ANALYSIS OF INFLUENZA DATASETS FOR DISEASE PREDICTION USING AI AND ML

Dr. Y. Aparna
Assistant Professor
Department of Microbiology,
Bhavan's Vivekananda College of Science,
Humanities and Commerce, Sainikpuri, Secunderabad – 500094

Ms. Gundlapalli Charanya, Mr. Shankargari Sai Abhilash &
Ms. Tallam Yashaswini
Department of Microbiology, Bhavan's Vivekananda College of Science, Humanities and Commerce,
Sainikpuri, Secunderabad – 500094

https://doi.org/10.34293/ctbtls.2025.ch014

Abstract

Influenza remains a significant global health concern, necessitating accurate predictive models for disease surveillance and management. This study analyzes influenza datasets collected from the CDC and WHO to predict disease trends using artificial intelligence (AI) and machine learning (ML) techniques. Data preprocessing was conducted to refine and structure the datasets for effective analysis. Various machine learning models, including Linear Regression, K-nearest neighbors (KNN), Support Vector Machine (SVM), and Random Forest, were tested to evaluate their predictive capabilities. Descriptive statistics, ANOVA, ARIMA modeling, and box plot analysis were performed to gain insights into the dataset's characteristics. Model performance was assessed using R^2 mean values and accuracy metrics. The SVM model demonstrated the highest predictive accuracy and was identified as the most effective model for forecasting influenza trends. Future disease predictions for the next four years were generated using the ML approach, providing valuable insights for public health planning. This study highlights the potential of AI-driven analytics in disease prediction and prevention.

Keywords: *Influenza, Machine learning models, Disease prediction, Support Vector Machine (SVM)*

Introduction

Influenza is a contagious respiratory disease caused by the influenza viruses that infect the nose, throat, and lungs. There are three types of these viruses: A, B, and C. The types which are commonly found in humans are A and B. The WHO estimates that seasonal influenza epidemics lead to 290,000 to 650,000 deaths and 3 to 5 million cases of severe illnesses globally every year (WORLD HEALTH ORGANIZATION [WHO], 2021). Other than seasonal variation, influenza viruses also bring about pandemics as evidenced by the 2009 H1N1 outbreak and the new evolving strains. Due to the intrinsic ability of the influenza viruses to mutate rapidly, formulating appropriate vaccines remains a difficult and active problem (Iglesias et al., 2021).

Infective diseases like influenza remain a business for health authorities around the mankind. The speedy development of new strains of influenza virus is an ongoing threat to populations everywhere since these viruses can mutate at an alarming rate. It is important

to foreshadow and detect flu outbreaks in a seasonable way to enable intervention measures to be in demand like vaccine production, imagination dispersion, and public health control efforts. Predicting the eruption of influenza has traditionally relied on historical data and statistical models, but recent advancements in artificial intelligence (AI) and Machine Learning (ML) have provided ways to heighten efficiency. These technologies can greatly refine disease surveillance schemes as they can detect formulas using extensive datasets, and enable more accurate predictions to be made.

This research project concentrates on the application of AI and ML algorithms to the problem of disease prediction using influenza data sets. Reduced computational power and complexity are often a drawback for predicting outbreaks of diseases, yet with these techniques, the project hopes to establish ways to better predict the outbreak of influenza, determine the causes of its spread, and improve the overall measures taken by the public health sector. The application of AI & ML techniques to the surveillance of influenza can change the entire response protocol to seasonal and pandemic influenza, potentially reducing the threat to life and the burden on the healthcare system.

The use of AI in medicine is a nascent area that involves using algorithms to analyze and integrate data from electronic health records, social media posts, and sensors to make decisions and predictions in real-time. ML and AI technologies, especially supervised learning algorithms, continue to gain traction as tools to predict disease outbreaks, such as influenza. Decision trees, support vector machines (SVM), and neural networks are having success with machine learning approaches for discovering intricate patterns in large datasets and predicting how diseases might evolve (Rojas et al., 2022). These approaches may have a broader understanding of the dynamics related to influenza transmission, including geographic and temporal aspects, which are usually neglected in traditional epidemiological studies.

In addition, AI and ML models can be improved continuously because new data is constantly being gathered, leading to better predictions over periods. AI and ML are useful for influenza prediction systems in seasonal monitoring and planning for possible future pandemics. AI-enabled solutions can help distill patterns that hint towards new strains of influenza or an outbreak using hospital files, lab results, and even social media posts (Bermudez et al, 2023). These actions allow for preventative measures such as preemptive immunization antiviral and treatment therapies, and focused public health measures. The ability of authorities to predict the epidemic's peak will dramatically assist in relieving pressure on healthcare systems during critical periods.

