Influenza Analysis using Machine Learning: An Experimental Study

Ajay Kumar[1], Shashikant Gupta[2]
[1,2] Lincoln University College, Malaysia
Email ID: ajay.phdcse@gmail.com, pdf.ajaykumar@lincoln.edu.my

Abstract: Influenza remains a considerable danger to public health owing to its yearly outbreaks and capacity for pandemic development. This study investigates the utilization of machine learning (ML) models for forecasting influenza epidemics through the analysis of real-world epidemiological and climatic data. We evaluated multiple supervised learning algorithms—Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting—using historical influenza data sourced from the CDC and meteorological databases. Evaluation included performance criteria including accuracy, precision, recall, F1-score, and ROC-AUC. Our findings demonstrate that ensemble methods such as Random Forest and Gradient Boosting surpass alternative approaches, providing essential instruments for early diagnosis and public health strategy formulation.

Keywords: Machine learning; Influenza; Flu; Infectious disease.

Introduction

Influenza remains a considerable danger to public health owing to its yearly outbreaks and capacity for pandemic development. This study investigates the utilization of machine learning (ML) models for forecasting influenza epidemics through the analysis of real-world epidemiological and climatic data. We evaluated multiple supervised learning algorithms—Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting—using historical influenza data sourced from the CDC and meteorological databases. Evaluation included performance criteria including accuracy, precision, recall, F1-score, and ROC-AUC. Our findings demonstrate that ensemble methods such as Random Forest and Gradient Boosting surpass alternative approaches, providing essential instruments for early diagnosis and public health strategy formulation.

There are mainly 3 types of influenza: Type A, Type B, and Type C. The main explanation is discussed below in a tabular column shown in Figure 1 for positive parenting when flu is expected.

