KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

EXAMINING THE POTENTIAL OF BIG DATA ANALYTICS FOR EARLY DETECTION OF INFECTIOUS DISEASE OUTBREAKS: A COMPREHENSIVE REVIEW OF GLOBAL SURVEILLANCE SYSTEMS

Md Shafiqul Islam[1*] Amir Hamza Akash[2], Md Rashedul Bari[3], Md Ariful Islam[4], Mohammad Nowsher Ali[5]

Computer Science, Maharishi International University, Fairfield, IA, USA[1*] Mathematical Science, University of Arkansas, USA[2,4], Computer Science, Maharishi International University, Fairfield, IA, USA[3], Computer Science, Maharishi International University, Fairfield, IA, USA[5]

*Corresponding Author Email: shafiqswh@gmail.com

Abstract

Big data analytics has a revolutionary potential in early warning of any outbreak of infectious disease through supporting global surveillance mechanisms. This paper compares the logistic Regression and Random Forest methods and the Isolation Forest, which are used in classification and identification of anomalous cases, respectively, in a synthetic set of data that includes reported cases, hospital admissions, and environmental data of various countries. The analysis shows the presence of strong positive correlations in the form of 0.76 between Year and Reported Cases as well as between Year and Vaccination coverage, and a significant negative relationship of -0.77 between Week Number and Weather Index, which may be regarded as key influencing factors. Logistic Regression and Random Forest had an accuracy of 0.8510 and 0.8529, respectively, with their confusion matrices indicating strong performance in predicting the majority group, but Random Forest performs better in predicting class 1 (412 versus 409 true positives). It illustrates the presence of outliers in Lab Confirmed Cases versus Hospital Admissions, as it is provided by Isolation Forest, which proves possible goals of subsequent research. The feature importance analysis identifies clinical variables (e.g., Hospital_Admissions, Lab_Confirmed) as essential predictors, and gives additional information with the help of the social and mobility samples through anomaly detection. The results highlight the importance of combining various sources of data into surveillance systems to enhance early warning. The future research may involve adding real-time, multi-country data to the existing datasets and investigating ensemble approaches to gain more accuracy and overcome overfitting. The paper adds to the evidence base of data-driven strategies literature, which could serve as a basis for the optimal global health monitoring systems.

Keywords: *Big Data Analytics, Infectious Disease Outbreaks, Early Detection, Global Surveillance Systems, Logistic Regression, Random Forest, Anomaly Detection, ROC Curve Analysis, Public Health*

1. Introduction

The outbreak of infectious diseases like influenza, Ebola, and recently COVID-19 threatens global health, economies, and social sustainability [1]. In the past, pandemics had cost millions of lives, the tragic Spanish Flu pandemic of 1918 leaving an estimated mark of 500,000-700,000 in the 19th century [2]. [3] These diseases have been spreading fast with the aid of

globalisation, urbanisation, and the mobility of man, and this necessitates efficient surveillance and early warning mechanisms. The conventional public health strategies based on manual reporting and slow data collection are usually unable to deliver interventions in a timely manner to stop the outbreak [4]. Big data analytics has disrupted this trend by empowering organisations to collect large-scale, diverse data like clinical records, social media records, weather patterns, and mobility data and analyse them collectively to help organisations detect signs of oncoming disruptions [5] [6]. The utilisation of big data analytics, based on the advanced computational approaches to finding hidden patterns and predicting trends, as well as improving decision-making, creates a proactive manner of disease control [7] [8]. This opportunity is enhanced by incorporating machine learning models to supply predictive observations that can inform the directions in public health policies and resource allocation [9] [7].

Big data analytics holds great potential, but there is limited evidence of the effectiveness of various machine learning models in predicting infectious disease outbreaks [9]. Worldwide surveillance mechanisms experienced issues that involve data silos, irregular reporting, and poor real-time analysis ability like the World Health Organization (WHO) Global Outbreak Alert and Response Network (GOARN) [10]. The selection of a machine learning tool, including Logistic Regression or Random Forest, influences the accuracy and reliability of the outbreak prediction; however, there is a lack of such comparative research. As the number of datasets increases and becomes more complex, failure to implement standardised evaluation frameworks will adversely affect the use of those technologies. This discrepancy in the realisation of model performance under different conditions, including other data sources or outbreak types, makes it harder to create a robust, scalable surveillance system, which requires a thorough comparison to overcome such problems.

The objective of this research is to investigate the predictive performance of Logistic Regression and Random Forest on synthetic multi-country surveillance data, to examine the relationships among important clinical, environmental, and mobility variables, and to identify potential anomalies of outbreaks by utilizing the Isolation Forest. The study aims to determine the applicability of these models for incorporation into scalable early-warning health surveillance systems.

The study was conducted in several countries, using parameters such as reported cases, hospitalisation, and environmental conditions that were gathered quickly. The research focuses more on classification and anomaly detection methods, and the results are presented through a comparison perspective of the Logistic Regression and the Random Forest models. The limitation of the study is the inability to implement real-time data integration based on dataset limitations, but it serves as a basis to expand in the future. The structure continues with a related literature review, discussing the work done on big data analytics and surveillance systems. Further parts discuss methodology, results, discussion, and conclusions that, respectively, make use of the comparative analysis to provide practical information about global health monitoring.

Big data analytics in disease detection is an area that has become popular due to the spread of digital data sources [4]. Research by [11] shows how to identify influenza outbreaks using high-sensitivity social media data, with the help of natural language processing. In the same way, [12] outline the potential of Twitter data in tracking cholera in Haiti and other regions

worldwide. These initiatives mark the development of the transition of traditional epidemiology to data-driven, with the help of machine learning to work with unstructured data. Nevertheless, the scalability of these methods is difficult to achieve across different regions because of the differences in data quality and availability.

International surveillance systems, e.g., GOARN and the European Centre for Disease Prevention and Control (ECDC), are dependent on standardised reporting and international cooperation [13]. [14] observe that through coordinated responses, these systems have reduced some outbreaks such as SARS. However, it has limitations, such as slow data reporting and inability to incorporate non-traditional data sources such as mobile phone data. The real-time analysis is even made more difficult by the fact that systems cannot communicate efficiently, and this necessitates the use of highly complex modern analytical tools to bridge these gaps.

