Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Predicting Flu Seasons using Machine Learning and Diverse Data

Shobhit Mattoo, Aaryan Diwan
Department of Computer Science and Engineering
Galgotias University, Greater Noida, India

Abstract **- The seasonal influenza has been dominating the healthcare systems because of its quick modes of spread and predictable annual outbreaks. Consistent prediction of influenza trends can assist hospitals and other public health agencies to prepare resources beforehand; although most of the current surveillance systems use delayed reports and few sources of data. The presented paper offers the framework of machine learning in predicting the activity of the flu by incorporating the information provided by several domains. The proposed method does not rely on one signal, but uses the previous data on influenza cases along with the weather and the features of human behaviour affecting the spread of the disease. A number of machine learning models, such as the Random Forest, Support Vector Machine, and Long Short-Term Memory (LSTM) networks, are tested on previous outbreak data. The findings indicate that multisource data integration is better in forecast accuracy and stability, and LSTM models are the most effective.**

Index Terms **- Keywords: Seasonal influenza, prediction of outbreaks, multisource data, machine learning, time-series analysis.**

I. INTRODUCTION

A. Background and Motivation

Seasonal influenza is a recurrent respiratory disease that is primarily brought about by influenza A and B virus and still affects millions of people globally annually. In addition to direct health impacts, influenza exerts a long-lasting strain on the healthcare systems because it causes more outpatient care, inpatient hospitalizations, and medical staff and resource demands. In severe seasons of influenza, hospitals tend to lack beds, equipment, and staff, which indicates the necessity of early preparedness and the effective distribution of resources.

Proper prediction of the influenza outbreaks is an important factor in decreasing this load. Vaccination, hospital staffing, and allocation of medical supplies are better planned and controlled by the healthcare authorities through early predictions. Nevertheless, it is still difficult to predict the trends of influenza because the process of transmitting the disease is complicated and dynamic. Several interacting factors such as environmental factors, population movement and human behavior pattern contribute to the spread of influenza and they differ depending on regions and seasons.

Temperature and humidity are two of the weather conditions which have a great influence on the survival and transmissibility of the virus. Extreme conditions of cold and dry temperatures have been known to increase viral longevity, and the population density and the number of social interactions per day affect the rate of the spread of infections. The conventional method of influenza surveillance is mainly based on clinical diagnosis and laboratory causes. These systems are also reliable but usually have delays in reporting hence most of the responses taken by the public health are in the form of reacting and not preventing.

The new opportunities to enhance the forecasting of influenza can be introduced by recent progresses of machine learning along with the increased availability of digital, environmental, and behavioral data. Machine learning models have the ability to identify subtle patterns and early warning signs that a traditional surveillance system alone would fail to detect because of the heterogeneous nature of the data they handle. Such transition to data-based forecasting can greatly improve the decision making process of the population.

B. Problem Statement

Notwithstanding all the research, most current forecasting models of influenza are still based on small data sets or individual sources, and often they consider only historic numbers of cases. This kind of approaches is not able to resonate with sudden outbreak changes due to variations in the environment and change in human behavior. Consequently, they tend to perform poorly in forecasting when there is an unforeseen increase in the level of infection.

Conventional statistical models often make the assumption of linearity or stationary behaviour which simplifies the behaviour of real-world disease. Such assumptions restrict their capacity to respond to complicated, nonlinear tendencies that occur in epidemiological data. Moreover, the use of models that do not consider the behavioral or environmental effects does not capture how influenza is transmitted as a multifactor process.

It becomes apparent that there is a necessity of having a combined forecasting framework that will be able to merge various sources of data and adjust to the constantly changing environment. Machine learning algorithms are in fact quite efficient at the same, since they are able to learn complicated relationships on noisy and high-dimensional data without having to rely on rigid predetermined assumptions. Utilization of these techniques can result in better and timely predictions of the influenza.

C. Research Objectives and Contributions

This study will contemplate a multisource predictive model of seasonal influenza, by incorporating epidemiological data, meteorological data and indicators of human behavior. The key

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

goals are to enhance the accuracy of its forecasts, to enhance its ability to detect outbreaks early in their progression, and to help effective proactive actions by its public health. The major contributions of the work are:

- Creation of an all-purpose, data-driven influenza prognostic model.
- Ensemble, kernel, and deep learning models were compared to each other.
- Comparison of the effect of data fusion between multiple sources on prediction.
- Evidence of the efficacy of the LSTM networks in the context of temporal disease patterns modeling.