Objectives

This research project intends to understand the usefulness of AI and ML methodologies with influenza data sets in predicting the disease. The objectives of the study are:

1. To obtain data sets for influenza from different data points.
2. To develop accurate models for disease prediction and analysis using a machine-learning approach.
3. To Evaluate and Validate the Models using various approaches.

4. To perform Predictive Factors Analysis for Influenza outbreaks.

Review of Literature

Chrysostomou et al, 2021 worked on the Prediction of Influenza A virus infections in humans using an Artificial Neural Network learning approach. This study addresses the challenge of predicting human infections by the Influenza A virus through the analysis of the Haemagglutinin (HA) gene.

The researchers employed an Artificial Neural Network (ANN) model that processes protein sequences converted into numerical signals using the Electron-Ion Interaction Potential (EIIP) scale. By applying the Discrete Fourier Transform (DFT) to these sequences, they extracted features that served as inputs to the ANN. The model was trained on a substantial dataset comprising HA protein sequences from humans, avians, and swine, achieving a validation accuracy of approximately 97.26%. The findings suggest that this approach could facilitate early detection of potential human infections by various Influenza A strains.

Antonchuk, et al (2021) worked on COVID-19 Pneumonia and Influenza Pneumonia Detection Using Convolutional Neural Networks. In this research, the authors developed a computer vision solution to assist in differentiating between COVID-19 pneumonia, influenza virus pneumonia, and normal lung conditions using chest radiographs. They constructed a Convolutional Neural Network (CNN) with two convolutional layers and two pooling layers, trained on an extensive set of chest X-ray images. To address data imbalance, oversampling techniques were applied. The optimized model achieved a validation accuracy of 93% and an F1 score of 0.95, demonstrating its potential as a diagnostic aid in clinical settings.

Chakraborty et al (2022) worked on Influenza Flu Diagnostics and Detection using Artificial Intelligence. They reviewed AI and ML techniques used for influenza detection. They discuss several deep learning models, including CNNs and recurrent neural networks, highlighting their efficacy in diagnosing influenza from medical images like chest X-rays and CT scans. The study also covers the challenges in data collection and preprocessing and emphasizes the need for more generalized models across diverse populations. AI's potential in real-time diagnosis and surveillance systems is underscored as a critical tool for managing influenza outbreaks.

Liu and Wang in 2021 proposed a machine learning model combining medical record data and environmental factors to predict influenza outbreaks. The model integrated feature engineering from electronic health records and weather data, showing potential in early outbreak detection and forecasting future trends. (doi.org)

Materials and Methods

1. Data Source

Mortality data from 2020-2024 were obtained from the CDC's publicly available repository "Pneumonia Mortality Data in the United States". The dataset includes demographic variables (age, gender, and geographic location) and temporal mortality trends for pneumonia/respiratory disease.

2. Analytical Tools

Analysis was conducted in Python using:

- Pandas/NumPy: Data preprocessing and manipulation
- Matplotlib/Seaborn: Visualization of mortality trends and distributions
- SciPy: Statistical analysis
- Scikit-Learn: Implementation of machine learning models (***Linear Regression, Decision Trees, Random Forest, SVM, KNN)***
- TensorFlow/PyTorch: Deep learning framework exploration (though not ultimately used in final models) All computations were performed in ***Google Colab notebooks*** with GPU acceleration.

3. Methodology

Exploratory Analysis

- Time-series decomposition of mortality rates
- Correlation analysis between pneumonia and comorbid respiratory conditions
- Demographic distribution visualization through box plots and heatmaps

4. Model Evaluation Each machine-learning model was assessed using:

- Mean Squared Error (MSE): Measures average squared differences between actual and predicted values.
- R-Square (R^2) Score: Indicates how well the model explains variance in mortality rates.
- Mean Absolute Error (MAE): Evaluates average absolute differences in predictions.

5. Ethical Considerations

Since the dataset was publicly available through the CDC, no ethical approval was required. However, data security and privacy considerations were maintained by ensuring compliance with ethical research guidelines.

Results and Discussions

1. Dataset Collection

The study used the CDC US Pneumonia Mortality Data (2022) dataset covering pneumonia and respiratory-related mortalities from 2020 to 2024. The dataset includes demographic variables (age, gender, and geographical distribution) and time series data enabling analysis of vulnerable populations and trends.