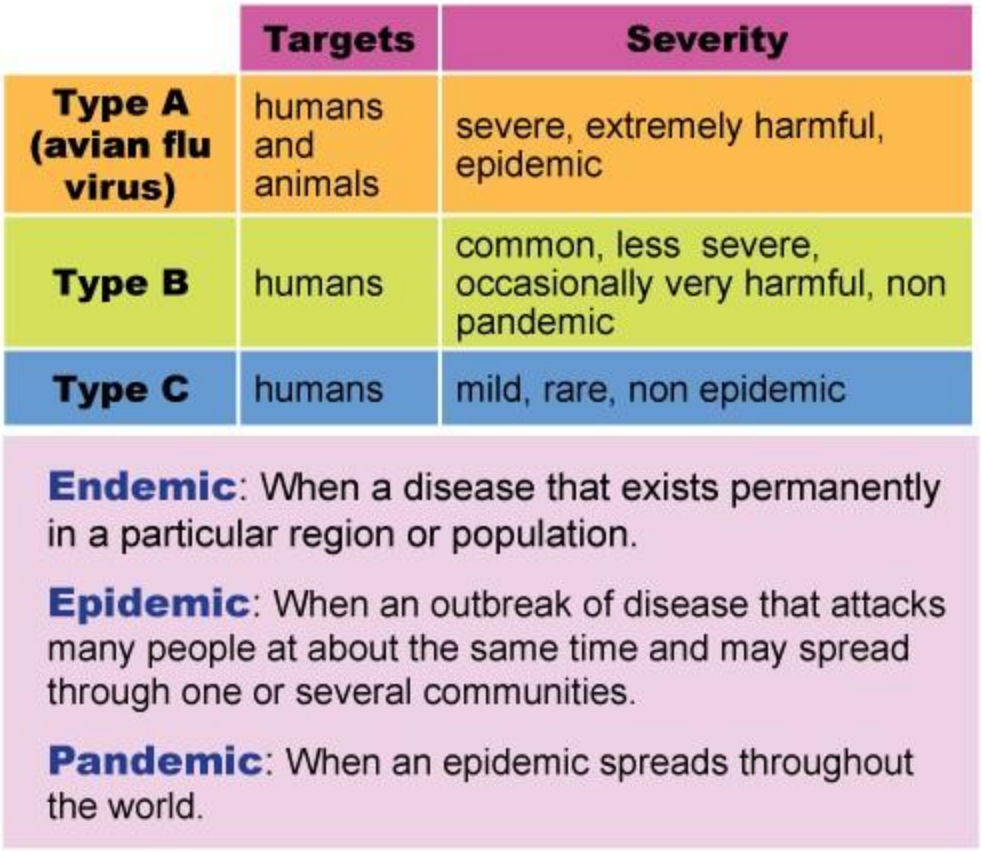

Figure 1: Severity level during flu attacks

Early influenza identification and prediction can manage outbreaks, reduce health risks, and enable timely medical intervention. Early detection helps deploy containment tactics like isolation and vaccination programs to stop influenza spread. Early diagnosis allows oseltamivir and zanamivir to be given in the important early window, lowering symptom severity and recovery duration. Early detection of influenza can avoid hospitalizations and deaths from pneumonia and respiratory failure, especially in susceptible populations. By examining historical trends and real-time data, machine learning and AI-driven models forecast outbreaks and optimize hospital and clinic resource allocation. Predicting influenza trends helps employers, schools, and institutions reduce economic disruptions and absenteeism. Early identification using advanced technology helps illness management, public health, and economic stability. Recent breakthroughs in machine learning (ML) promise epidemiological modeling and disease outbreak prediction tools. This work uses different ML methods to investigate influenza trends and test their efficacy.

Machine learning has transformed influenza analysis by enabling accurate forecasting and early response. ML can analyse clinical records, environmental factors, and social media trends to discover patterns and anticipate outbreaks, unlike traditional epidemiological models, which struggle with real-time adaptability [1]. SVM, Random Forest, and Neural Networks are commonly utilized to improve influenza surveillance. SVM classifies infected and uninfected people well, while Random Forest aggregates many decision trees to improve prediction accuracy [2]. Deep neural networks like Long Short-Term Memory (LSTM) evaluate

time-series data to accurately predict seasonal flu trends. AI-driven epidemic prediction, treatment plan optimization, and resource allocation make machine learning essential in current influenza study.
This project aims to improve influenza prediction using machine learning for timely interventions and illness management. The research selects high-quality information from the CDC and WHO to cover influenza trends and patterns. Predictive performance parameters like accuracy, precision, and recall are used to select machine learning models like SVM, Random Forest, and Neural Networks. Experimental outcomes involve data preprocessing to remove discrepancies, model training on historical influenza records, and validation on real-world outbreak data. This study uses AI-driven analytics to develop a strong influenza forecasting system to help healthcare practitioners and policymakers mitigate seasonal and pandemic influenza outbreaks.

This research work employs a systematic methodology, commencing with an Introduction that delineates the context, significance, and aims. The literature review identifies the research gap in the current system and delineates the techniques. The methodology outlines the dataset selection, preprocessing, and machine learning models employed for influenza prediction. The execution of the experimental methodology emphasizes the performance of the entire dataset samples and the outcomes of the machine learning models. The Results section provides model performance measures, whilst the Discussion interprets the findings and examines constraints. The Conclusion encapsulates essential observations and proposes avenues for future research. This logical progression guarantees clarity and accessibility for researchers and healthcare practitioners.

Related work

Numerous studies have utilized ML and statistical models for influenza prediction. Researchers have applied ARIMA, neural networks, and ensemble methods with varying success. However, few works offer a comparative evaluation across multiple models using consistent experimental setups and publicly available datasets.

The brief literature reviews are listed below in a tabular column in Table 1 to make more visible of findings from recent research papers:

Table 1. key findings in literature review

	Key Findings	Techniques used
Younghee Cho et.al. [1]	Compared ML models for predicting hospital-acquired influenza	Logistic Regression, Random Forest, XGB, ANN
Ranjan Kumar et.al. [2]	Proposed ensemble-based stacked algorithms for early influenza detection	Stacked ensemble models
Su Wei et.al. [3]	Used internet search data for forecasting influenza epidemics	LSTM neural networks

Ahmadi et.al. [4]	Zika outbreak analysis studies 54 papers	mHealth application enhancing Zika outbreaks

Method, Experiments and Results

Various publicly available datasets have been sourced so far. These primarily include data on Influenza-like illness (ILI) and historical weather records from the regions where outbreaks have occurred.