II. LITERATURE REVIEW

The models used to do early influenza forecasting were mostly based on compartmental epidemiology of SusceptibleInfected-Recovered (SIR) and Susceptible-Exposed-InfectedRecovered (SEIR). These models subdivide the populations into discrete compartments and model the disease progression based on the use of differential equations. Although these are useful in the overall dynamics of transmission, they presuppose homogeneous mixing, and constant parameters which restricts their application in real life.

Influence forecasting was then applied to the use of timeseries statistical models, with AutoRegressive Integrated Moving Average (ARIMA) being one of them. These models only reproduce the temporal correlation of historical data, and they tend to be incapable of nonlinear behavior and external effects. Consequently, their predictive capabilities are poor especially when it comes to making multi-week predictions.

In an attempt to overcome these constraints, scientists initiated the investigation of machine learning algorithms that can be applied to estimate complicated relationships in epidemiological data. The popularity of the Random Forest models lies in their strength, capacity to handle noisy data, and the possibility to interpret their results by feature importance scores. The Support Vector Machines proved to be effective in high dimension space and used in short term influenza prediction.

Traditional models of machine learning are however, not normally good at capturing long term temporal dependencies in the spread of disease. The solution to this dilemma was the development of deep learning methods, especially Long Short-Term Memory networks. The LSTM networks are characterized by memory cells and gating processes which enable them to remember a long sequence of data hence they are the best in modeling time-series epidemiological data.

It has been demonstrated that LSTM-based models can better predict trends in influenza-like illnesses than traditional methods, particularly when external sources of data are used. The studies also highlight the significance of data integration, which is composed of multiple sources, such as climatic data and behavioral aspects like active search on the internet. These supplementary signals alleviate reporting time and enhance the rapid outbreak detection.

III. SYSTEM ARCHITECTURE

The suggested system is developed based on a modular and scalable architecture that reflects the real-life influenza forecasting processes. Multiple sources are used in the collection of data such as epidemiological surveillance systems, meteorological databases, and online behavioral platforms. All the data sources represent a different facet of the process of influenza transmission, which makes the picture of the dynamics of an outbreak more complete.

Prior to the process of model training, every dataset is processed through data cleaning, temporal alignment, and normalization. Temporal alignment makes sure that data in various sources refers to the same weekly basis, and normalization makes sure that no single feature takes an advantage in the model training.

Upon preprocessing, the important temporal and seasonal characteristics are ascertained and presented to machine learning models. The architecture is designed in a modular manner which makes it easy to add other data sources or forecast models over time. The standard statistical measures are used to assess model outputs so as to compare these methods fairly and consistently.

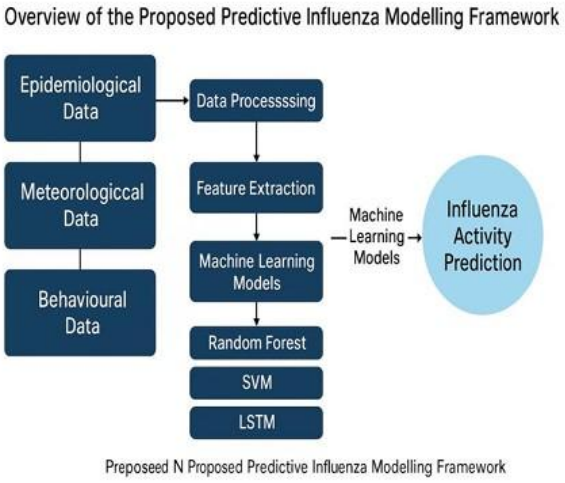

Fig. 1. Overview of the proposed predictive influenza modelling framework

IV. METHODOLOGY

The central data on the study is the number of cases of influenza per week. These data are combined with the meteorological factors like the average temperature and the relative humidity, which affect the survival of the virus. Online search behavior on flu symptoms yields behavioral measures of proxies of public health awareness and early infection trends.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

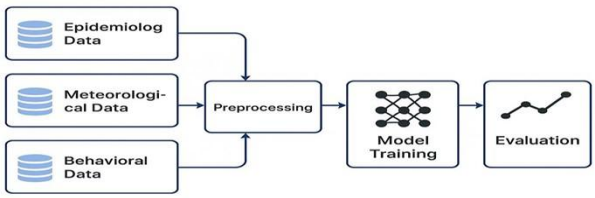

Fig. 2. System architecture diagram showing data input, preprocessing, model training, and evaluation

Preprocessing of data includes the processing of missing data, scale of numerical data, and alignment of all data to a common weekly schedule. The steps are essential to assure stability and quality of the data.