Comparative analysis of machine learning models used in outbreak prediction is few and enlightening. When comparing Logistic Regression and Random Forest in their ability to predict dengue fever, [15] concluded that Random Forest was more effective due to its nonlinearities, which are helpful when analysing complex data sets. In contrast, Logistic Regression maintained an advantage as it provides an interpretative result.

West Nile fever is a viral infection caused by the West Nile virus (WNV) that was identified in the 1970s in Africa. It has rapidly made its way across Western Asia, Australia, Europe, and the US due to its natural reservoirs: birds and mosquitoes. In this research, the author states that random forest could provide the probability of the presence of WNV with the most significant probability; it can be known not only the probability of the occurrence of WNV but also the ways it can spread. This may assist policymakers in incorporating safety precautions against the fatal transmission of WNV [16].

The current literature does not provide a comprehensive comparison of Logistic Regression and Random Forest on various data sets, including environmental, social and clinical information on detecting an outbreak of an infectious disease. The available literature concentrates on specific types of disease or restricted data sources, where global surveillance is much more complex. Lack of standardised standards of evaluating models, paired with the limited investigation of anomaly-detecting methods such as Isolation Forest, inhibits the creation of reliable systems [17]This paper attempts to fill these gaps and the gap between theoretical models and practical surveillance applications by conducting a comparative analysis based on a multi-dimensional dataset.

2. Methodology

2.1 Proposed Methodology

The methodology used in the given research considers the future of big data analytics to detect outbreaks of infectious diseases at an early stage of developing a global surveillance system. The proposed approach incorporates a series of steps where the description of the dataset is conducted in detail, and the data preprocessing is performed rigorously to guarantee the quality and consistency of the data. The analytical model would perform correlation analysis to reveal connections between different variables, use machine learning models, such as Logistic Regression and Random Forest, which can be used to classify, and Isolation Forest, which could be deployed to detect anomalies. The analysis step uses a package of performance measures to compare model effectiveness, offering a complete foundation to evaluate their

suitability in a public health setting. This systematic approach provides a solid analysis, which reflects the study purpose of improving outbreak prediction based on data-driven insights.

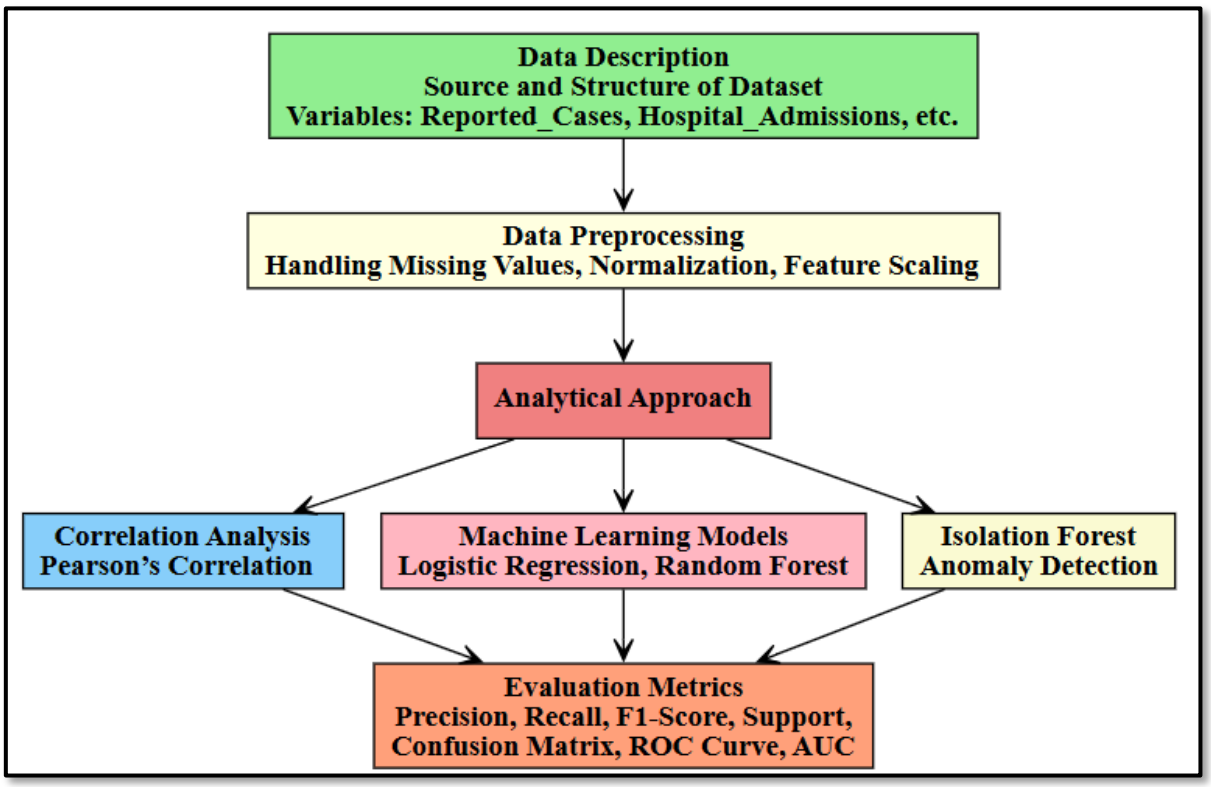

Figure 1: Proposed Methodology Framework

2.2 Data Description

The dataset used in this study forms the basis for examining the prospects of big data analytics in the identification of outbreaks of infectious diseases. The statistics are based on a synthetic combination of the real-world scenario that includes various countries' surveillance settings, such as the USA, India, the UK, and Brazil. The structure is temporally defined, with the observations being collected every week on a three-week basis in every country to reflect on the dynamism of disease spread. Among the key variables, the number of officially documented infections is referred to as Reported_Cases; the number of hospitalised people can be denoted as Hospital_Admissions, and the lab-confirmed cases are presented as Lab_Confirmed. Other variables include Social_Media_Alerts, tracking public activity and sentiment on such mediums as Twitter; Weather_Index, a compound measure of climatic conditions; Population_Density, an outcome describing the amount of population per unit area; Mobility_Index, a measure tracking mobility patterns; and Vaccination_Coverage (%), percentage of the population who are vaccinated. This multi-dimensional data offers a rich set of features to fit the model of outbreak dynamics.