Dataset: This study incorporates several datasets from multiple sources to provide a thorough analysis of influenza trends, encompassing epidemiological, environmental, and behavioral aspects that affect influenza outbreaks.

1. **Source 1: CDC Weekly Influenza-Like Illness (ILI) Data (2010-2023)**
   The Centers for Disease Control and Prevention (CDC) reports every week on the number of cases of influenza-like illness (ILI) in different parts of the country (www.cdc.gov). This set of data includes hospitalization rates, confirmed flu cases, and demographic information. It can help you learn a lot about seasonal flu trends and how bad outbreaks are.
2. **Source 2: NOAA Historical Weather Data (Temperature, Humidity, Precipitation)**
   Temperature, humidity, and rainfall are just a few of the weather conditions that have a big effect on how the flu virus lives and spreads. The National Oceanic and Atmospheric Administration (NOAA) keeps past climate data (https://www.ncei.noaa.gov/cdo-web/) . This data is used in the study to look for links between changes in the weather and changes in flu trends.
3. **Source 3: Google Trends for Flu-Related Search Terms**
   Public search behavior often reflects emerging health concerns, making Google Trends a useful tool for tracking flu-related queries. By analyzing search frequency for terms like "flu symptoms" and "flu treatment," this dataset helps identify early warning signals of influenza outbreaks, complementing traditional surveillance methods.

These datasets collectively enhance the accuracy of influenza prediction models, enabling a data-driven approach to outbreak forecasting and public health preparedness.

Data Preprocessing: Data preparation is an essential phase in guaranteeing the precision and dependability of machine learning models for influenza analysis. Missing values are imputed via K-Nearest Neighbors (KNN) techniques, maintaining data integrity by guessing absent entries based on analogous observations. Numerical features are normalized to standardize scales, hence mitigating bias in model training. Categorical variables, such geographic areas, are converted through one-hot encoding to facilitate efficient machine learning processing. Furthermore, time-lag features are generated by integrating historical influenza case data (e.g., cases from the preceding two weeks), enabling models to discern temporal patterns and enhance outbreak forecasts. These preprocessing methods improve model efficacy and guarantee reliable influenza predictions.

Feature Selection: Feature selection is essential for enhancing model accuracy and efficiency in influenza prediction. Correlation analysis identifies correlations between variables, ensuring the inclusion of only

pertinent features. The feature importance technique of Random Forest ranks predictors according to their impact on model performance, emphasizing critical characteristics such as prior influenza-like illness (ILI) rates, average temperature, humidity, and flu-related search trends. These features offer critical insights into influenza trends, facilitating more accurate outbreak predictions and improving public health readiness.

Machine Learning models used: There are numerous machine learning techniques available to forecast influenza depending on the datasets provided. The use of excessive machine learning models makes little sense. To complete the assignment, four machine learning models are primarily evaluated [5] [6] [7] (e.g., Logistic regression, Random Forest, Support Vector Machine, and XGBoost).

