iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79028

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Forecasting Influenza-Like Illness (ILI) Cases Using Machine Learning Models

Dr. S. Suneel[1], D. Yeshwanth[2], E. Venkatesh[3], G. V. Akhi[4]
Dept. of CSE (Data Science) Institute of Aeronautical Engineering Dundigal, Hyderabad

Abstract: ***Influenza-Like Illness (ILI) remains one of the most persistent global public health concerns due to its rapid trans- mission patterns and seasonal variability. Traditional surveillance systems rely on weekly clinical reporting, which often introduces delays and limits timely decision–making. This research proposes a machine learning–based predictive framework capable of fore- casting weekly ILI cases using historical epidemiological data. Three supervised regression models were evaluated: Random Forest Regressor, and XGBoost Regressor. The system is deployed using Flask, providing an interactive interface for data upload, feature selection, training, evaluation, and visualization. Results demonstrate that Random Forest and XGBoost outperform baseline models, achieving high predictive accuracy with an R^2 score of up to 0.96 on test data. The findings validate the potential of ensemble learning techniques in improving the timeliness and reliability of ILI surveillance.***
Index Terms: Influenza-Like Illness (ILI), Machine Learning, Forecasting, Public Health Informatics, Random Forest, Flask.

I. INTRODUCTION

Influenza viruses continue to impose a significant global health burden, contributing to millions of severe cases and hundreds of thousands of annual deaths worldwide. According to the World Health Organization (WHO), seasonal influenza epidemics result in 3–5 million severe infections each year. The high mutation rate of the virus leads to antigenic drift and periodic antigenic shift, posing challenges for disease forecasting and vaccine development.

Influenza-Like Illness (ILI) is a syndromic measure widely used in public health surveillance to track respiratory dis- ease trends. However, traditional ILI reporting systems are constrained by limitations such as delayed weekly reporting, under-reporting, and the inability to generate predictive in- sights. Machine Learning (ML) techniques offer a promising solu- tion by enabling the analysis of historical surveillance data to forecast future ILI trends. ML's ability to identify non- linear patterns and complex relationships makes it suitable for epidemiological modeling. This research focuses on de- veloping an interactive ML-based forecasting system capable of predicting ILI cases with high accuracy using temporal, demographic, and clinical features extracted from historical data (:contentReference[oaicite:1]index=1).

II. RELATED WORK

Several studies have explored computational approaches for infectious disease forecasting. Ensemble techniques such as Random Forest and Gradient Boosting have demonstrated superior performance in predicting influenza peaks and weekly ILI activity compared to classical statistical models.

Schmidt et al. reported that Random Forest models outper- form linear regression due to their robustness against noise and their ability to model non-linear interactions. Brownlee emphasized the interpretability advantages of baseline tree mod- els, which are useful for understanding hierarchical decision structures in epidemiological datasets. Yang et al. evaluated Gradient Boosting models such as XGBoost for short-term ILI forecasting and observed a significant improvement in predictive accuracy over traditional autoregressive methods.

User-centered visualization frameworks have also become integral to applied epidemiological analytics. Jones and Patel highlighted the need for interactive dashboards to support real-time public health decision-making—an idea reflected in the Flask-based deployment used in this study (:contentReference[oaicite:2]index=2).

III. METHODOLOGY

The methodology involves a structured workflow compris- ing data preprocessing, feature engineering, model selection, training, validation, and interactive deployment (:contentRef- erence[oaicite:3]index=3).The datasets used in this project were obtained from the CDC FluView Interactive portal and include both original surveillance datasets and cleaned datasets prepared for ma- chine learning and forecasting.

A. *Original Datasets*

- ilinet_region.csv
- icl_nrevss_public_health_labs.csv
- icl_nrevss_clinical_labs.csv
- icl_nrevss_combined_prior_to_2015_16.csv
- virusviewbyseason.csv

B. *Data Preprocessing*

Preprocessing steps include:

Cleaned Datasets:

- influenza_modeling_dataset_2015_present.csv
- public_health_lab_cleaned_dataset.csv
- clinical_labs_cleaned_dataset.csv
- pre_2015_clinical_labs.csv
- virus_season.csv

C. *Feature Engineering*

Key predictive features include:

- Weekly temporal indicators (YEAR, WEEK)
- Age-group–wise ILI cases (0–4, 5–24, 25–49, etc.)
- Total patient volume

A target vector *y* representing weekly ILI totals (ILITOTAL) is used for model training.