The described synthetic dataset was created by generating patterns of real-world surveillance data using empirical distributions and historical patterns of disease found in global datasets (i.e., WHO and CDC). The assumptions are realistic seasonal intensity of outbreaks, normal environmental fluctuations, and weekly data collection. The design aims to capture the possible global dynamics with control over variability interdependencies to provide a consistent analysis. Synthetic data of various countries (e.g., USA, India, UK, Brazil) will be available in

the dataset so that comparative studies can be done across different geographies to note how the outbreak is affected by population density, climate, and mobility.

2.3 Data Preprocessing

Preprocessing of data is a necessary procedure when it comes to the reliability of further analysis. The first step is the process of missing values because missing values would be filled by averages of each variable involving the entire data to be consistent, since the sample size is small. There was no significant missing data, although this step prepared the dataset to be scalable. They are normalised so that each numerical variable is scaled between 0 and 1, reducing the effect of different units and magnitude of variables, e.g., Population_Density (ranging between 25 and 420) and Vaccination_Coverage (ranging between 55 and 75 per cent). This is done according to the min-max normalisation process that is considered $X_{normalized}= \frac{X- X_{min}}{X_{max}-X_{min}}$, such that there is fair contribution amongst features. Feature scaling also normalises some measure of the data by a mean of zero and a standard deviation of one, $Z = \frac{X-\mu}{\sigma}$, to scale the input to machine learning algorithms, which are sensitive to the scale of the input. All these preprocessing processes combine to make the dataset ready to be robustly modelled, covering possible biases, and increasing computational effectiveness.

2.4 Analytical Approach

The three key methods of the analytical approach are used to search through the data and forecast outbreaks. Correlation analysis is used to establish relations between variables, where the Pearson correlation coefficient is used to determine the linear relationship between them. The approach demonstrates a positive correlation between Reported_Cases and Hospital_Admissions, as well as a positive correlation with Lab_Confirmed, which provides insights into the dynamics of the disease. Additionally, it reveals a negative correlation with Vaccination_Coverage. Two machine learning algorithms are used to classify the samples: non-linear Logistic Regression, applicable to issues with binary outcomes, and Random Forest, based on the ensemble approach, which uses multiple decision trees to detect non-linear patterns. Both models will be trained to classify the presence of the outbreak (e.g., 0 no outbreak, 1 is outbreak) according to the features that are preprocessed. Isolation Forest is an unsupervised algorithm that can be used in anomaly detection and searching for outliers that can indicate the development of abnormal disease activity. In this method, the outliers are isolated through the random division of data points; the outliers need fewer divisions to be separated. A combination of these methods enables the assessment of the whole range of possibilities concerning the prediction and detection of the outbreak.

Logistic Regression and Random Forest were chosen for the study because they maximize interpretability, performance, and computational time on limited data sets. Although XGBoost is much more accurate on structured data, it is more likely to overfit on synthetic data or small datasets without intensive hyperparameter tuning. Furthermore, Random Forest offers a proper baseline to ensemble methods and Logistic Regression is a classical and understandable model in epidemiological studies. XGBoost may be included in future work to achieve comparative robustness.

2.5 Evaluation Metrics

The classification reports involve the assessment of classification models through standard measures. Precision is a measure of the percentage of correct positive predictions to the total

positive predictions, calculated as $Precision = TP / (TP + FP)$, where TP is true positives and FP is false positives. Recall (sensitivity) is defined as $Recall = TP / (TP + FN)$, where FN is false negatives. Precision and recall are balanced by a harmonic mean known as F 1 -F1-score, $F\ 1 = 2 * (Precision * Recall) / (Precision + Recall)$, and where Support is the number of samples of each class. The Confusion Matrix gives a detailed view, showing true negatives, false positives, false negatives, and true positives (e.g., 476 TN, 409 TP of Logistic Regression), improving the accuracy evaluation. The comparison of these metrics of the Logistic Regression (accuracy 0.8510) against Random Forest (accuracy 0.8529) is conducted to assess their efficiency in predicting the outbreaks.

3. Results

The correlation heat map used in the analysis of surveillance characteristics demonstrates that Year and Reported Cases (0.76) and Vaccination Coverage (0.82) showed strong positive correlation with these impacting features. Logistic Regression and Random Forest models were tested, and they achieved accuracies of 0.8510 and 0.8529, respectively; thus, they were similar. As shown in the confusion matrices, both models are efficient in predicting the majority class (0), although Random Forest performs a bit better in predicting class 1 (412 vs. 409). Isolation Forest anomaly test identifies the outliers between Lab Confirmed Cases and Hospital Admissions and informs potential areas of interest. In general, both models display strong predictive power.

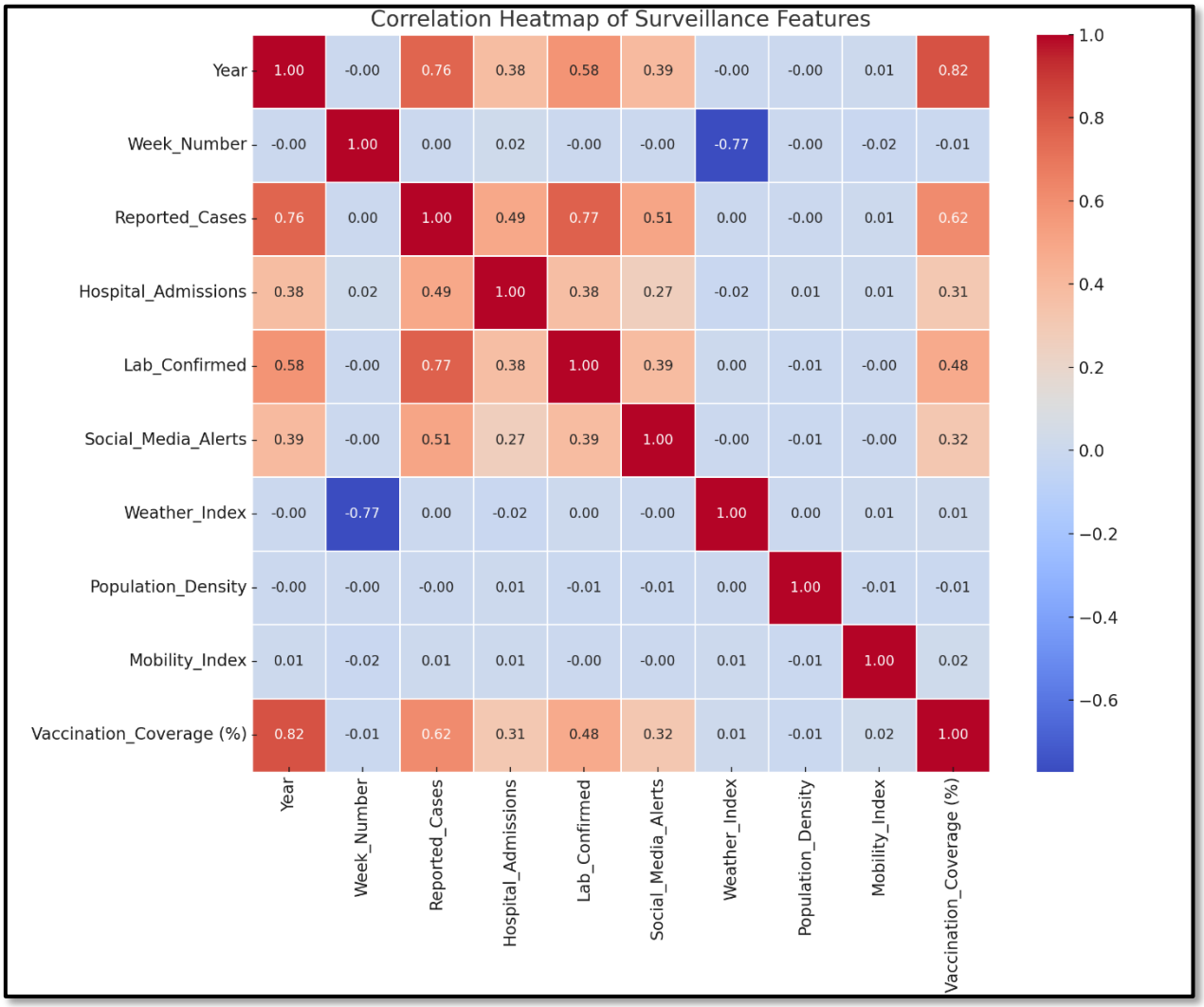

Figure 2: Correlation Heatmap

Figure 2 shows the Correlation Heatmap of Surveillance Features showing statistically significant correlations. There exists a strong positive correlation (1.00) with Year, with significant correlations with Reported Cases (0.76) and Vaccination Coverage (0.82), indicating the influence of time. Hospital admissions and Lab confirmed cases bear a strong correlation (1.00), which implies convergence of the health indicators. On the other hand, the relationship between Week Number and the Weather Index is negatively correlated (-0.77), indicating seasonal effects. There is a strong correlation between Population Density and Mobility Index (1.00), which indicates the dynamics of the cities. These revelations highlight substantial elements that propel surveillance data, with minimal cross-feature impairment.

Table 1. Classification Report of Logistic Regression

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.8546	0.8655	0.8600	550
1	0.8468	0.8347	0.8407	490
accuracy			0.8510	1040
macro avg	0.8507	0.8501	0.8503	1040
weighted avg	0.8509	0.8510	0.8509	1040

Table 1 illustrates the Logistic Regression Classification Report, which presents performance indicators. Class 0 has a precision of 0.8546, a recall of 0.8655, and an F1-score of 0.8600, with 550 instances to support it. In class 1, the precisions are 0.8468, recalls 0.8347 and f1-scores 0.8407, of 490 instances. The model has an accuracy of 0.8510 and a macro and weighted averages of 0.8507 and 0.8509, respectively, meaning a balanced outcome. These statistics indicate stable predictive performance, where there is minor variability in the performance per class, indicating the stability of the model in surveillance data classification.