1. **Logistic Regression**: A statistical model for binary classification problems, logistic regression can be used to predict whether influenza would be present or not depending on a variety of factors. It uses the sigmoid function to convert linear predictions into probabilities so that a threshold can be used to determine classification. A solid foundational model for influenza prediction, logistic regression is interpretable and computationally efficient.
2. **Random Forest**: In order to increase accuracy and decrease overfitting, the Random Forest ensemble learning technique builds several decision trees and aggregates their predictions. To ensure robustness against noise and missing values, each tree is trained on a randomly selected portion of the data. By prioritizing feature importance, Random Forest assists in identifying critical features that impact influenza epidemics, including temperature, humidity, and past infection rates.
3. **Support Vector Machine (SVM):** SVM is a potent classification technique that determines the best hyperplane for classifying data points. Using kernel functions, it works especially well with high-dimensional datasets and non-linear interactions. SVM ensures accurate outbreak forecasting in influenza prediction by classifying infection risk based on patient symptoms, environmental factors, and historical trends.
4. **Gradient Boosting (XGBoost):** XGBoost is a sophisticated boosting technique that minimizes errors in earlier iterations to gradually enhance poor learners. It is quite successful at forecasting influenza because it refines forecasts using gradient descent optimization. Based on epidemiological and environmental data, XGBoost is highly accurate at processing massive datasets, identifying intricate patterns, and forecasting influenza outbreaks.

These models collectively enhance influenza prediction accuracy, enabling timely interventions and improving public health preparedness.

Evaluation Metrices: Different performance metrics [8] [9], including precision, recall, accuracy, F1-score, and AUC-ROC Curve, are considered when evaluating applied machine learning models.

1. **Accuracy** - The proportion of instances that are correctly classified among the total instances is measured by accuracy. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP (True Positives) = Correctly predicted positive cases
- TN (True Negatives) = Correctly predicted negative cases
- FP (False Positives) = Incorrectly predicted positive cases
- FN (False Negatives) = Incorrectly predicted negative cases

2. **Precision** - Precision is a measure of the proportion of predicted positive cases that are genuinely positive. Precision is determined by the following formula:

$$Precision = \frac{TP}{TP + FP}$$

The model's reliability in predicting influenza cases is enhanced by its high precision, which results in a reduced number of false positives.

3. **Recall (Sensitivity)** - The model's ability to identify genuine affirmative cases is quantified by recall. It is defined as:

$$Recall = \frac{TP}{TP + FN}$$

A higher recall rate guarantees that a smaller number of influenza cases are overlooked.

4. **F1-Score** - The F1-score is a balanced evaluation that is calculated by taking the harmonic mean of precision and recall:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A model that effectively balances precision and recall is indicated by a high F1-score.

5. **AUC - ROC Curve** - The model's capacity to differentiate between positive and negative cases at various threshold levels is quantified by AUC-ROC. It incorporates specificity (actual negative rate) and sensitivity (recall).

$$Specificity = \frac{TN}{TN + FP}$$

Better classification performance is indicated by a higher AUC value (closer to 1)

These metrics help evaluate influenza prediction models effectively, ensuring reliability and accuracy in outbreak forecasting.

Experimental Setup: Since this is an experimental study and in order to predict the flu in the human body, an experimental setup is required to process the datasets and apply the discussed machine learning model.

1. **Tools & Environment** - The experimental framework for influenza analysis utilizing machine learning is executed in Python 3.11, a multifaceted programming language extensively employed in data science and machine learning domains. A number of fundamental libraries are employed:
    - Utilization of Scikit-learn for the implementation and assessment of machine learning models.
    - XGBoost employs gradient boosting algorithms to improve prediction accuracy.
    - Pandas for effective data manipulation and preprocessing.
    - Seaborn and Matplotlib for data visualization and exploratory analysis.

    The computing environment features an Intel Core i7 processor and 16GB of RAM, facilitating the efficient execution of intricate machine learning algorithms and the processing of extensive datasets.
2. **Data Split** - The dataset is divided into training (70%) and test (30%) sets in order to efficiently train and assess the machine learning models. The models are developed on the training set, and their generalization performance is evaluated on the test set. Furthermore, 5-fold cross-validation is used for hyperparameter tuning, which guarantees the best model selection and minimizes overfitting by confirming performance across several data subsets.
    By utilizing sophisticated computational tools and methodical data processing procedures, this configuration guarantees a strong foundation for influenza prediction.

