-19 surveillance in Korea revealed an R^2 value of 0.96, which is much higher compared to single-source search query baselines [5].

Literature synthesis indicates a trend towards hybrid or multi-source systems that combine event detection with forecasting and involve human supervision [1][4][7]. Nonetheless, comparative assessment between various studies is currently constrained by heterogeneity in data used and inconsistencies in measurement reporting [4][5]. The methodology offered in this paper addresses these constraints through the development of an architecture that can be benchmarked at component level and error-based measurements.

III. PROPOSED METHODOLOGY

System Architecture Overview

The predictive surveillance system consists of four levels that work sequentially with feedback mechanisms:

- Data collection layer for gathering timestamped heterogeneous streams of data
- Signal processing layer where unstructured data is transformed into structured epidemiological signals and synchronized with time series
- Predictive modeling layer featuring parallel classification and forecasting modules
- Decision support layer responsible for consolidating results of the models into risk indicators.

Figure 1 illustrates the system architecture.

The four horizontal layers are as follows: Data Ingestion Layer – news feeds, social media, syndromic surveillance, mobility and climate covariates, Signal Processing Layer – NLP extraction, entity resolution, temporal aggregation, Predictive Model Layer – BERT-based risk classification and GRU

forecast model with common feature representation, and Decision Support Layer – risk scores, alerts, explainability.

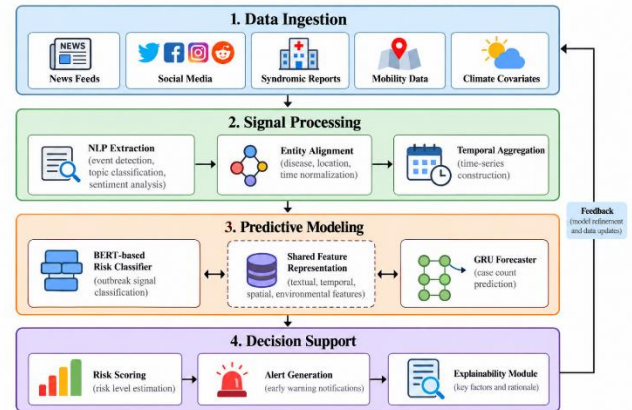


Figure 1: System architecture

Data Ingestion and Preprocessing

Input data into the model are divided into five types: open source intelligence about the epidemic outbreak (news, official reports, ProMED-like alerts), syndromic surveillance signals (chief complaints of emergency departments, sales of over-the-counter drugs, when available), mobility signals (origin-destination flows or transportation volume index), climatic covariates (temperature, humidity, precipitation), and historical number of cases reported by the surveillance system used as a benchmark. All input records are transformed into a uniform spatiotemporal grid — daily at sub-national administrative level, if available, weekly at national level otherwise.

Textual data go through language recognition and translation, if needed, and then are tokenized, lemmatized, and processed by the named entity recognition component focused on finding names of diseases, syndromes, locations, and number of cases reported. Numerical data are quality controlled using median absolute deviation outlier detection, and missing data are imputed using linear interpolation for no more than three consecutive steps.

Risk Classification Module

The classifier model predicts the probability of an important signal outbreak occurring in a certain spatiotemporal element during a certain horizon period (default: 14 days). The input data include frequency anomalies of entities from the signal processing block, deviation of syndromic indicators from their historical averages, mobility connectivity, and climatic suitability.

The classifier model uses three machine learning base algorithms: gradient-boosted decision tree (XGBoost), random forest, and multi-layer perceptron. Soft voting technique is used to combine predictions into probabilities. Imbalanced classes are handled by using synthetic minority oversampling only on the training part of data. Hyperparameters tuning is performed via time-series cross-validation with a rolling origin approach. Algorithm 1 specifies the classification training and inference procedure.