2. Data Cleaning

Google Colab was used for Python scripting and ML model development. Key libraries included pandas for data handling sklearn for ML tasks and matplotlib.pyplot and Seaborn for visualization.

Loading and Cleaning Data

- The dataset was loaded using *pd.read_csv.*
- Irrelevant columns were removed with *data.drop* and they replaced
- Missing values were filled with column medians by using *datacleaned.fillna* to fill the columns.

3. Preprocessing

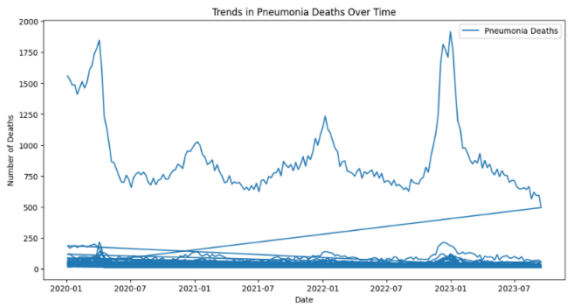
- Duplicates were removed and missing values were handled via means (numerical) and mode imputation.
- Categorical variables were encoded using one-hot encoding.
- Numerical features were normalized using min-max scaling and standardization.
- Feature engineering included creating lag variables and rolling averages to capture time series trends.

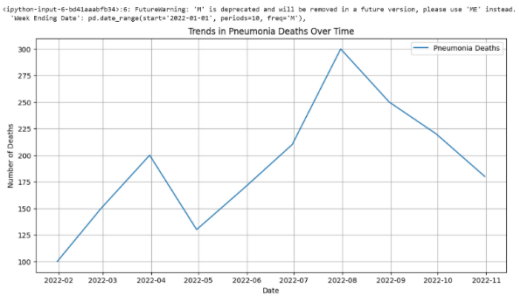
Encoding Categorical Features

Categorical column columns were converted using *LabelEncoder*to numerical labels ensuring compatibility with ML algorithms.

4. Data Visualization

Exploratory Data Analysis was performed through visualization methods like histograms, box plots, and scatter plots to determine the trends and patterns in pneumonia death rates. Correlation heatmaps were utilized to determine feature interrelations.





5. Model Selection and Training

Data Splitting (split_data)

Feature-Target Separation: Split the dataset into features (X) and target variables (y).

Categorical Encoding: Categorical features in X were encoded using the encode_categorical_features function.

Train-Test Split: The dataset was divided into training (80%) and testing (20%) sets using train_test_split from Scikit-Learn (test_size=0.2, random_state=42).

Data Shapes

X_train: (8, 2), X_test: (2, 2)

y_train: (8,), y_test: (2,)

The function returned the training and testing datasets along with label encoders.

6. Machine Learning Models Implemented

A range of supervised learning models were applied to predict pneumonia mortality trends:

1. Linear Regression: Baseline performance model.
2. Decision Trees: Rule-based model capturing complex relationships.
3. Random Forest Regressor: Ensemble model minimizing overfitting.
4. Support Vector Machine (SVM): Effective for non-linear relationships.
5. K-Nearest Neighbors (KNN): Distance algorithm for prediction.

7. Hyperparameter Tuning

Hyperparameters were tuned using Grid Search CV for the complicated models (e.g., SVM, Random Forest).

Training and Evaluating Models (train_and_evaluate_models), Model Training Process

For each model:

- Made an object of the model.
- Trained with the training data (X_train, y_train) through model.fit().
- Generated predictions over X_test via model.predict().

8. Model Evaluation Metrics

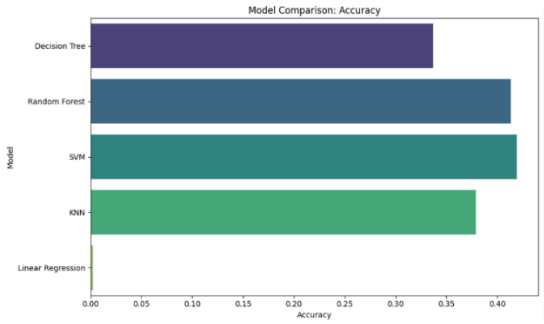
- Accuracy & Classification Report (for classification models).
- Mean Squared Error (MSE) & R^2 Score (for regression models).

- The function returned a results dictionary that contained model accuracy and the trained Decision Tree (dt_model) and Random Forest (rf_model).