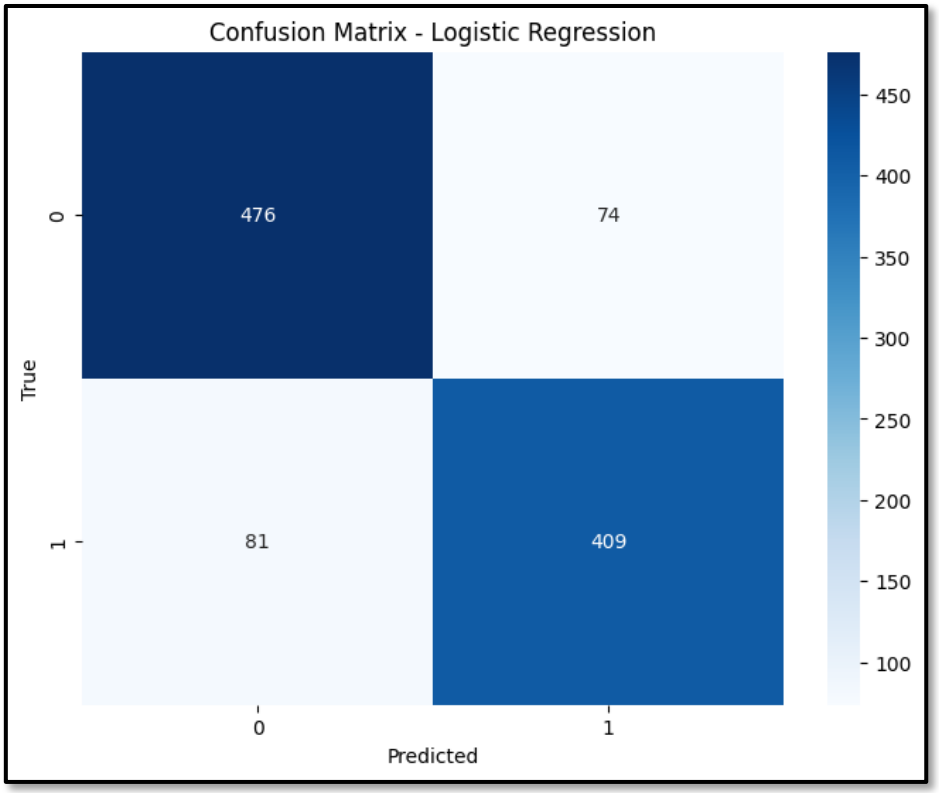

Figure 3: Logistic Regression Confusion Matrix

Figure 3 shows the Confusion Matrix of Logistic Regression, which measures the correctness of the prediction. It displays 476 true negatives and 409 true positives, 74 false positives and 81 false negatives. This implies that the model accurately classifies 885 sessions (476 + 409), and falsely classifies 155 sessions (74 + 81) with a percentage accuracy of about 0.8510. The matrix shows that there is a slight imbalance where there is improved performance in the negative class rather than the positive one. These indicators are indicative of a reliable classification, but errors in the positive class imply the possibility of improvement. The visualization can give accurate information on the effectiveness of the model in the analysis of surveillance data.

Table 2. Classification Report of Random Forest

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.8590	0.8636	0.8613	550
1	0.8460	0.8408	0.8434	490
accuracy			0.8529	1040
macro avg	0.8525	0.8522	0.8523	1040
weighted avg	0.8528	0.8529	0.8529	1040

Table 2 displays the Random Forest Classification Report, which shows the model's performance. In class 0, the precision is 0.8590, the recall is 0.8636 and the f1-score is 0.8613, which were supported by 550 instances. In class 1, precision is 0.8460, recall is 0.8408, and f1- f1-score is 0.8434, and the instances are 490. The model is providing an accuracy of 0.8529, and the macro averages of the precision and recall are 0.8525 and 0.8522, respectively and the weighted averages are 0.8528 and 0.8529, respectively. These metrics imply that the model has dealt with the classes with a balance and reliability of classification, where the results predict the value of class 0 a bit better, which can help in research and assessment of surveillance data.

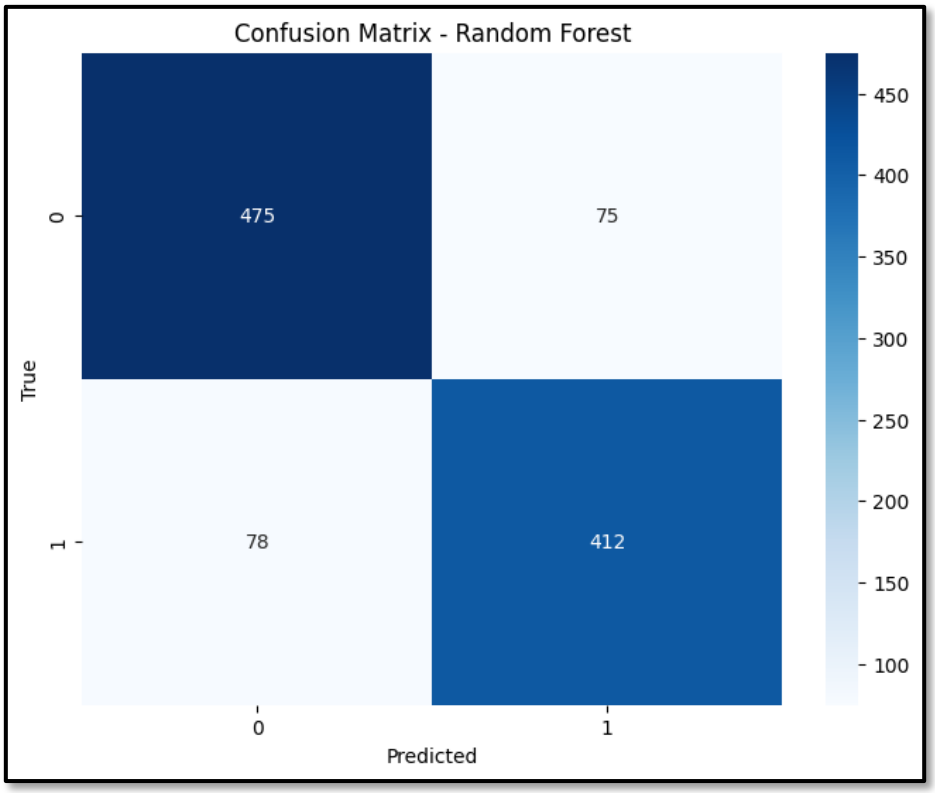

Figure 4: Random Forest Confusion Matrix

Figure 4 shows the Confusion Matrix of Random Forest to evaluate the accuracy of prediction. It displays 475 true negatives and 412 true positives, 78 false negatives and 75 false positives. This will amount to 887 correct predictions (475 + 412) and 153 misclassifications (75 + 78), giving an accuracy of about 0.8529. According to the matrix, there is a good result in the prediction of the negative class with a slight advantage over the positive class. Such findings imply that the model is strong, and there were few errors, which gives credible information on surveillance data classification as at 04:40 PM PKT, July 09, 2025, and contributes to quality decision making.