The field of feature engineering is aimed at representing the temporal patterns in terms of lagged infection counts, rolling averages, and seasonal indicators. Lagged variables are used to make models consider delayed effects in the transmission of disease, whereas rolling averages average short-term variations.

There are three machine learning models analyzed. Random Forest Regression is adopted to represent the nonlinear relationships and is interpretable by analysing the feature importance. The support vector machine regression is used due to its strength on high-dimensional spaces. Long Short-Term Memory networks are also used to represent sequential dependencies and temporal patterns of influenza data in the long term.

V. RESULTS AND FINDINGS

The tool of proposal influenza forecasting is tested in terms of the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or the coefficient of determination R^2.. The combination of these measures evaluates the mean prediction error, the sensitivity to the large deviations and the goodness of fit in general. A smaller MAE and RMSE value means a better accuracy whereas a higher value of R2 is a good sign of a better agreement between predicted and observed influenza trends.

The Long Short-Term Memory (LSTM) network is the best performer in terms of all metrics as compared to other models evaluated. The LSTM model has the lowest MAE and RMSE rates, which proves that it is effective to consider temporal dependencies and reduce prediction errors. It also achieves the greatest R2 scores meaning that it has a high explanatory power and is strongly modeled in the trends of influenza over time.

Random Forest regression is a competitive model that scores second in most of the measures of evaluation. Its ensemble nature enables it to provide nonlinear models of the relationship between influenza cases and external variables in the form of meteorological variables and behavioral indicators.

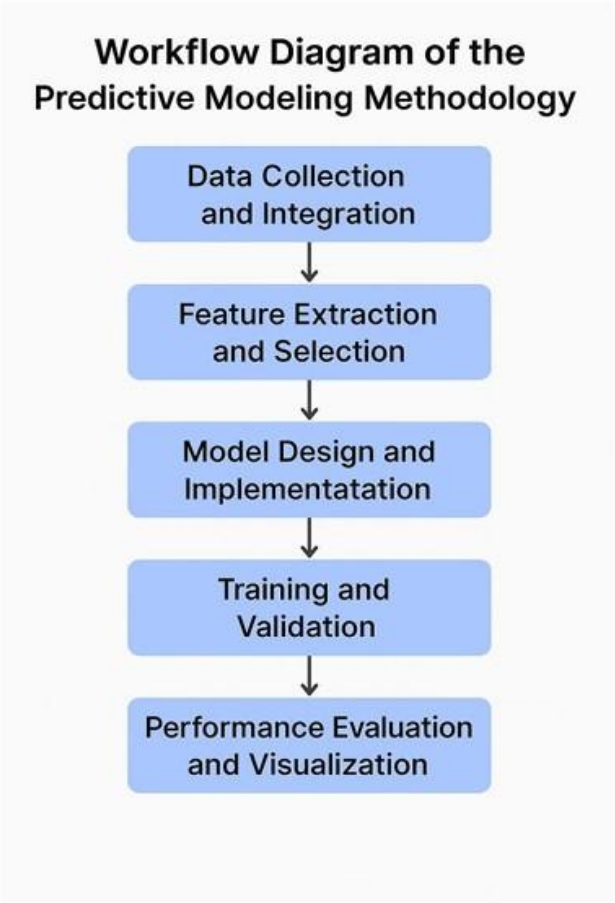

Fig. 3. Workflow diagram of the predictive modelling methodology

Although its accuracy is slightly decreased compared to LSTM model, the Random Forest has stable predictions and useful interpretability based on the analysis of feature importance.

SVM regression demonstrates relatively decreased predictive power. Due to its lack of extrapolative capture of longterm temporal relationships, SVM models are not very effective in influenza forecasting, especially model prediction with a time horizon of several weeks.