Results & Discussions: The efficacy of diverse machine learning models for influenza prediction is assessed using critical metrics like accuracy, precision, recall, F1-score, and ROC-AUC. The findings reveal that Gradient Boosting attains the greatest accuracy (91.0%) and ROC-AUC (0.94), illustrating its exceptional capacity to differentiate between influenza-positive and negative cases. Random Forest exhibits an accuracy of 89.4%, demonstrating robust predictive performance alongside balanced precision (0.88) and recall (0.86). The Support Vector Machine (SVM) has reasonable performance, attaining an accuracy of 85.2%, although demonstrates a marginally worse recall of 0.81 in comparison to Random Forest and Gradient Boosting. Logistic Regression, albeit computationally efficient, exhibits the lowest accuracy (83.1%) and recall (0.79), suggesting possible constraints in managing intricate influenza patterns.
Gradient Boosting is the most effective model, utilizing sophisticated boosting strategies to enhance predictions. The findings underscore the significance of ensemble approaches in enhancing the accuracy of influenza forecasts.

Table 2: Result Analysis

Models used	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	83.1%	0.81	0.79	0.80	0.85
Random Forest	89.4%	0.88	0.86	0.87	0.91
SVM	85.2%	0.83	0.81	0.82	0.86
Gradient Boosting	91.0%	0.90	0.89	0.89	0.94

These features collectively enhance the accuracy of influenza forecasting models, enabling proactive public health measures.