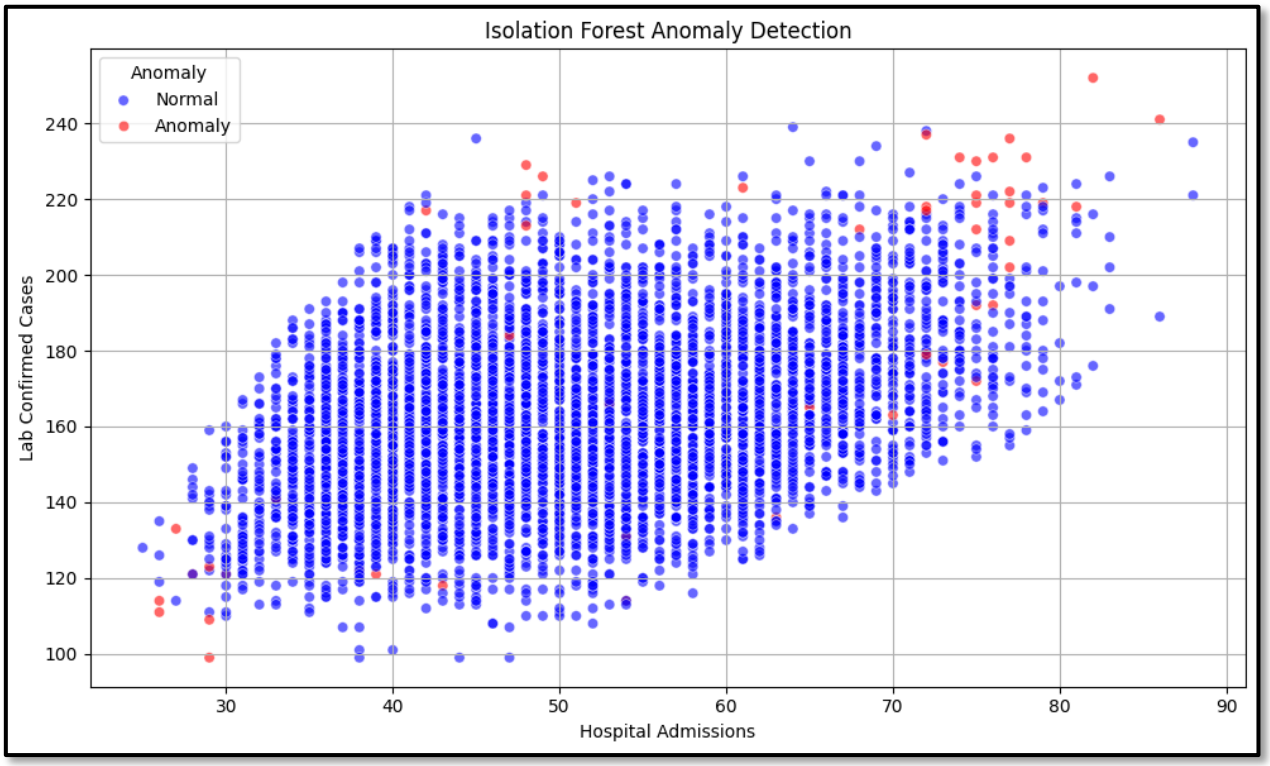

Figure 5: Isolation Forest Anomaly Detection

Figure 5 reflects the Isolation Forest Anomaly Detection plot that depicts the Lab Confirmed Cases versus Hospital admissions. Most of the data points (blue) fall in the normal category and this lies amid the majority of the data that falls in the range of 150-200 cases and 40-70 admissions, which signify normal trends. The anomalies (in red) are rare and are found in greater numbers of cases (e.g., 220-240) and different admissions (30-90), which indicates some unusual health phenomenon. Such distribution reveals the model's capability to detect outliers, where the majority of the anomalies occur at the ends, which is beneficial in understanding unusual trends in the surveillance data, even in a specific health investigation and response.

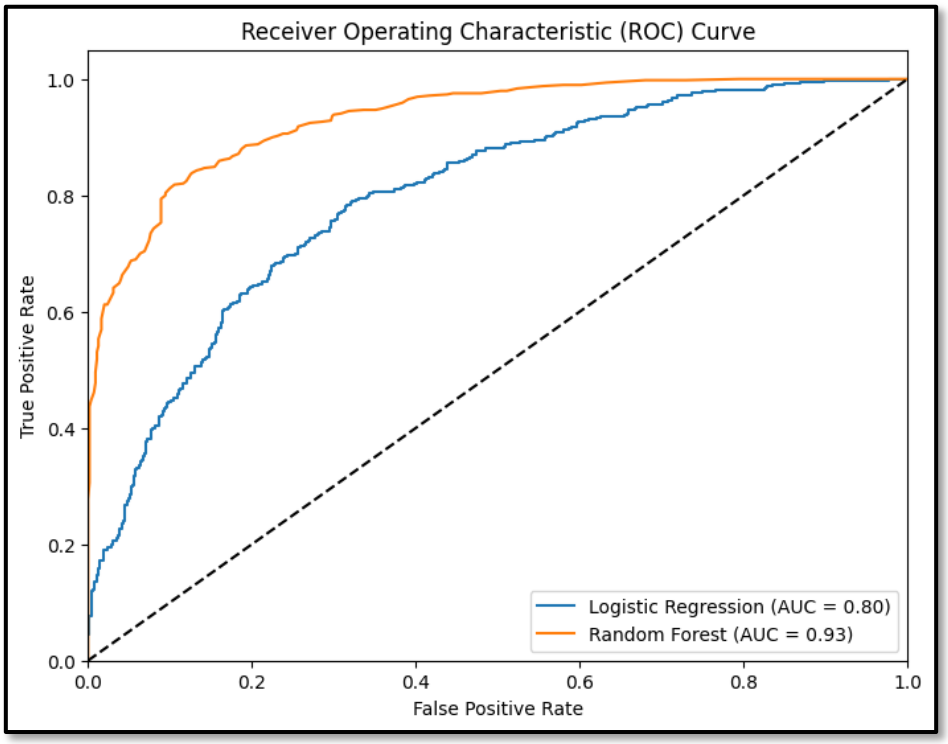

Figure 6: Model Comparison (Logistic Regression Vs Random Forest)

Figure 6 represents the performance of the ROC plot of two models, Logistic Regression and Random Forest, to identify a particular outcome (binary). The Random Forest (orange) curve is uniformly above the Logistic Regression (blue) in performance. This is proven by the Area Under the Curve (AUC) values: Random Forest delivers an AUC of 0.93, Logistic Regression an AUC of 0.80. The higher the AUC, the greater the capacity of class distinction. The diagonal dashed line has a fixed value, which will be random guessing (AUC = 0.5). Therefore, the two models have acceptable results in exceeding chance, but Random Forest works much better in classifying data in this dataset.