The benefit of data integration on multisources is also a major finding of this study. Models that are trained based on a combination of epidemiological, meteorological and behavioral data are better than those that are based on individual inputs. Weather factors are used to predict the seasonal variations, and the behavioral indicators based on online search activity are frequently used as a warning signal of the onset of increasing influenza activities. The combination of these different data sources enhances more accuracy and robustness of prediction and the sensitivity to noises and reporting delays decreases.

Comprehensively, the findings indicate that deep learningbased solutions, especially LSTM networks, are quite useful in seasonal influenza prediction. Multisource data fusion improves the stability of prediction and detection of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

early risk, which is useful to apply machine learning models to practice in the surveillance and planning of public health.

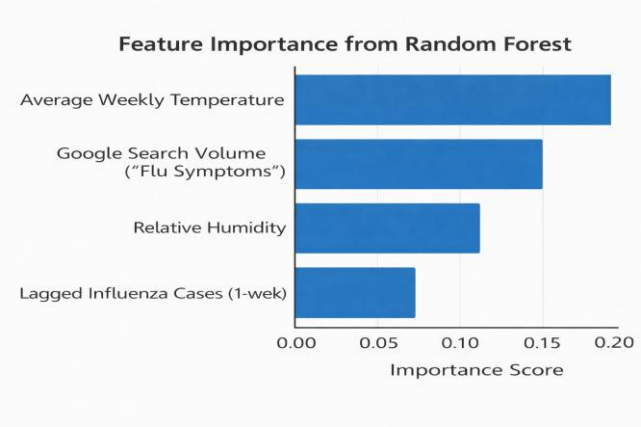

Fig. 4. Feature importance ranking obtained from the Random Forest model

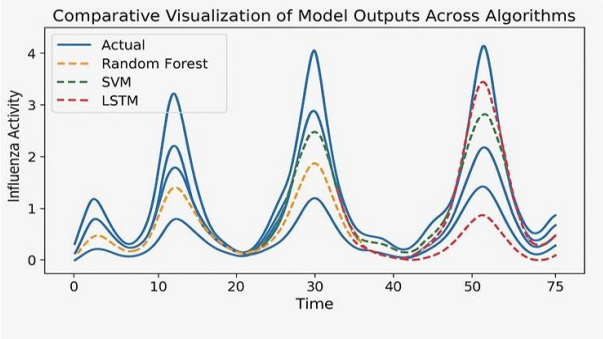

Fig. 5. Comparative visualization of model outputs across algorithms

VI. DISCUSSION

This research results in the realization of the efficiency of machine learning methods and especially deep learning models in the process of predicting seasonal patterns of influenza when various data sources are combined. The effectiveness of the Long Short-Term Memory (LSTM) network in its long-standing good performance has shown that the time-dependence in epidemiological time-series data can be well modeled. Measurement of the transmission of influenza changes with time and is subject to the time lag effects of both the environment and behavior and therefore memory based models are particularly appropriate in this task.

Among the most important lessons of this piece of work is the usefulness of multisource data integration. The conventional use of influenza forecasting relies solely on the number of cases in the past and this restricts its responsiveness to the abrupt variations in the transmission patterns. The proposed framework also includes meteorological variables and behavioral indicators, which will represent a wider picture of the actual factors in the real world that influence the spread of influenza. The behavioral indicators like the online search activity of flu symptoms are often forewarners of the increase in the level of infection and therefore the forecast can be made to predict the outbreaks, even before the confirmed cases reports are announced.

The weather also has a significant role in explaining seasonal patterns. Humidity and temperature have an effect on viral survival and efficiency of transmission and consideration of them enhances the capability of the model to capture variability due to season. This is because epidemiological, climatic and behavioral data together improve the predictive accuracy and sensitivity, minimizing the sensitivity to noise as well as the delay in reporting which are typical of surveillance systems.

Although LSTM models have been shown to be accurate in their predictions, the Random Forest models have complementary advantages. Their ensemble form enables them to effectively capture nonlinear relationships and their analysis of feature importance can be of great use in terms of their interpretability. This interpretability can be of vital importance in public health in both understanding what factors contribute the most to the spread of the disease, as well as confidence in the data-driven decision-making. Whereas the Support Vector Machine models are less accurate in this study, they show steady performance and can potentially find use in an environment with limited computational capabilities.

In practical terms, the findings indicate that machine learning-based prediction systems can be used to help in proactive population health planning. Timely and high-quality predictions allow medical services to organize resources, plan vaccination processes, and introduce specific interventions in time before the outbreak reaches its peak. Nevertheless, the implementation of such systems in practice must be regarded with the attention to the availability of data, maintenance of the model, and ethical issue regarding the privacy of data.

On the whole, this research paper shows that the combination of multisource information and sophisticated machine learning algorithms can enhance seasonal forecasts of influenza to a considerable extent. The issue of the trade-off of predictive accuracy and interpretability is still relevant, yet the benefits of the deep learning methods, in particular, LSTM networks, allow turning them into a promising instrument to improve the state of preparedness and public health surveillance.

VII. CONCLUSION

The paper proposes an extensive machine learning-driven model of predicting seasonal influenza through the combination of epidemiological, meteorological and behavioral data obtained across several sources. The proposed system will allow a greater number of factors to be factored into influenza transmission and more accurate and timely predictions due to the abandonment of conventional single-source surveillance methods. The findings show that multisource data fusion is very useful in improving the accuracy and stability of the forecasts especially in cases where there is delay in reporting and in cases where real-world data is noisy.

The Long Short-Term Memory (LSTM) networks are the best performing of all the evaluated models in all evaluation metrics. The fact that they can model temporal dependencies allows them to model both short term fluctuations and long term seasonality in influenza activity. Although the models based on the Random Forests are helpful in interpretability and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

high accuracy, and Support Vector Machines demonstrate the ability to provide stable baseline performance, LSTM networks become the most useful models when it comes to time-series disease forecasting in the context of this study.

The research results demonstrate the practical usefulness of machine learning in assisting early warning systems and proactive decision-making in relation to the population health. Effective influenza projections can also guide health professionals and policy makers to allocate resources optimally to plan on how to conduct their vaccinations as well as taking timely measures before the outbreaks are in full swing.

Future research can be aimed at the addition of real-time data streams, enhancement of the model generalization to the requirements of other geographic areas, and investigation of new models of deep learning, including attention-based or transformer models. The issue of applying the proposed framework to other seasonal infectious diseases could also make the framework even more relevant and effective. On the whole, the study proves the opportunities of data-driven solutions to enhance surveillance of people health and enhance their readiness to the reoccurrence of an infectious disease outbreak.

REFERENCES

[1] World Health Organization, “Influenza (Seasonal),” World Health Organization, Geneva, Switzerland, 2018.

[2] Centers for Disease Control and Prevention, “Influenza Overview,” CDC, Atlanta, GA, USA, 2014.

[3] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Q. Liu, H. Zheng, and X. Zhang, “Influenza trends prediction using LSTM networks,” in *Proc. Int. Symp. Bioinformatics Research and Applications (ISBRA)*, Beijing, China, 2018, pp. 1–10.

[8] Y. Yang, Y. Gao, and J. Zhang, “Influenza-like illness prediction using LSTM neural networks,” *Journal of Supercomputing*, vol. 76, no. 4, pp. 2301–2315, 2020.

[9] S. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: Detecting influenza epidemics using Twitter,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011, pp. 1568–1576.

[10] S. Xue, Z. Zhang, and Y. Li, “Regional influenza surveillance using Twitter data,” *PLoS ONE*, vol. 12, no. 12, e0189186, 2017.

[11] Y. Zhang, C. Zhang, and Z. Liu, “Computational prediction of influenza epidemics,” *Virologie*, vol. 10, no. 3, pp. 145–152, 2006.

[12] Z. Ertem, M. S. Brownstein, and D. V. Zeng, “Optimal multi-source forecasting of seasonal influenza,” *PLoS Computational Biology*, vol. 14, no. 9, e1006236, 2018.

[13] W. Zhu *et al.*, “Deep-learning model for influenza prediction from multisource heterogeneous data,” *Journal of Medical Internet Research*, vol. 25, e46812, 2023.

[14] S. Punarselvam *et al.*, “Enhancing seasonal influenza prediction through advanced time series machine learning models,” *Journal of Neonatal Surgery*, vol. 14, 2051, 2025.

[15] X. Zhang *et al.*, “Developing a machine learning prediction model for daily disease trends,” *China CDC Weekly*, 2025.