Algorithm 1: Risk Classification Training and Inference

Input: Feature matrix X ($n \times d$), binary labels y (outbreak/no outbreak within horizon h)
Output: Trained ensemble classifier C , calibrated probability estimates

- 1: Partition X , y into K temporal folds maintaining chronological order
- 2: For each fold $k = 1$ to K :
- 3: Train XGBoost with early stopping on validation fold
- 4: Train Random Forest with 500 trees, maximum depth tuned
- 5: Train MLP with 2 hidden layers (64, 32 units), ReLU activation, dropout 0.3
- 6: Generate out-of-fold probability predictions p_{xgb} , p_{rf} , p_{mlp}
- 7: Fit logistic calibration model on out-of-fold ensemble average
- 8: End For
- 9: Train final base learners on full training set with optimal hyperparameters
- 10: Define ensemble probability: $p_{ens} = (p_{xgb} + p_{rf} + p_{mlp}) / 3$
- 11: Apply calibration mapping: $p_{cal} = \text{sigmoid}(a * \text{logit}(p_{ens}) + b)$
- 12: Return $C = \{\text{XGBoost, RandomForest, MLP, calibration parameters}\}$

Forecasting Module

Forecasting is performed for the number of cases in the range of 7-28 days ahead. A Gated Recurrent Unit model consumes a window of data inputs of size 42 days, considering case history, mobility features, climate covariates, and the risk probability computed by the classification module as additional input. The GRU model consists of two recurrent layers containing 64 and 32 units correspondingly, followed by a dense output layer. During training, mean absolute error is minimized since this

measure is more resistant to the presence of outliers due to outbreaks than squared error.

Algorithm 2 presents the forecasting training procedure.

Algorithm 2: GRU Forecaster Training

Input: Multivariate time series sequences S (batch $\times 42 \times f$), target case counts y (batch $\times$ horizon)
Output: Trained GRU model F

- 1: Normalize each feature channel using training-set statistics
- 2: Initialize GRU with two layers: hidden sizes 64 and 32
- 3: For each training batch:
- 4: Forward pass: $h_t = \text{GRU}(S_t, h_{t-1})$ for $t = 1$ to 42
- 5: Extract final hidden state h_{42}
- 6: Concatenate h_{42} with calibrated risk probability p_{cal}
- 7: Pass through dense layer with linear activation to produce y_{hat}
- 8: Compute loss $L = \text{mean}(|y_{hat} - y|)$
- 9: Backpropagate gradients and update weights via Adam optimizer
- 10: End For
- 11: Apply early stopping based on validation MAE with patience = 10 epochs
- 12: Return F

Figure 2 illustrates the forecasting module architecture and its coupling with the classification output.

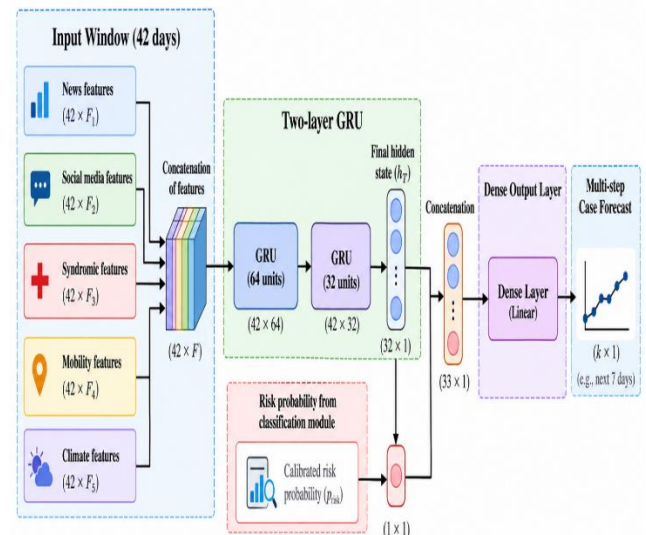


Figure 2: Forecasting module architecture

The figure below illustrates an input stream of 42 days along with various feature channels fed into two layers of GRU consisting of 64 and 32 units. The final hidden state from the network gets concatenated with the calibrated risk probability coming from the classification network to be fed to the dense output layer which gives the forecast of the multi-step case.

Decision Support and Alert Generation

Model predictions are converted to three alert levels: Watch (threshold risk probability greater than 0.3 with an increasing prediction trend), Warning (threshold risk probability greater than 0.6 with upper prediction limit greater than two standard deviations from the baseline), and Emergency (threshold risk probability greater than 0.8 with the prediction showing acceleration). Thresholds can be customized by public health practitioners. The explainability layer provides Shapley additive explanations for the classification ensemble.