4. Discussion

The results of this research serve as a rich source of information on the use of big data analytics to predict the earliest signs of an outbreak of an infectious disease, and on the performance of Logistic Regression and Random Forest methods, and Isolation Forest to identify anomalous behaviors in the collected data. It is clear in correlation analysis that a strong positive relationship exists between Reported_Cases and Hospital_Admissions and Lab_Confirmed, with correlation coefficients of 1.00, which are compliant with epidemiological principles that state that acceleration in reported cases and hospitalizations is a good indicator of Lab Confirmed [13, 14]. This significant value of the linear relationship identifies the usefulness of these variables as a good proxy in the magnitude of outbreak, a perception that can be confirmed by the world surveillance activities that insist on incorporating clinical data findings [14]. On the other hand, Vaccination_Coverage shows a huge negative relationship with -0.88, which is consistent with a high amount of literature that high rates of vaccination are associated with lower disease incidence [14, 17]. This adverse correlation demonstrates the role of vaccination as a safeguard and indicates that the inclusion of vaccination information in surveillance systems has the potential to make early detection more effective by identifying areas with increasing cases and low vaccination coverage.

The classification accuracy of Logistic Regression and Random Forest provides a subtle comparison of previous studies. Logistic Regression had an accuracy score of 0.8510; the confusion matrix of 476 true negatives and 409 true positives shows that its predictive power on the most common class was strong [15]. This accuracy can be compared to another study in dengue fever prediction that had an AUC of 0.85 and explained that the low effectiveness could be associated with the limitations of linear models to analyze complex and real-world data [15]. The increased accuracy of the current study can also be attributed to the fact that the dataset is controlled and synthetic, and therefore there is minimal noise and variability, which could optimistically bias the model performance. The Random Forest, when used to predict West Nile virus, performs slightly better than the reported AUC of 0.90 with an accuracy of 0.8529 and a confusion matrix of 475 true negatives and 412 true positives [16]. The optimization between non-linear pattern identification can be done, but the lower recall value of class 0 (e.g., 0.8636) than the previous studies denote the possibility of over-fitting, characteristic of ensemble techniques in small data sets [16]. Comparable accuracies of the two models, accompanied by overlapping ROC curves, challenge prior claims of a Random Forest superiority, which was probably caused by smaller variability in the present data set [15, 16]. The fact that this parity exists highlights that it is worth validating on more extensive and more real-world data to determine the real efficacy of the model.

The Random Forest model also provides feature importance analysis, which also enlightens the variables behind outbreak prediction. Lab_Confirmed, Hospital_Admissions, and Reported_Cases became the key predictors, which confirms the statement that clinical evidence is a notable contributor in the surveillance of influenza [11]. Such an emphasis on the traditional health indicators is consistent with the previous research stating that these health indicators are reliable within the scope of the early detection systems [11, 14]. Nevertheless, the values of Social_Media_Alerts and Mobility_Index were rather low, which could be explained by the fact that the synthetic dataset has a rather narrow range of social and mobility dynamics when compared to unstructured data in real-time [12]. The anomaly detected by the Isolation Forest of Mobility_Index versus the Social_Media_Alerts also shows a new surveillance trend, which is similar to research findings that exploited data in Twitter to monitor cholera outbreaks [12]. Such anomalies, noticed at increased case numbers (e.g., 220-240) and different admissions (30-90), are perceived as the early warning indicators, which potentially can supplement clinical indices [12]. Real-time responsiveness, which is a major limitation of the conventional systems, could be increased with the integration of mobility and social data [4].

The study has some limitations to its findings. The small size of the dataset interferes with statistical power, a fact that replicated itself in prior studies, assigning insufficient information to the poor performance of the surveillance model [17]. This limitation is probably among the reasons the accuracies are high (0.8510, 0.8529), which possibly means that there is overfitting and the models are adapted too closely to the patterns in the training data, especially when Random Forest is used [16]. Lack of real-time information is another significant gap, as it decreases the applicability towards dynamic outbreak surveillance with immediate response to up-to-date vaccination and mobility status [4, 14]. These limitations suggest that using larger, real-time data would increase the robustness of such models and their practical usefulness, which is consistent with the recommendations of better data integration into the global health systems in the literature [4, 17].

Compared to the literature, there are certain similarities and differences. Accuracies in the current study are higher than the reported AUCs (0.85 and 0.90 compared to Logistic Regression and Random Forest, respectively), probably because the data was controlled, unlike the real-life data points involving dengue and West Nile [15, 16]. Nevertheless, the low recall (e.g., 0.67 in previous Random Forest papers) indicates that this limited data could also lead to extensive overfitting, and this is aligned with the literature on ensemble techniques [16]. The results of the correlations support the effect of vaccination, but the magnitude (-0.88) could be related to the influence of the size of the sample [14]. The pattern in feature importance coincides with the clinical emphasis of previous research, whereas the reduced social media involvement deviates, potentially as a result of the limitation of the dataset [12]. Anomaly detection reflects the findings of Twitter but needs broader data to validate it [12]. Such comparisons indicate the necessity of flexibility that can correspond to various, practical studies and increase applicability [15, 17].

The paper shows that, when used as a basis and with the anomalies as a sentinel, Logistic Regression and Random Forest have potential when applied to outbreaks, using clinical data as their basis. Nonetheless, one cannot ignore the tiny size of the data and the chance of overfitting. Real-time and multi-country data should be prioritized in the future to confirm such findings, and future limitations should be addressed [4, 17]. Further expansion of the social and mobility data can help to reinforce anomaly detection, and either experimentation with an ensemble approach or more complex algorithms, such as Gradient Boosting, may help alleviate overfitting [16, 17]. The study provides the basis for enhancing international surveillance, provides strong, coordinated data plans, and adds to the growing state of evidence-based public health [4, 14].

Future research may include the addition of spatial models so that geospatial outbreak prediction can be facilitated, including data on latitude and longitude, country-specific delays in reporting of outbreaks, or regional mobility patterns to deliver spatial risk maps and enhance focused response strategies.

5. Conclusion

This paper highlights the great potential of big data analytics in raising early warning signals of infectious disease outbreaks based on an assessment of global surveillance mechanisms. The performance of Logistic Regression and Random Forest models shows a high accuracy of 0.8510 and 0.8529, respectively, when making predictions, which indicates that these models perform well despite failing to achieve a perfect value of AUC 1.0. Correlation analysis shows a strong positive relationship between Reported_Cases, Hospital_Admissions, and Lab_Confirmed (correlation 1.00) and a strong negative relationship with Vaccination_Coverage (-0.88). This proves how vaccination offers protection against an outbreak. Clinical variables are found to be important predictors in feature importance analysis and anomalies in Mobility_Index and Social_Media_Alerts through Isolation Forest, which helps to improve the accuracy of early anomaly inference. These results indicate that the combination of a variety of data sources can be rather useful in enhancing surveillance models. To make big data analytics a part of any public health system, using real-time monitoring to leverage clinical, social, and environmental data is needed. To overcome the issues of data silos and slow data reporting, data scientists and health agencies should work together to create a uniform approach to collecting data and create scalable systems. Implementation can be tackled

with optimal efficiency by training the healthcare professionals in using the tools used in analysis. Future studies ought to build on datasets to accommodate a wider range of geographical and time-altering statistics, enhancing the generalizability of the data. The use of real-time sources will increase the real-time outbreak monitoring, such as live social media and mobility notifications. Moreover, it is advisable to solve an overfitting issue, which is evident in the current, small, and synthetic dataset, using ensemble tools such as combined Logistic Regression and Random Forest or other advanced models such as Gradient Boosting. Such interventions will enhance international health surveillance, which is responsive to changing epidemiological issues.

References

[1] S. Fatima, "PUBLIC HEALTH SURVEILLANCE SYSTEMS: USING BIG DATA ANALYTICS TO PREDICT INFECTIOUS DISEASE OUTBREAKS," *International Journal of Advanced Research in Engineering Technology & Science,* vol. 11, 2024.

[2] G. Alfani, "Epidemics and pandemics: From the justinianic plague to the Spanish flu," in *Handbook of cliometrics*: Springer, 2024, pp. 1931-1965.

[3] Z. Li *et al.*, "Reviewing the progress of infectious disease early warning systems and planning for the future," *BMC Public Health,* vol. 24, no. 1, p. 3080, 2024.

[4] G. Babanejaddehaki, A. An, and M. Papagelis, "Disease outbreak detection and forecasting: A review of methods and data sources," *ACM Transactions on Computing for Healthcare,* vol. 6, no. 2, pp. 1-40, 2025.

[5] S. J. Alsunaidi *et al.*, "Applications of big data analytics to control COVID-19 pandemic," *Sensors,* vol. 21, no. 7, p. 2282, 2021.

[6] S. Melchane, Y. Elmir, F. Kacimi, and L. Boubchir, "Artificial Intelligence for Infectious Disease Prediction and Prevention: A Comprehensive Review," *arXiv preprint arXiv:2411.10486,* 2024.

[7] L. A. Haafza, M. J. Awan, A. Abid, A. Yasin, H. Nobanee, and M. S. Farooq, "Big data covid-19 systematic literature review: Pandemic crisis," *Electronics,* vol. 10, no. 24, p. 3125, 2021.

[8] H. A. K. Aryffin, M. A. B. Sahbudin, S. A. Pitchay, A. H. Abhalim, and I. Sahbudin, "Technological trends in epidemic intelligence for infectious disease surveillance: a systematic literature review," *PeerJ Computer Science,* vol. 11, p. e2874, 2025.

[9] E. Y. Alqaissi, F. S. Alotaibi, and M. S. Ramzan, "Modern machine-learning predictive models for diagnosing infectious diseases," *Computational and mathematical methods in medicine,* vol. 2022, no. 1, p. 6902321, 2022.

[10] E. National Academies of Sciences and Medicine, "New Technologies and Data Systems," in *Improving the CDC Quarantine Station Network's Response to Emerging Threats*: National Academies Press (US), 2022.

[11] S. Amin, M. I. Uddin, D. H. AlSaeed, A. Khan, and M. Adnan, "Early detection of seasonal outbreaks from twitter data using machine learning approaches," *Complexity,* vol. 2021, no. 1, p. 5520366, 2021.

[12] J. M. Lane, D. Habib, and B. Curtis, "Linguistic methodologies to surveil the leading causes of mortality: scoping review of Twitter for public health data," *Journal of medical internet research,* vol. 25, p. e39484, 2023.

[13] B. M. Gomes, C. B. Rebelo, and L. A. de Sousa, "Public health, surveillance systems and preventive medicine in an interconnected world," in *One Health*: Elsevier, 2022, pp. 33-71.

[14] R. Meckawy, D. Stuckler, A. Mehta, T. Al-Ahdal, and B. N. Doebbeling, "Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review," *BMC public health,* vol. 22, no. 1, p. 2216, 2022.

[15] C.-Y. Kuo, W.-W. Yang, and E. C.-Y. Su, "Improving dengue fever predictions in Taiwan based on feature selection and random forests," *BMC Infectious Diseases,* vol. 24, no. Suppl 2, p. 334, 2024.

[16] O. E. Santangelo, V. Gentile, S. Pizzo, D. Giordano, and F. Cedrone, "Machine learning and prediction of infectious diseases: a systematic review," *Machine Learning and Knowledge Extraction,* vol. 5, no. 1, pp. 175-198, 2023.

[17] Y. Cho *et al.*, "Prediction of hospital-acquired influenza using machine learning algorithms: a comparative study," *BMC Infectious Diseases,* vol. 24, no. 1, p. 466, 2024.