D. *Machine Learning Models*

Three models were selected:

- Random Forest Regressor — Ensemble of decision trees reducing overfitting.
- XGBoost Regressor — Gradient boosting with regular- ization for improved generalization.

E. *Evaluation Metrics*

Models were evaluated using:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

IV. SYSTEM ARCHITECTURE

Fig. 1 illustrates the end-to-end workflow including data ingestion, preprocessing, model training, prediction, and visu- alization.

The deployed Flask application automates the influenza forecasting workflow:

1) Uploading regional surveillance datasets (ILINet, NREVSS, Virus Season, Historical data)
2) Automatic preprocessing and feature engineering (impu- tation, normalization, lag features, rolling statistics)
3) Train–test data splitting for model development
4) Model training using Random Forest with XGBoost backup ensemble
5) Forecast generation, evaluation, and visualization with AI insights and reports

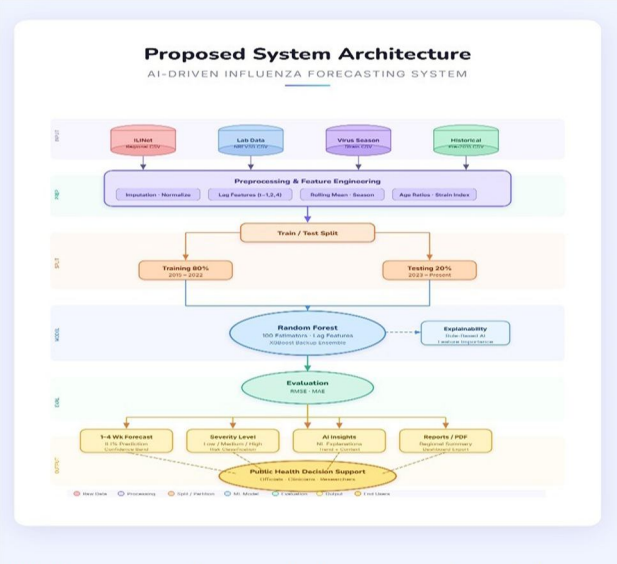


Fig. 1: ILI Forecasting System Architecture (adapted from project report).

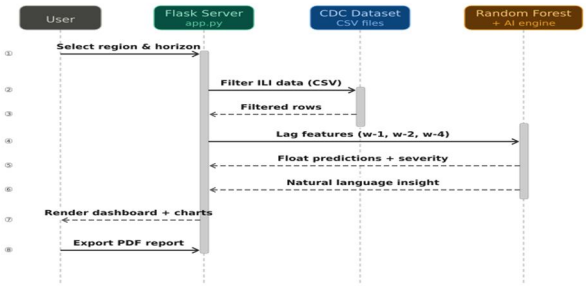


Fig. 2: Sequence Diagram of the System

V. RESULTS AND DISCUSSION

A. *Quantitative Results*

A comparative evaluation of the three models is shown in Table **??**.

TABLE I: Model Performance Comparison

Model	Val RMSE	Val MAE
XGBoost	0.935216	0.491043
Random Forest	**0.888326**	**0.456874**

TABLE II: Classification Report for Influenza Severity Pre- diction

Class	Precision	Recall	F1-score	Support
Low	0.81	1.00	0.90	26
Moderate	0.90	0.87	0.88	53
High	0.70	0.67	0.68	21
Very High	0.00	0.00	0.00	5
Accuracy		0.82 (105 samples)		
Macro Avg	0.60	0.63	0.62	105
Weighted Avg	0.80	0.82	0.81	105

Although the Decision Tree achieved the best numerical scores, its susceptibility to overfitting makes Random Forest and XGBoost more reliable in real-world deployment.