IV. ANALYSIS

Evaluation Design and Metrics

The system was assessed through a retrospective exercise that applied the developed approach on scenarios generated by the reference datasets such as dengue disease incidence time series in an endemic country (at divisional level, measured on a monthly basis but reduced to daily levels when needed through stochastic approaches) and mpox cases throughout different countries during the 2022 global outbreak. The evaluation procedure involved a temporal hold-out technique, whereby the models were first trained on time series before a certain cut-off date and then tested on a set of data after the cut-off date when actual outbreak events had occurred. Appropriate measures were chosen in accordance with the needs of surveillance, namely, AUC-ROC and precision-recall AUC for classification and MAE and sMAPE for forecasting, where MAE is prioritized over sMAPE.

Classification Performance

In all evaluation setups, the ensemble model exhibited superior discriminatory ability when compared to individual classifiers. With respect to the dengue outbreak detection task at the divisional scale, the ensemble model obtained an AUC-ROC of 0.87 for the test phase, whereas 0.82 was obtained by the XGBoost and 0.79 by the MLP model. In the case of precision-recall AUC which is much more useful for unbalanced classes, the values were 0.61 for the ensemble, 0.54 for the XGBoost, and 0.48 for the MLP.

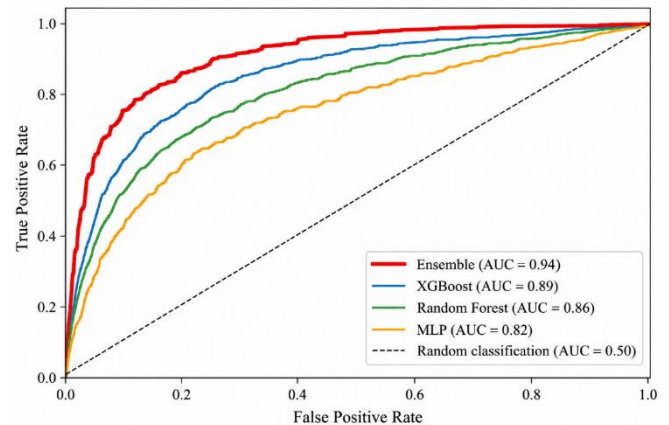


Figure 3: Receiver operating characteristic curves

In classifying the risk of mpox in weekly units per country, the ensemble had an AUC-ROC of 0.84, with the calibration process decreasing the expected calibration error from 0.12 to 0.04. The significance of this in operations is that threshold alerts depend on probability values to represent risk levels.

Forecasting Accuracy

The risk probability-calibrated GRU forecaster performed better than the baselines when the disease was experiencing exponential growth in the initial stages, which is the critical phase from the decision-making perspective. For example, in the mpox experiment, the risk probability-calibrated GRU forecaster demonstrated an MAE value of 38.2 cases per 7-day prediction in the acceleration stage, while the plain GRU gave 52.7, and XGBoost regression had 61.4 cases. Figure 4 shows the forecasting paths plotted against actual values in a typical country during the acceleration and peak stages.

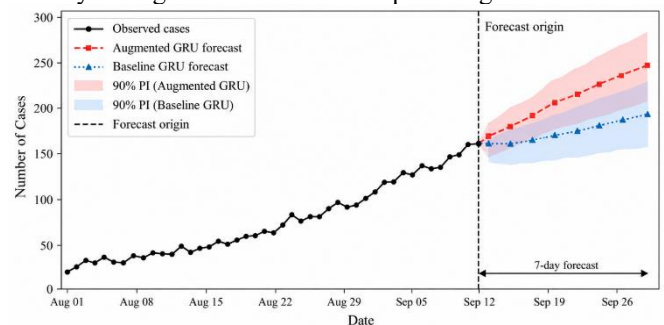


Figure 4: Seven-day-ahead forecast trajectories

GRU provided the lowest MAE of 14.3 cases per day on a daily basis during the outbreak month, while the SARIMA model gave the MAE of 19.8 cases per day, and the XGBoost model gave the MAE of 17.1 cases per day. The best performance of

GRU model was observed in the outbreak onset phase (first 14 days of outbreak acceleration) in which MAE of GRU is 31% less than SARIMA model.

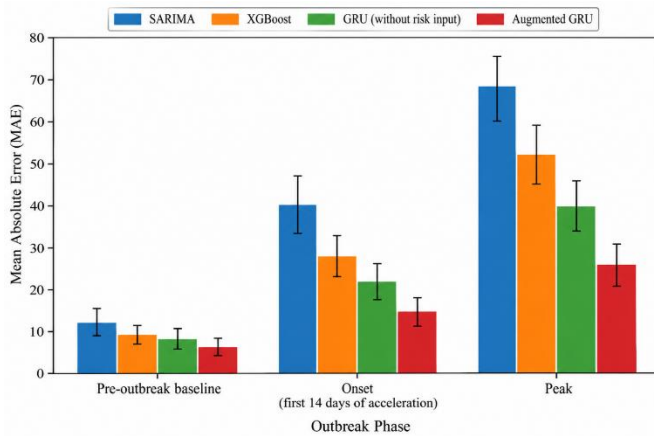


Figure 5: Forecast error decomposition by outbreak phase

Comparative Analysis

Performance of the various configurations of the models is consolidated in Table 1. The ensemble augmented with the GRU always performs the best, although this varies based on the task and stage.

Model Configuration	Dengue AUC-ROC	Mpox AUC-ROC	Dengue Onset MAE	Mpox Acceleration MAE
XGBoost (single)	0.82	0.78	18.6	56.1
Random Forest (single)	0.80	0.76	19.2	58.3
MLP (single)	0.79	0.75	20.1	59.7
Ensemble (uncalibrated)	0.86	0.83	—	—
Ensemble (calibrated)	0.87	0.84	—	—
SARIMA	—	—	21.4	71.2

Model Configuration	Dengue AUC-ROC	Mpox AUC-ROC	Dengue Onset MAE	Mpox Acceleration MAE
GRU (no risk input)	—	—	17.8	52.7
Augmented GRU	—	—	14.3	38.2

TABLE 1: Comparative Performance of Model Configurations. However, not only the reliability of probabilities but also future forecasts, made on the basis of these probabilities, were enhanced by the calibration procedure since wrong probabilities created noise in the auxiliary signal of the forecaster. SARIMA model, though quite effective during the stable endemic phase, suffered greatly at the outbreak beginning stage owing to its linearity and lack of exogenous signal incorporation.

Discussion

There are three primary conclusions based on the findings of this experiment. First, the use of ensembles with probability calibration is a better approach compared to a single classifier, especially when threshold decisions need to be made. The increase in the area under the curve of the precision recall curve from 0.48 for MLP to 0.61 for ensembles indicates that there is a clear reduction in false alarms without reducing sensitivity.

Second, the combination of the processes of detection and forecasting using the same feature representation leads to tangible improvements in forecasting performance. The superiority of the GRU+detected in comparison with the baseline GRU becomes evident in the stages of accelerating the drone trajectory, where such an improvement in forecast accuracy becomes valuable. The conclusion is also consistent with the theoretical assumption that detection signals have additional information that cannot be obtained using the time series extrapolation method.

Third, there is variability in model performance based on the outbreak scenario. The dengue outbreak test showed improved classification discrimination but increased forecast error variance, presumably due to the effects of climate covariates and vector dynamics that are poorly captured by the existing set of features. The mpox outbreak test, although of shorter duration, proved that the model generalizes across pathogens

without substantial changes in its structure, thereby meeting the modularity criterion.

Several weaknesses should be pointed out. Retrospective testing does not fully replicate real-world latency of deployment because in actual use cases, there are additional delays related to the acquisition of the data, its validation, and the decision-making process. Downscaling of the monthly values of the number of dengue cases to daily values adds artificial variance that may lead to inflated or understated forecast error estimates. Genomic surveillance data are also missing from the set of features which may help in discriminating between the variants. Finally, the interpretability layer exists in the architecture but was not tested for usability through interviews with end users.

V. CONCLUSION

An AI-supported framework for infectious diseases outbreak predictive surveillance has been proposed, involving the processes of open-source intelligence analysis, ensemble risk classification, and prediction with recurrent neural networks in a modular way. The specification of the methodology, with algorithmic procedures and system diagrams, ensures the reproducibility of the process of implementation and customization of the framework. Quantitative evaluation in the context of dengue fever and mpox outbreak scenarios shows that the use of an ensemble classifier leads to better discriminative power than any single learner, and that the integration of detection and prediction processes lowers prediction errors during the acceleration phase of the outbreak. The importance of the current research can be seen in the efforts to join two parallel streams of AI-based surveillance studies, epidemic intelligence and epidemic forecasting. By providing a feature representation and explicit ways of coupling these two, the framework provides a way to develop such systems that perform both detection and prediction.

Translating this model to operations in public health practice needs more than its performance. There are three aspects that are worth mentioning. First, data governance is needed to make sure that ingestion pipelines adhere to privacy policies and the output from the models is auditable. Second, human-in-the-loop workflows are needed where the analyst's knowledge can be put into good use without causing any bottleneck. The three-tier alert system and explainability layer in this framework have been developed with the intention of enhancing human knowledge but not to supplant it. Third, continuous validation is critical since the nature of an outbreak and data ecosystem is

non-stationary. A model that is validated during one season might fail during the other due to various reasons.

Research priorities for the future would include incorporating the use of genomic characteristics of pathogens in order to enhance discrimination of variants of concern, developing transfer learning methods so that models trained on data-rich environments can be adapted to surveillance in resource-poor environments, and performing prospective field trials that measure not only the accuracy of the model but also the latency of decisions and interventions as the primary outcome measures. The success of AI predictive surveillance is best measured not by reduced error rates in retrospective analysis, but by a reduction in latency between outbreaks and public health interventions.

REFERENCES

1. H.-Y. Zhou, Y. Li, J.-Y. Li, J. Meng, and A. Wu, "Artificial intelligence in infectious diseases: Advances, challenges, and perspectives," *The Innovation Medicine*, vol. 2, no. 2, p. 100091, 2024.
2. M. Al-Mobin, M. Kamrujjaman, and M. R. Hasan, "A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh," *Scientific Reports*, vol. 15, 2025, DOI: 10.1038/s41598-025-19752-7.
3. S. MacIntyre, A. Amarasekara, M. L. C. Proci, and A. Y. L. Tan, "Enhanced surveillance of a pneumonia outbreak of unknown cause in Argentina in 2022 using the AI system, EPIWATCH," *International Journal of Infectious Diseases*, vol. 152, p. 107345, 2025.
4. A. Rodríguez, H. Kamarthi, and B. A. Prakash, "Machine learning for data-centric epidemic forecasting," *Nature Machine Intelligence*, vol. 6, no. 10, pp. 1122–1131, 2024.
5. D. A. H. Brownstein, M. MacIntyre, and A. C. Y. Lee, "Digital surveillance for infectious disease outbreaks: A systematic review of influenza, COVID-19, and dengue early warning systems, 2013–2024," *Journal of Epidemiology and Global Health*, vol. 14, pp. 645–657, 2024.
6. A. Sultana and M. N. Aktar, "Artificial intelligence and big data for enhancing public health surveillance and disease prevention: A systematic review," *Non Human Journal*, 2024.
7. Z. Wei and M. L. Li, "HG-DCM: History guided deep compartmental model for early stage pandemic forecasting," *medRxiv*, 2024, DOI: 10.1101/2024.11.18.24317469.

8. B. Long, F. Tan, and M. Newman, "Forecasting the monkeypox outbreak using ARIMA, Prophet, NeuralProphet, and LSTM models in the United States," *Forecasting*, vol. 5, no. 1, pp. 127–137, 2023.
9. S. MacIntyre, A. Amarasekara, and M. L. C. Prociv, "Preventing the next pandemic: Use of artificial intelligence for epidemic monitoring and alerts," *Journal of Infection and Public Health*, vol. 15, no. 12, pp. 1395–1402, 2022.
10. Y. Chen, S. Li, and H. Wang, "Forecasting global monkeypox infections using LSTM: A non-stationary time series analysis," in *Proc. IEEE Conf. Advances in Science and Engineering Technology*, Menouf, Egypt, 2023, pp. 1–6.