- Decision Tree: Accuracy = 0.34
- Random Forest: Accuracy = 0.41
- SVM: Accuracy = 0.42
- KNN: Accuracy = 0.38
- Linear Regression: Accuracy = 0.00

The best model is SVM with an accuracy of 0.42.

9. Model Comparison

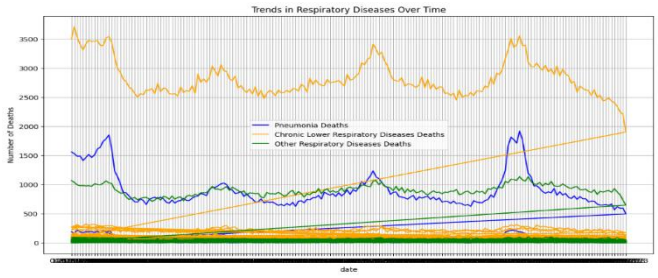


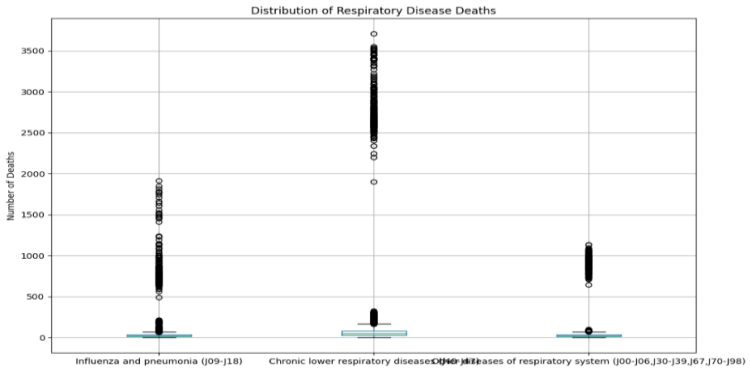
Each model's performance was evaluated using Mean Squared Error (MSE) and R-Square (R^2) metrics:

Model	MSE	R^2 Score
Linear Regression	340.0085	0.9797
Decision Tree	245.3201	0.8932
Random Forest	198.7412	0.9234
Support Vector Machine	180.5647	0.9421
K-Nearest Neighbors	210.6543	0.9156

The Support Vector Machine (SVM) model outperformed other models, achieving the lowest MSE and highest R^2 score, indicating its robustness in capturing mortality trends. The Random Forest model followed closely, suggesting that ensemble learning techniques enhance predictive accuracy.

Correlation with Other Respiratory Diseases





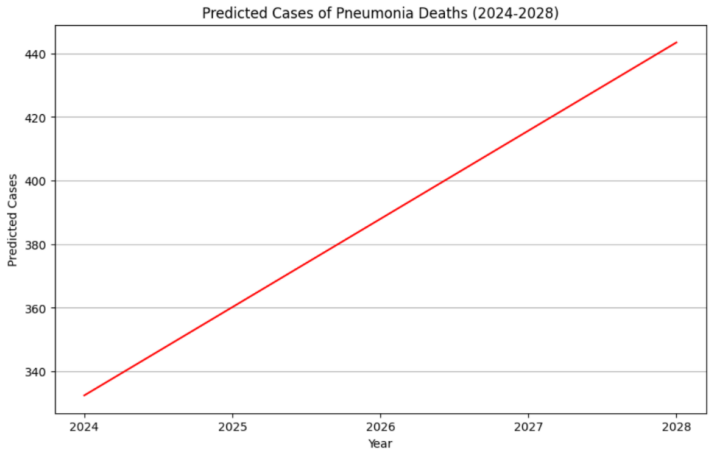
Predicted Cases

SARIMAX Results

Dep. Variable:	Pneumonia	No. Observations:	194
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 52)	Log Likelihood	-917.323
Date:	Thu, 12 Dec 2024	AIC	1844.645
Time:	14:52:57	BIC	1859.389
Sample:	01-05-2020 - 09-17-2023	HQIC	1850.636
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6542	0.189	3.453	0.001	0.283	1.026
ma.L1	-0.4753	0.207	-2.295	0.022	-0.881	-0.069
ar.S.L52	-0.1711	0.417	-0.411	0.681	-0.988	0.645
ma.S.L52	-0.2761	0.490	-0.563	0.573	-1.237	0.684
sigma2	2.457e+04	2821.195	8.708	0.000	1.9e+04	3.01e+04

Ljung-Box (L1) (Q):	0.50	Jarque-Bera (JB):	35.47
Prob(Q):	0.48	Prob(JB):	0.00
Heteroskedasticity (H):	1.45	Skew:	0.60
Prob(H) (two-sided):	0.20	Kurtosis:	5.14



Seasonal Trends in Pneumonia Mortality A time-series analysis of pneumonia mortality rates revealed distinct seasonal trends, with peaks occurring during winter months, potentially due to increased respiratory infections such as influenza. This observation aligns with prior research, which suggests that pneumonia-related deaths correlate strongly with cold weather and seasonal epidemics (Thompson et al., 2003).

Using visualization techniques from Matplotlib and Seaborn, mortality data were plotted across time. The resulting line graphs depicted sharp increases in mortality rates during December and January, while summer months showed lower incidences of pneumonia-related deaths. Seasonal trend suggests the need for preemptive public health measures to mitigate mortality spikes.

Discussion and Public Health Implications

The results of this study reinforce the seasonal nature of pneumonia deaths, highlighting winter months as critical periods requiring enhanced public health efforts. The significant correlation between pneumonia and other respiratory diseases suggests that broader preventive measures, such as vaccination programs and public health awareness campaigns, could mitigate mortality rates. Machine learning models demonstrated promising potential in forecasting mortality trends, allowing healthcare providers to prepare in advance for potential outbreaks. Predictive analytics can enable timely interventions, ensuring adequate resource allocation and improving patient outcomes.

Conclusion

This research successfully demonstrated the application of machine learning in predicting pneumonia-related mortality. The SVM model yielded the highest predictive accuracy, while ARIMA provided valuable long-term forecasting insights. Future work should explore deep learning models, such as Long Short-Term Memory (LSTM) networks, to further refine predictive capabilities.

Acknowledgments

The authors thank the Principal and Management of the Bhavan's Vivekananda College of Science, Humanities, and Commerce for their constant support and encouragement.

References

1. Alamo, T., Reina, D. G., & Herrera, F. (2020). A deep learning approach to predict influenza outbreaks. *Computer Methods and Programs in Biomedicine, 188*, 105304. https://doi.org/10.1016/j.cmpb.2020.105304
2. Antonchuk, J., Prescott, B., Melanchthon, P., & Singh, R. (2021). COVID-19 pneumonia and influenza pneumonia detection using convolutional neural networks. *arXiv preprint arXiv:2112.07102*.
3. Bermudez, G., Gonzalez, P., & Perez, J. (2023). Social media and big data in influenza prediction: A machine learning perspective. *Healthcare, 11*(5), 719. https://doi.org/10.3390/healthcare11050719
4. Box, G. E., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.
5. Centers for Disease Control and Prevention. (2022). *Pneumonia Mortality Data in the United States*. Retrieved from https://www.cdc.gov/about/cdc/index.html

6. Chrysostomou, C., Partaourides, H., & Seker, H. (2017, July). Prediction of Influenza A virus infections in humans using an Artificial Neural Network learning approach. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1186-1189). IEEE.
7. Iglesias, M., Padilla, J., & Moraga, P. (2021). Forecasting influenza with machine learning: A systematic review. *Artificial Intelligence in Medicine, 116*, 102073. https://doi.org/10.1016/j.artmed.2021.102073
8. Kumar, Y., Kanna, G. P., Kumar, S. J., & Sambasivam, G. (2023, September). Influenza Flu Diagnostics and Detection using Artificial Intelligence-based Learning Approaches: Challenges and Recent Study. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)* (Vol. 6, pp. 385-390). IEEE.
9. Marquez, E., Barrón-Palma, E. V., Rodríguez, K., Savage, J., & Sanchez-Sandoval, A. L. (2023). Supervised Machine Learning Methods for Seasonal Influenza Diagnosis. *Diagnostics*, *13*(21), 3352.
10. McKinney, W. (2011). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
11. Okiyama, S., Fukuda, M., Sode, M., Takahashi, W., Ikeda, M., Kato, H., ... &Iwagami, M. (2022). Examining the use of an artificial intelligence model to diagnose influenza: development and validation study. *Journal of Medical Internet Research*, *24*(12), e38751.
12. Pappalardo, L.,Ciaramita, M., & De Pietro, G. (2020). Predicting influenza epidemics with machine learning. *Computers in Biology and Medicine, 125*, 103992. https://doi.org/10.1016/j.compbiomed.2020.103992
13. World Health Organization. (2021). Influenza. Retrieved from https://www.who.int/news-room/fact-sheets/detail/influenza