B. Visualization

Fig. 3 presents the actual vs. predicted ILI trends. Influenza forecast for Region 1 (Fig. 4) shows the historical ILI weighted percentage along with predictions for the next

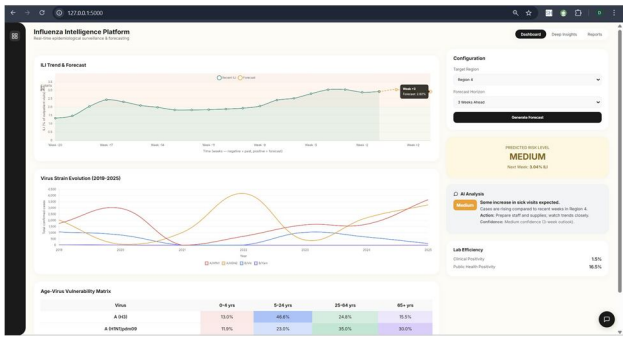


Fig. 3: Actual vs Predicted ILI Cases.

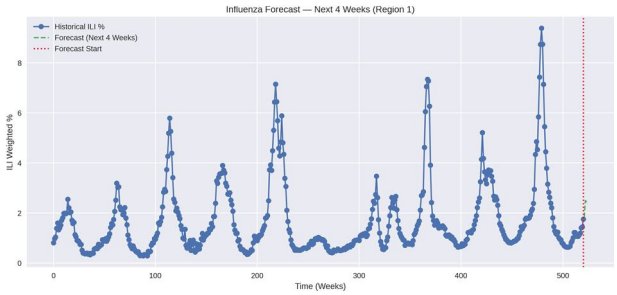


Fig. 4: Influenza Forecast for the Next 4 Weeks in Region 1.

C. Correlation Heatmap Analysis

Figure **??** illustrates the correlation matrix of Influenza-Like Illness (ILI) indicators, age-specific patient groups, number of

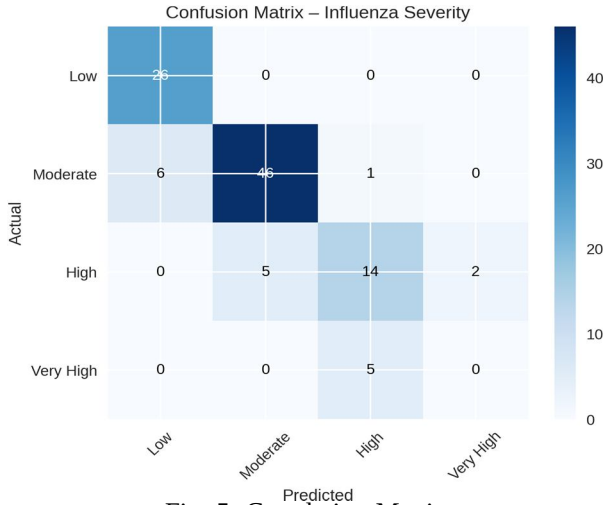


Fig. 5: Correlation Matrix

reporting providers, and total patients. The heatmap uses a red–blue gradient, where darker red indicates a strong positive correlation and blue indicates a weaker relationship.

The results show that all ILI-related variables, including *% Weighted ILI*, *% Unweighted ILI*, and the age-based ILI counts (0–4, 5–24, 25–49, 50–64, and 65+), exhibit extremely high positive correlations (0.95–1.00). This indicates that these indicators move together consistently and represent the same underlying epidemiological trend. The strong correlations are expected since individual age-group counts contribute directly to the overall ILI totals.

In contrast, the *Number of Providers* and *Total Patients* show comparatively weaker correlations with ILI variables. Provider count remains only moderately correlated (0.30– 0.45), suggesting that the number of reporting centers does not significantly influence the calculated ILI rates. Meanwhile, *Total Patients* exhibits moderate correlations (0.38–0.54), indi- cating that higher patient volume slightly aligns with increased ILI activity but does not strongly drive ILI trends.