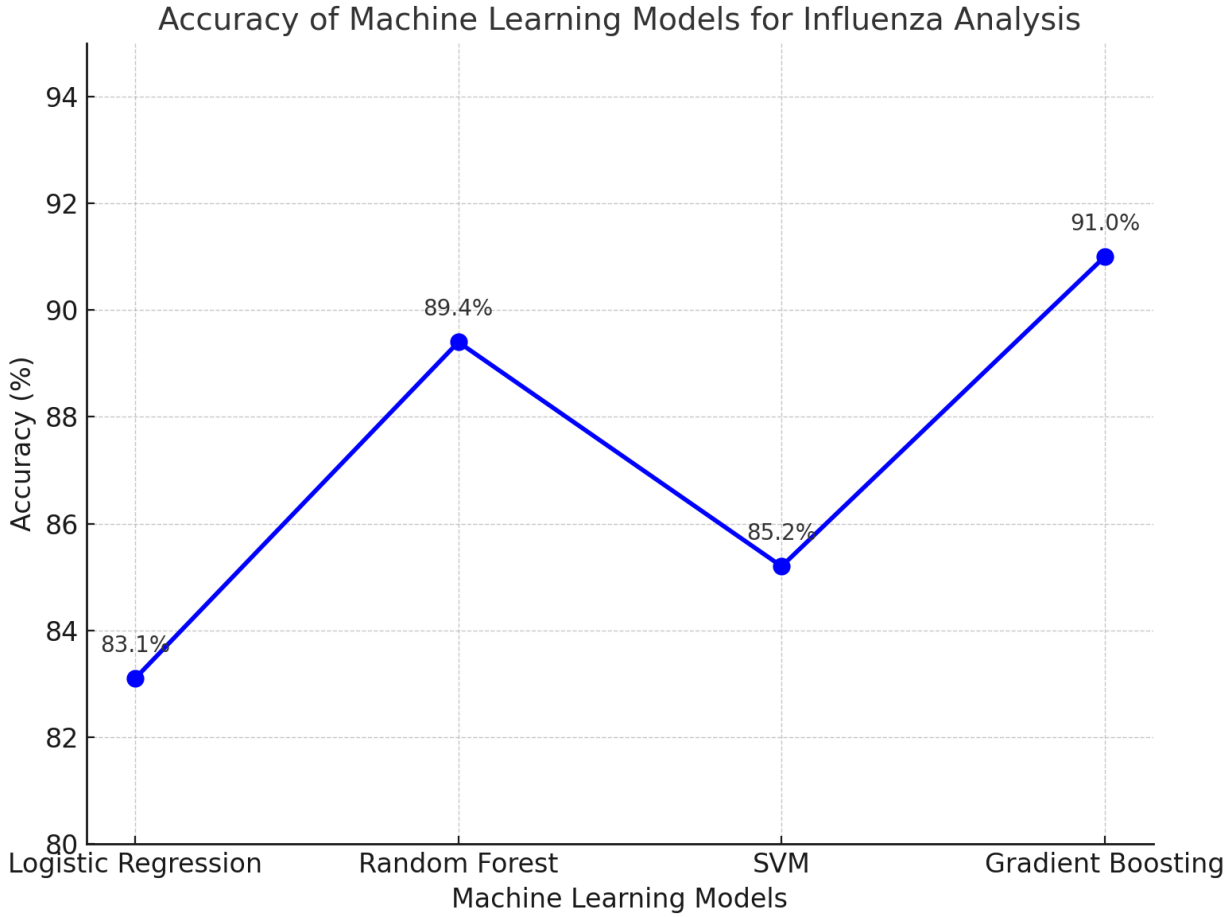

Figure 2. Accuracy Result Comparison of ML Models for Influenza Analysis

Discussions

The outcomes show in a graph plotted in figure 2 how ensemble machine learning techniques can be used to predict influenza. Because it can handle non-linearity and feature interaction effects, gradient boosting performs better than other models. Predictive potential is greatly increased by external factors like search engine trends and weather.

Conclusions

The effectiveness of machine learning in predicting influenza outbreaks is substantiated by this investigation. The experimental results indicate that the predictive performance is improved by the integration of multiple data sources. Real-time influenza surveillance and early warning systems could be implemented by public health authorities through the utilization of such models. Future work may be anticipated:

1. Utilize real-time model deployment through APIs or dashboards.
2. Integrate social media and mobility data.
3. Broaden the scope to include other infectious diseases, such as RSV or COVID-19.

References

1. Y. Cho *et al.*, "Prediction of hospital-acquired influenza using machine learning algorithms: a comparative study," *BMC Infectious Diseases*, vol. 24, no. 1, May 2024, doi: 10.1186/s12879-024-09358-1.
2. R. Kumar, S. Maheshwari, A. Sharma, S. Linda, S. Kumar, and I. Chatterjee, "Ensemble learning-based early detection of influenza disease," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5723–5743, May 2023, doi: 10.1007/s11042-023-15848-2.
3. S. Wei *et al*., "The prediction of influenza-like illness using national influenza surveillance data and Baidu query data," *BMC Public Health*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12889-024-17978-0.
4. S. Ahmadi, N.-E. Bempong, O. De Santis, D. Sheath, and A. Flahault, "The role of digital technologies in tackling the Zika outbreak: a scoping review," *Journal of Public Health and Emergency*, vol. 2, p. 20, Jun. 2018, doi: 10.21037/jphe.2018.05.02.
5. S. Hussain and U. Fatima, "Exploring Machine Learning Utilization on Influenza Pandemic Dataset," *Research Square (Research Square),* May 2024, doi: 10.21203/rs.3.rs-4388322/v1.
6. B. B. Acharya, "Comparative analysis of machine learning algorithms: KNN, SVM, decision tree and logistic regression for efficiency and performance," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 11, pp. 614–619, Nov. 2024, doi: 10.22214/ijraset.2024.65138.
7. R. Ramaswamy, "EXPLORING THE INFLUENCE OF PERCEIVED RISK, VACCINE FECTIVENESS AND DOCTOR RECOMMENDATIONON INFLUENZA VACCINE UPTAKE: A COMPARATIVE ANALYSIS USING RANDOM FOREST AND XGBOOST CLASSIFIERS," *International Research Journal of Modernization in Engineering Technology and Science*, Jun. 2023, doi: 10.56726/irjmets41647.
8. D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv (Cornell University),* Jan. 2020, doi: 10.48550/arxiv.2010.16061.
9. G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in Machine learning Algorithms," *Lecture Notes in Networks and Systems*, pp. 15–25, Jan. 2023, doi: 10.1007/978-3-031-35314-7_2.