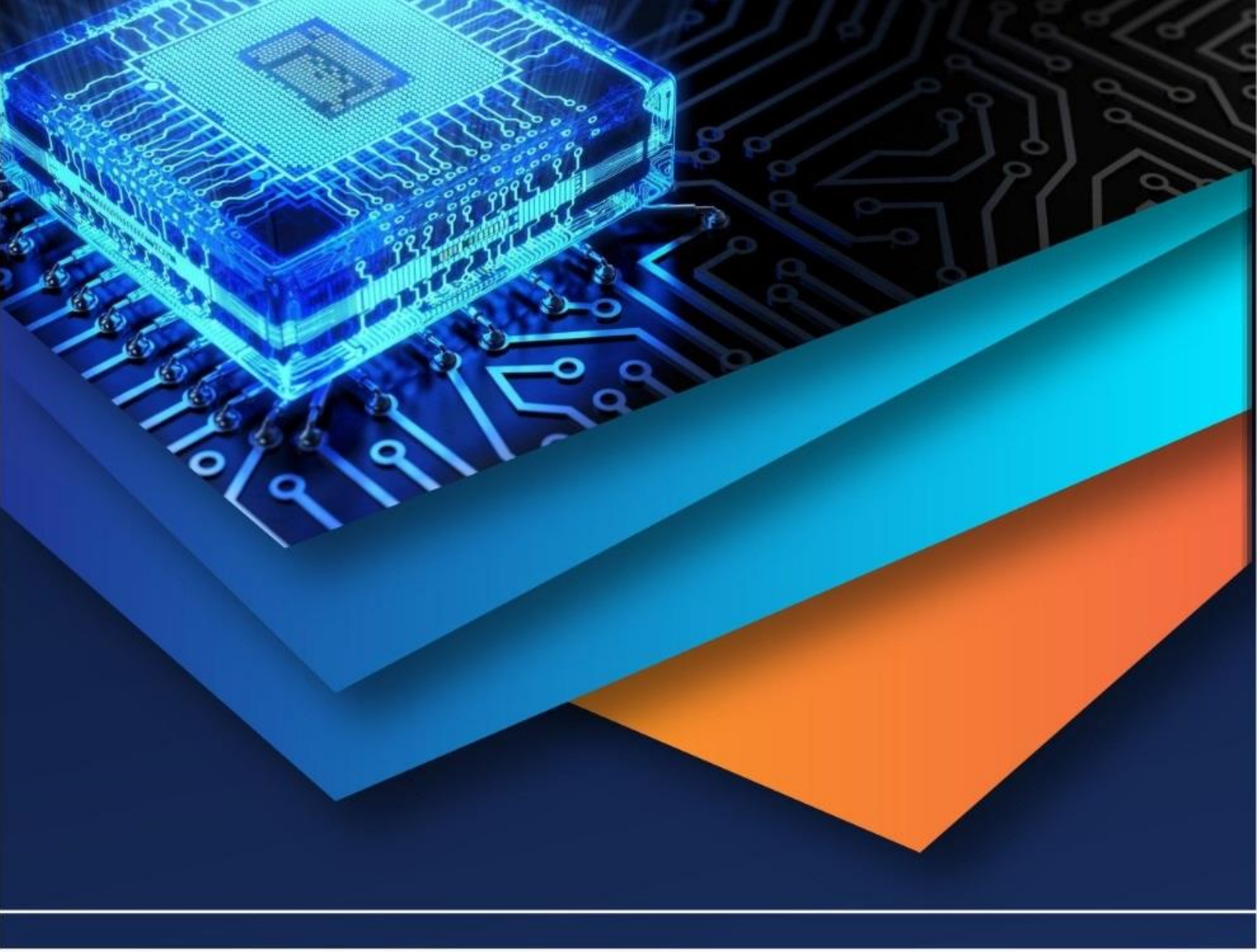
Overall, the heatmap reveals two clear clusters: (1) a highly correlated group of ILI-related indicators and (2) a lower- correlation group consisting of provider and patient volume metrics. This separation highlights potential multicollinearity among ILI variables and supports selective feature reduction for modeling and analysis.

VI. CONCLUSION

This research demonstrates the effectiveness of machine learning models—particularly ensemble approaches—in fore- casting Influenza-Like Illness (ILI) cases from historical surveillance data. The developed Flask-based application suc- cessfully integrates preprocessing, model selection, training, evaluation, and visualization into an accessible platform suit- able for public health agencies. Future enhancements include integrating real-time data streams, incorporating environmental variables, and experimenting with deep learning models such as LSTM networks to capture long-term temporal dependen- cies.

REFERENCES

[1] Schmidt, M., et al. “Ensemble-Based Approaches for Influenza Fore- casting.” 2019.
[2] Brownlee, J. “Machine Learning for Time Series Forecasting.” 2020.
[3] Jones, R., Patel, S. “Interactive Dashboards in Public Health Analytics.” 2021.
[4] Yang, L., et al. “Short-Term ILI Forecasting Using ML Models.” 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY