Factors In A Country That Could Significantly Increase The Probability Of Respiratory Disease Outbreak

Andreas Zurhaar
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—This thesis investigates predictive modeling of flu outbreaks using machine learning techniques, including Linear Regression, Gradient Boosting, Random Forest, Feedforward Neural Networks, Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. The primary aim is to identify key preconditions for predicting flu outbreaks and assess model effectiveness.**

Data preprocessing ensures feature robustness by calculating and imputing average changes over multiple years. Feature importance analyses highlight Mean Income Per Day, Population, Public Healthcare Expenditure, and Positive Flu Tests as critical factors.

RNN and LSTM models, evaluated through comparisons of predicted and actual flu cases, demonstrate suitability for time series predictions, though RMSE variability suggests the need for careful evaluation and potential ensemble integration.

Ensemble models, particularly Gradient Boosting and Random Forests, show superior predictive performance. Gradient Boosting exhibits a consistent learning curve, while Random Forest achieves a high R-squared value and low mean squared error, validating ensemble methods.

Accurate flu predictions enhance public health preparedness and response, underscoring the importance of socio-economic and health-related features in outbreak prediction. This research validates RNNs for temporal forecasting and confirms ensemble models' accuracy improvement, informing public health strategies and future research.

I. Introduction

Respiratory diseases pose a significant global health challenge, affecting millions and heavily burdening healthcare systems worldwide. The urgency of understanding and mitigating factors contributing to these outbreaks has been underscored by recent pandemics like COVID-19. This thesis investigates country-specific preconditions that may contribute to respiratory disease outbreaks, utilizing advanced machine learning models to predict their likelihood.

Existing literature primarily explores conditions during ongoing outbreaks, focusing on real-time data analysis and immediate response strategies. For instance, [2] highlighted the importance of examining environmental and social conditions during outbreaks for effective management. Additionally, studies such as "Under the Weather" (2001) have explored the role of climate in common cold outbreaks, yet they lack a comprehensive scope of preconditions.

Recent years have seen increased attention on the application of machine learning in disease prediction. Techniques like recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks have shown promise in time-series analysis and prediction. Despite these advancements, there is a notable gap in the literature regarding the use of ensemble models for predicting respiratory disease outbreaks and comparing the effectiveness of different predictive models.

This research sits at the intersection of epidemiology and machine learning, leveraging state-of-the-art predictive analytics to address a critical public health issue. The study employs a systematic approach to identify multiple preconditions leading to respiratory disease outbreaks and uses advanced models, including RNNs, LSTMs, and ensemble methods like Random Forests. By integrating data from diverse sources across multiple countries, this research aims to develop robust and reliable predictive models.

The primary problem this thesis addresses is whether identifiable conditions within a country increase the chances of a respiratory disease outbreak. Understanding these conditions is essential for developing effective predictive models and preventive strategies. To tackle this issue, the research poses the following questions:

- Which preconditions are most relevant for predicting a possible outbreak?
- Are recurrent neural networks suited for predicting respiratory disease outbreaks?
- Can ensemble models increase prediction accuracy compared to neural networks?

This thesis aims to bridge the gap in proactive prediction and prevention measures by forecasting the likelihood of future outbreaks based on various preconditions, ultimately contributing to better public health preparedness and response.

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor(s): [Linda Rieswijk, Rachel Cavill]

II. Literature Review

This literature review examines existing research on factors contributing to respiratory disease outbreaks, the application of machine learning in disease prediction, and the comparative effectiveness of different predictive models.

A. Identifiable Conditions Influencing Respiratory Disease Outbreaks

Respiratory disease outbreaks are influenced by environmental, socio-economic, and healthcare-related conditions.

1) Environmental Factors: Air pollution, including PM2.5, NO2, and SO2, significantly contributes to respiratory illnesses [7]. Climate change also affects virus transmission rates, with cold and dry conditions linked to increased influenza survival and transmission [5].

2) Socio-Economic Factors: High population density facilitates pathogen transmission. Lower-income populations often have limited healthcare access and live in areas with poorer air quality [8]. Socio-economic disparities also influence vaccination rates and healthcare-seeking behavior [7].

3) Healthcare-Related Factors: The robustness of healthcare infrastructure affects outbreak management. Well-established healthcare systems ensure timely diagnosis and effective treatment, whereas inadequate facilities lead to delayed responses and higher transmission rates [5].

B. Machine Learning in Disease Prediction

Machine learning techniques, especially for time-series analysis, show promise in predicting disease outbreaks. RNNs and LSTMs are effective in handling sequential data and capturing temporal dependencies [6].

1) Recurrent Neural Networks (RNNs): RNNs are suitable for predicting disease outbreaks based on historical data, analyzing patterns in disease incidence and environmental conditions [9].

2) Long Short-Term Memory (LSTM) Networks: LSTMs, an extension of RNNs, address the vanishing gradient problem and capture long-term dependencies, making them effective for long-term outbreak predictions [5].

C. Ensemble Models in Predictive Analytics

Ensemble models, combining multiple algorithms, improve prediction accuracy. Methods like Random Forests and Gradient Boosting Machines have shown enhanced performance in various predictive tasks [6].

1) Random Forests: Random Forests, which aggregate predictions from multiple decision trees, are robust and provide high accuracy in classification tasks. They have been used to predict disease outbreaks by analyzing diverse datasets [8].

2) Gradient Boosting Machines and AdaBoost: These boosting algorithms iteratively improve accuracy by focusing on misclassified instances, useful for handling imbalanced datasets common in disease outbreak data (academic.oup.com, 2023).

D. Comparative Analysis

Studies comparing RNNs, LSTMs, and ensemble models indicate that while RNNs and LSTMs excel in capturing temporal patterns, ensemble models often outperform in overall prediction accuracy due to leveraging multiple algorithms' strengths [5]. Metrics like accuracy, precision, recall, and F1 score show ensemble models typically providing superior results in outbreak prediction [8].

E. Conclusion

The literature highlights the multifaceted nature of respiratory disease outbreaks, influenced by environmental, socio-economic, and healthcare-related factors. Machine learning models, especially RNNs, LSTMs, and ensemble methods, offer promising avenues for predicting these outbreaks. Further research is needed to refine these models and validate their effectiveness in diverse settings. This thesis aims to contribute by identifying key preconditions for respiratory disease outbreaks and evaluating the predictive performance of various machine learning models.

III. Methods

This study employs a mixed-methods approach, integrating both quantitative and qualitative data to identify preconditions for respiratory disease outbreaks and to develop predictive models. The research is structured into several key phases: data acquisition, feature engineering, model training, evaluation, and ethical considerations.

A. Data Acquisition

Eight countries were selected for their diverse and representative sample: the United States, China, Sweden, the United Kingdom, Australia, Brazil, Russia, and Japan. These countries were chosen for their accessible relevant data and variance in environmental, socio-economic, and healthcare conditions.

1) Data Sources: Data were collected from multiple reputable sources:

- World Health Organization (WHO): Influenza case frequency data.
- Our World in Data: Global data on health and socio-economic factors.
- Microsoft Power BI: Environmental and climate data.
- National Health Databases: Country-specific health and disease incidence data.

2) Variables and Measurements: The study focuses on a range of variables influencing respiratory disease outbreaks:

- Environmental Factors: Air quality (PM2.5, PM10, NO2, SO2), temperature, humidity.
- Socio-Economic Factors: Population density, income levels, education, healthcare access.
- Healthcare Infrastructure: Number of hospitals, healthcare expenditure, vaccination rates.
- Travel Patterns: International travel frequency, urbanization rates.

B. Feature Engineering

Feature engineering was performed to preprocess and transform the raw data into a suitable format for machine learning models:

- Data Cleaning: Removal of duplicates, handling missing values, normalization of data.
- Temporal Features: Creation of time-series data for variables such as air quality and temperature.
- Interaction Terms: Development of interaction terms to capture the combined effects of variables, such as population density and air pollution levels.

C. Model Training

1) Machine Learning Techniques: Several machine learning techniques suitable for time-series analysis and prediction were employed:

- Data Splitting: The dataset was split into training (70%), validation (15%), and test (15%) sets to ensure robust model evaluation.
- Hyperparameter Tuning: Grid search and cross-validation were used to optimize the hyperparameters of the machine learning models.
- Training: Models were trained on historical data, capturing the relationships between identified preconditions and respiratory disease outbreaks.

D. Model Evaluation

1) Evaluation Metrics: Several metrics were used to assess the performance of the predictive models:

- Accuracy: The proportion of correct predictions made by the model.
- Precision: The proportion of true positive predictions among the predicted positives.
- Recall: The proportion of true positive predictions among the actual positives.
- F1 Score: The harmonic mean of precision and recall.
- Root Mean Squared Error (RMSE): Measures the differences between predicted and actual values.

2) Validation and Testing: Models were evaluated on the validation set during hyperparameter tuning and tested on the test set to assess generalization capabilities. The performance of RNNs, LSTMs, and ensemble models was compared to identify the most accurate predictive model.

E. Data Visualization

To facilitate interpretation and communication of findings, data visualizations were developed using tools such as Tableau and Matplotlib. These visualizations include:

- Temporal Trends: Graphs showing trends of key variables over time.
- Prediction Plots: Comparing predicted and actual values of respiratory disease incidences.

F. Ethical Considerations

The study adheres to ethical standards concerning data privacy and security. Personal identifying information was anonymized, and data were handled in compliance with privacy regulations such as the General Data Protection Regulation (GDPR). The ethical implications of using predictive analytics tools were considered, emphasizing the importance of enhancing public health outcomes without infringing on individual privacy rights.

IV. Experiments

A. Objective

The primary objective of these experiments is to identify the most critical features for predicting respiratory disease outbreaks. By comparing the performance of various machine learning models when key features are excluded, we aim to determine which features consistently show importance across all models. This analysis is crucial for enhancing the accuracy of predictive models and improving public health interventions.

B. Experimental Design

To achieve this objective, we have designed a series of experiments involving multiple machine learning models and systematically excluding each feature to observe the impact on model performance. The models used in these experiments include Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), Linear Regression, Random Forest, Neural Networks (NN), and Gradient Boosting Regressor (GBR).

1) Data Preparation: Data from yearly records (2009-2019) were read and concatenated into a single comprehensive DataFrame. Missing values were handled by calculating average changes across years and using these averages to fill gaps, maintaining dataset consistency.

2) Normalization: All features were normalized using MinMaxScaler, transforming data to a range between 0 and 1 for effective model training. The 'Country' feature was excluded as it did not significantly contribute to prediction and introduced noise.

3) Feature Importance Analysis: To identify important features for predicting respiratory disease outbreaks, each feature was sequentially excluded from the dataset. Models were trained without the excluded feature, and the impact was recorded by evaluating the Root Mean Squared Error (RMSE) of predictions.

4) Models Used:

a) Recurrent Neural Network (RNN): RNNs handle sequential data by maintaining a 'memory' of previous inputs. Two SimpleRNN layers (50 units each) were used, followed by a Dense layer for final prediction. The model used mean squared error loss and Adam optimizer.

b) Long Short-Term Memory (LSTM) Network: LSTMs capture longer-term dependencies in data. The model used two LSTM layers (50 units each), followed by a Dense layer. Compiled with mean squared error loss and Adam optimizer.

c) Linear Regression: Linear Regression assumes a linear relationship between features and the target variable. Implemented using scikit-learn, it serves as a baseline for performance comparison.

d) Random Forest: Random Forest constructs multiple decision trees during training and outputs the mean prediction. Implemented using scikit-learn with 100 estimators for robust predictions.

e) Neural Network (NN): A feedforward neural network with two hidden layers (128 and 64 neurons, ReLU activation) was used to capture complex relationships between features and target variable.

f) Gradient Boosting Regressor (GBR): Gradient Boosting builds models sequentially, each correcting errors of the previous ones. Implemented with 100 estimators to incrementally improve prediction accuracy.

5) Evaluation Metrics: Root Mean Squared Error (RMSE) measures the standard deviation of residuals and is sensitive to large errors, making it suitable for assessing the impact of feature exclusion on model accuracy.

6) Comparison and Analysis: Each model was evaluated by calculating the RMSE on training and test datasets. RMSE values were recorded for each feature exclusion scenario, determining feature importance based on model performance deterioration.

7) Results Visualization: Results were visualized using tools such as Tableau and Matplotlib, including temporal trends, and prediction plots to facilitate interpretation and communication of findings.

V. Results

By comparing the performance of various models, we identified the importance of individual features in predicting respiratory disease outbreaks. The analysis provided insights into which features are consistently important across different models, aiding in more accurate and reliable predictions.

A. Linear Regression

OLS Regression Results

Dep. Variable:	y	R-squared:	0.686
Model:	OLS	Adj. R-squared:	0.636
Method:	Least Squares	F-statistic:	13.66
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	1.97e-14
Time:	16:52:12	Log-Likelihood:	-73.885
No. Observations:	88	AIC:	173.8
Df Residuals:	75	BIC:	206.0
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.388e-17	0.065	2.15e-16	1.000	-0.129	0.129
x1	1.3940	0.927	1.504	0.137	-0.452	3.240
x2	0.0295	0.195	0.151	0.880	-0.359	0.418
x3	-0.0699	0.079	-0.890	0.376	-0.226	0.087
x4	0.1865	0.361	0.517	0.607	-0.533	0.906
x5	0.7953	0.373	2.135	0.036	0.053	1.537
x6	-0.3219	0.167	-1.925	0.058	-0.655	0.011
x7	-0.3614	0.240	-1.505	0.137	-0.840	0.117
x8	-0.5595	0.261	-2.143	0.035	-1.080	-0.039
x9	-1.0869	0.834	-1.303	0.197	-2.749	0.575
x10	0.0135	0.071	0.190	0.850	-0.128	0.155
x11	-0.1461	0.103	-1.418	0.160	-0.351	0.059
x12	0.0157	0.107	0.147	0.884	-0.197	0.229

Omnibus:	22.948	Durbin-Watson:	1.888
Prob(Omnibus):	0.000	Jarque-Bera (JB):	169.931
Skew:	0.261	Prob(JB):	1.26e-37
Kurtosis:	9.788	Cond. No.	39.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 1. Ordinary Least Squares (OLS) for Linear Regression.

Figure 1 shows the OLS regression results, indicating the relationship between observed and predicted values. The table includes estimated coefficients, standard errors, t-values, and p-values to assess each predictor's significance.

Feature	VIF
Population (millions)	99.644381
GDP (Billions)	16.523791
Tourist Trips	1.808483
International Passengers (Millions)	18.958051
Mean Income Per Day (Dollars)	89.513476
Public Healthcare Expenditure (Percentage)	47.494307
Median Age	107.318413
Growth Rate	7.054193
Air Pollution	149.583804
Water Quality Deaths (Percentage)	1.224144
Positive Flu Tests Min	2.860578
Positive Flu Tests Max	11.397720

TABLE I
VARIANCE INFLATION FACTOR (VIF) FOR FEATURES

Table I lists each feature and its VIF value, indicating the presence of multicollinearity. High VIF values suggest high multicollinearity, affecting the stability and interpretation of model coefficients.

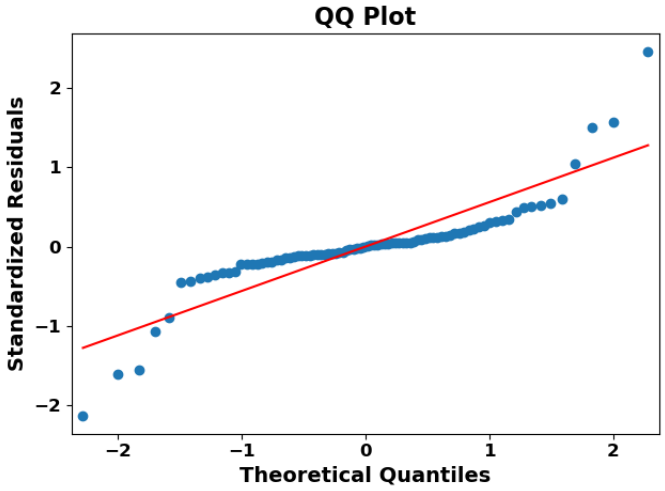

Fig. 2. Q-Q Plot of Residuals.

Figure 2 compares the quantiles of the residuals to a theoretical normal distribution, assessing residual normality. Deviations suggest departures from normality, impacting model assumptions.

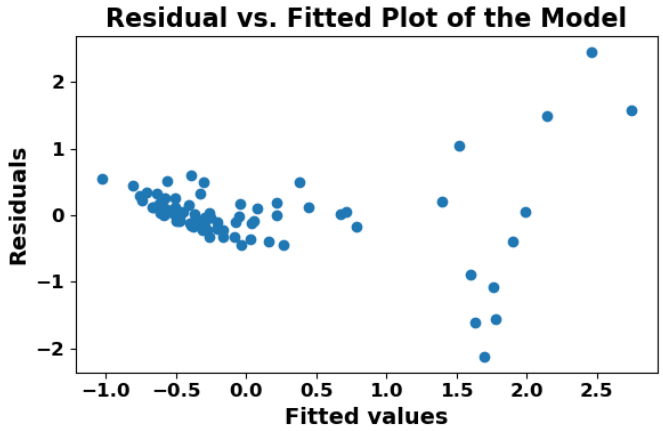

Fig. 3. Residuals vs. Fitted Values.

Figure 3 shows residuals against fitted values to assess model fit. Ideally, residuals should be randomly dispersed around zero, indicating a good fit. Patterns suggest issues like non-linearity or heteroscedasticity.

B. Gradient Boosting

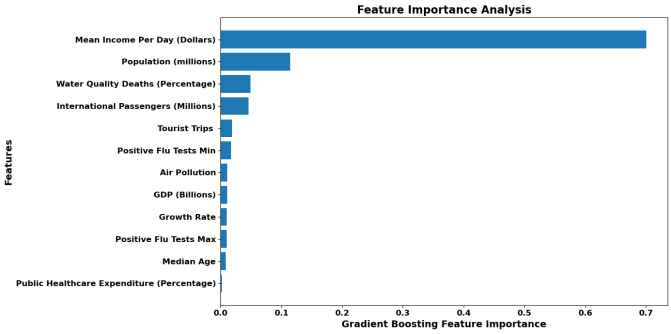

Fig. 4. Gradient Boosting Model Feature Importance.

Figure 4 displays the relative importance of each feature as determined by the gradient boosting model, highlighting the contribution of each feature to predictions.

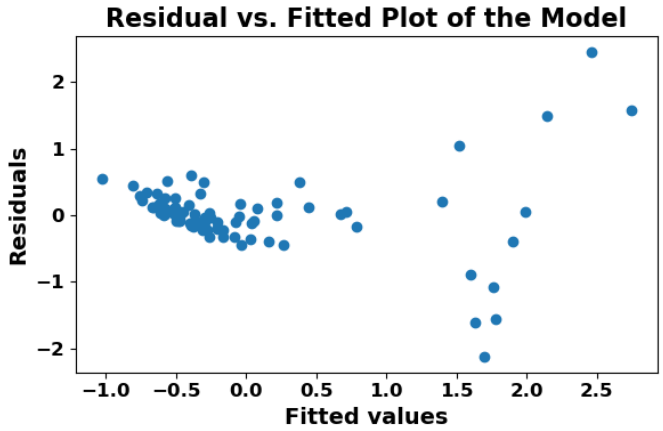

Fig. 5. Residuals vs. Fitted Values.

Figure 5 shows the residuals against fitted values for the gradient boosting model, assessing model fit. Random dispersion around zero indicates a good fit.

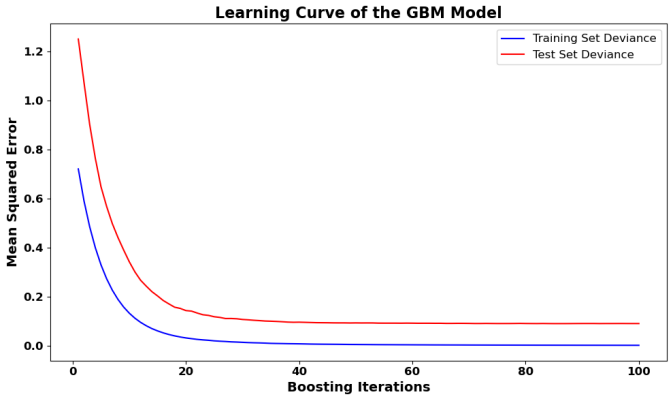

Fig. 6. Learning Curve of the Gradient Boosting Model.

Figure 6 illustrates the model's learning process by plotting training and test set deviance against boosting iterations, highlighting potential overfitting.

C. Random Forest

Metric	Value
Mean Squared Error	0.142889
R-squared	0.902413
Out-of-Bag Score	0.682524

TABLE II
RANDOM FOREST MODEL PERFORMANCE METRICS

Table II presents the performance metrics for the random forest model, including MSE, R-squared value, and OOB Score, evaluating the model's accuracy and generalization capability.

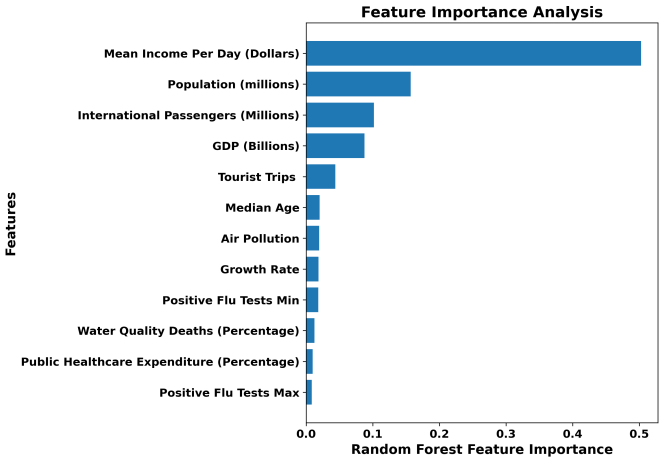

Fig. 7. Random Forest Model Feature Importance.

Figure 7 displays the feature importance determined by the random forest model, indicating each feature's contribution to predictions.

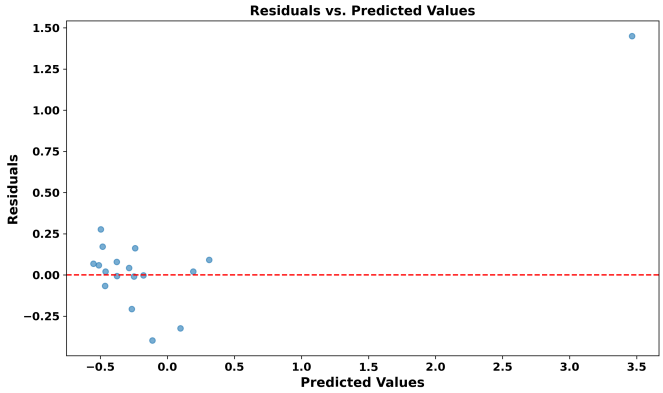

Fig. 8. Residuals vs. Fitted Values.

Figure 8 shows residuals against fitted values for the random forest model, assessing fit. Random dispersion around zero indicates a good fit.

D. Feed Forward Neural Network

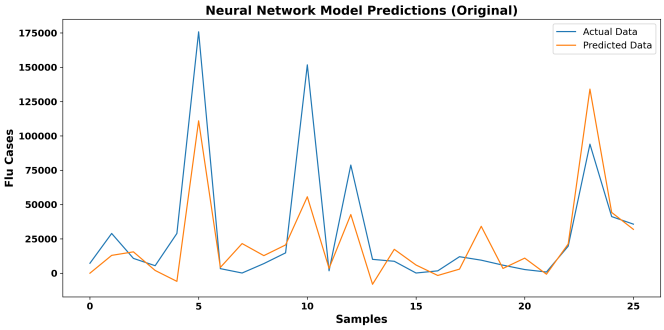

Fig. 9. Neural Network Model Predictions.

Figure 9 shows predicted values from the feed forward neural network compared to actual test data, visually assessing model performance on the original data scale.

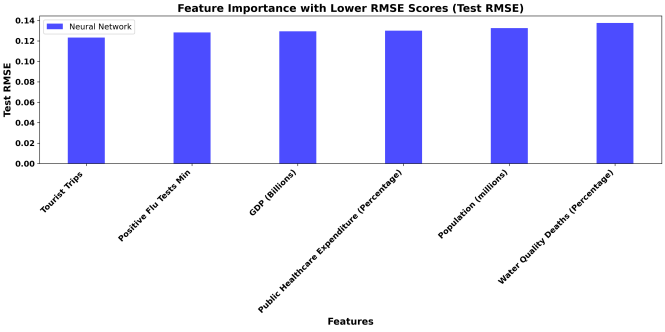

Fig. 10. Feature Importance with Lower RMSE Scores (Test RMSE).

Figure 10 displays features resulting in lower averaged RMSE scores over 100 iterations when excluded, indicating their critical contribution to model accuracy.

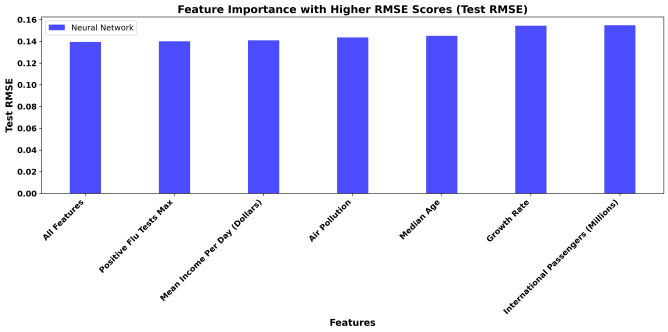

Fig. 11. Feature Importance with Higher RMSE Scores (Test RMSE).

Figure 11 displays features resulting in higher averaged RMSE scores over 100 iterations when excluded, indicating their critical contribution to model accuracy.

E. RNN

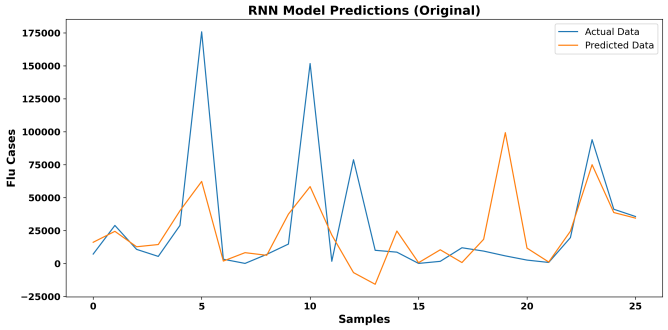

Fig. 12. RNN Model Predictions (Original).

Figure 12 shows predicted values from the RNN compared to actual test data, visually assessing model performance on the original data scale.

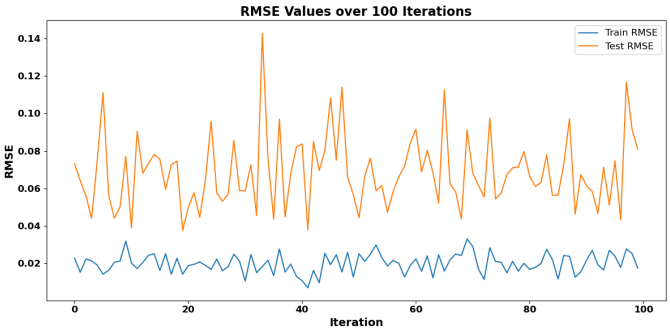

Fig. 13. RNN RMSE Over Iterations.

Figure 13 shows RMSE values of the RNN for both training and test sets across 100 iterations, illustrating performance stability and consistency.

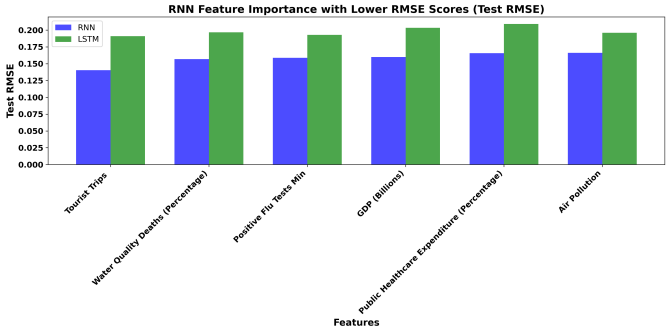

Fig. 14. Feature Importance with Lower RMSE Scores Over Iterations (Test RMSE).

Figure 14 displays features resulting in lower averaged RMSE scores over 20 iterations when excluded from the RNN, indicating their critical contribution to model accuracy.

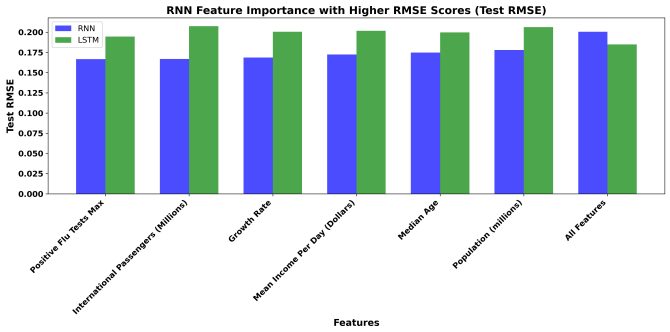

Fig. 15. Feature Importance with Higher RMSE Scores Over Iterations (Test RMSE).

Figure 15 displays features resulting in higher averaged RMSE scores over 20 iterations when excluded from the RNN, indicating their critical contribution to model accuracy.

Feature	RNN Test RMSE
Tourist Trips	0.1403
Water Quality Deaths (Percentage)	0.1569
Positive Flu Tests Min	0.1587
GDP (Billions)	0.1600
Public Healthcare Expenditure (Percentage)	0.1656
Air Pollution	0.1661
Positive Flu Tests Max	0.1668
International Passengers (Millions)	0.1670
Growth Rate	0.1687
Mean Income Per Day (Dollars)	0.1724
Median Age	0.1749
Population (millions)	0.1781
All Features	0.2006

TABLE III
FEATURE IMPORTANCE (TEST RMSE) FOR RNN MODEL

This table shows the Test RMSE values for each feature when omitted from the Recurrent Neural Network (RNN) model, as well as the RMSE for the model when using all features. Lower RMSE values indicate more critical features for the model's accuracy.

F. LSTM

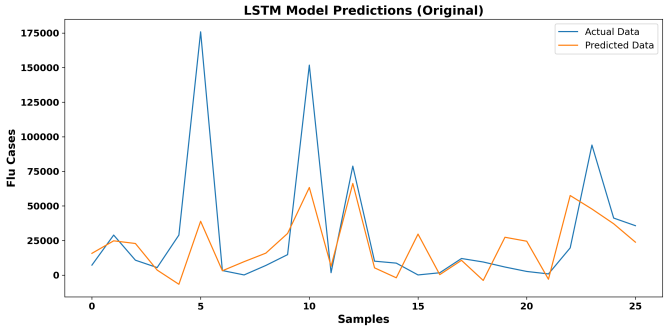

Fig. 16. LSTM Model Predictions (Original).

Figure 16 shows the predicted values from the LSTM model compared to the actual test data, providing a visual assessment of the model's performance on the original data scale.

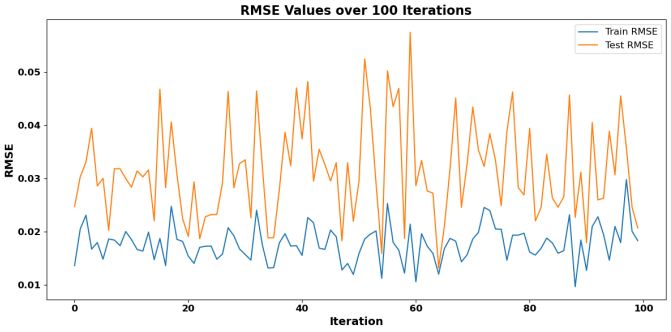

Fig. 17. LSTM RMSE Over Iterations.

Figure 17 shows the RMSE values of the LSTM for both training and test sets across 100 iterations, illustrating the model's performance stability and consistency.

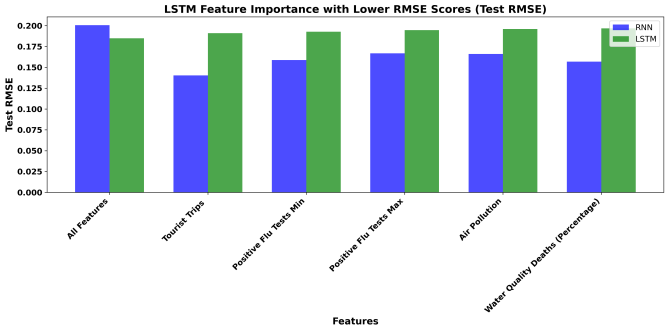

Fig. 18. Feature Importance with Lower RMSE Scores Over Iterations (Test RMSE).

Figure 18 displays features resulting in lower averaged RMSE scores over 20 iterations when excluded from the LSTM network, indicating their critical contribution to the model's accuracy.

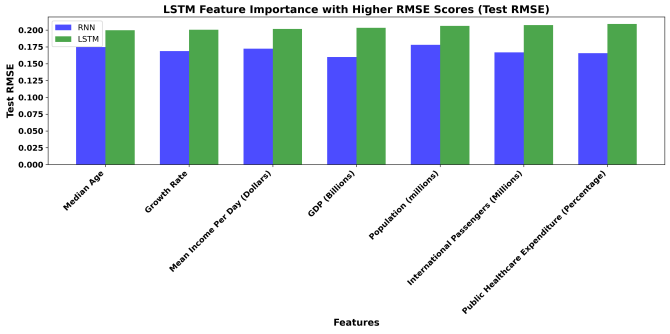

Fig. 19. Feature Importance with Higher RMSE Scores Over Iterations (Test RMSE).

Figure 19 displays features resulting in higher averaged RMSE scores over 20 iterations when excluded from the LSTM network, indicating their critical contribution to the model's accuracy.

Feature	LSTM Test RMSE
Tourist Trips	0.1911
Positive Flu Tests Min	0.1928
Positive Flu Tests Max	0.1947
Air Pollution	0.1960
Water Quality Deaths (Percentage)	0.1967
Median Age	0.1998
Growth Rate	0.2005
Mean Income Per Day (Dollars)	0.2017
GDP (Billions)	0.2034
Population (millions)	0.2062
International Passengers (Millions)	0.2074
Public Healthcare Expenditure (Percentage)	0.2092
All Features	0.1849

TABLE IV
FEATURE IMPORTANCE (TEST RMSE) FOR LSTM MODEL

This table shows the Test RMSE values for each feature when omitted from the Long Short-Term Memory (LSTM) model, as well as the RMSE for the model when using all features. Lower RMSE values indicate more critical features for the model's accuracy.

VI. DISCUSSION

A. Linear Regression

The OLS regression model explains 68.6% of the variability in the dependent variable ($R^2 = 0.686$), with an adjusted R^2 of 0.636, indicating a good model fit. The F-statistic (13.66, $p < 0.001$) confirms the model's statistical significance. Key predictors include Public Healthcare Expenditure (positive coefficient, $p = 0.036$) and Air Pollution (negative coefficient, $p = 0.035$), indicating their significant impact on the dependent variable.

Model diagnostics show non-normal residuals (Omnibus test statistic = 22.948, $p = 0.000$) and no significant autocorrelation (Durbin-Watson statistic = 1.888). The Jarque-Bera test confirms non-normality (statistic = 169.931, $p < 0.001$). The condition number of 39.9 suggests no severe multicollinearity.

The VIF analysis reveals significant multicollinearity, especially for Population (VIF = 99.644), Air Pollution (VIF = 149.584), and Median Age (VIF = 107.318). High VIF values indicate potential instability in coefficient estimates.

In summary, the OLS regression model demonstrates a substantial explanatory power, with Public Healthcare Expenditure and Air Pollution significantly impacting the dependent variable. However, non-normal residuals and multicollinearity issues suggest areas for model refinement. Addressing multicollinearity by removing or combining highly correlated predictors may improve the model's interpretability and stability.

B. Gradient Boosting

The gradient boosting model highlights Mean Income Per Day and Population as the most crucial features. The learning curve shows rapid initial learning with the training set deviance leveling off, indicating effective learning from the training data. The test set deviance also decreases, suggesting good generalization, though some overfitting is indicated by the gap between training and test set deviance.

The residuals vs. fitted values plot shows some clustering and outliers, suggesting potential areas for model improvement. These visualizations collectively provide a comprehensive understanding of the gradient boosting model's performance and the importance of various features in the dataset.

In summary, the gradient boosting model effectively identifies key predictors and generalizes well, although some overfitting is present. The feature importance analysis and learning curve indicate the model's strong performance, with Mean Income Per Day and Population being the most critical features.

C. Random Forest

The Random Forest model provides valuable insights into feature importance and model performance. The analysis reveals that Mean Income Per Day (Dollars) is the most crucial feature, significantly influencing the model's predictions. Population (millions) is the second most important feature, indicating its significant impact on the model's accuracy. Other contributing features include International Passengers (Millions), GDP (Billions), and Tourist Trips, though to a lesser extent. Features like Positive Flu Tests Max, Public Healthcare Expenditure (Percentage), and Water Quality Deaths (Percentage) have minimal importance in this model.

The residuals vs. predicted values plot diagnoses the goodness of fit, showing residuals (errors) plotted against the predicted values. Ideally, residuals should be randomly dispersed around zero, indicating a good fit. The plot indicates some clustering around zero with a few significant outliers, suggesting areas where the model does not fit well and potential improvements in model fitting or data preprocessing are needed.

The performance metrics for the Random Forest model include a Mean Squared Error (MSE) of 0.142889, indicating the average squared difference between observed and predicted outcomes, with lower MSE indicating better performance. The R-squared value of 0.902412 suggests that approximately 90.24% of the variability in the dependent variable is explained

by the model, indicating a strong fit. The Out-of-Bag (OOB) Score of 0.682524 reflects the model's accuracy on data not used during training, indicating good generalization performance.

In summary, the feature importance analysis shows that Mean Income Per Day (Dollars) and Population (millions) are the most critical features. The residuals plot reveals areas of poor fit, suggesting potential improvements. The MSE, R-squared value, and OOB score indicate strong model performance and good generalization capability.

D. Feed Forward Neural Network

The Feed Forward Neural Network (NN) model offers comprehensive insights into feature importance and model performance, as demonstrated by various graphs and analyses.

The "Neural Network Model Predictions (Original)" graph shows the comparison between actual flu cases and predicted flu cases using the original data. The model's predictions closely follow the trend of the actual data, effectively capturing the overall pattern of flu cases. However, deviations, particularly at peaks, suggest the model might struggle with extreme values.

The "Feature Importance with Higher RMSE Scores (Test RMSE)" graph displays features that resulted in higher RMSE scores when excluded, indicating their importance. Positive Flu Tests Max and Mean Income Per Day (Dollars) are among the top features, critical for accurate predictions. Other important features include Air Pollution, Median Age, Growth Rate, and International Passengers (Millions).

The "Feature Importance with Lower RMSE Scores (Test RMSE)" graph shows features that resulted in lower RMSE scores when excluded, indicating their lesser importance. Tourist Trips, Positive Flu Tests Min, GDP (Billions), Public Healthcare Expenditure (Percentage), Population (millions), and Water Quality Deaths (Percentage) have relatively lower importance.

The "Neural Network Model Predictions (Normalized)" graph compares actual and predicted flu cases using normalized data. The model's predictions closely follow the actual data trend, indicating effective pattern capture. Deviations, especially at peaks, suggest areas for model improvement.

The Feed Forward Neural Network model demonstrates strong predictive ability, indicated by the close alignment between actual and predicted values in both original and normalized data. The feature importance analysis reveals that Positive Flu Tests Max, Mean Income Per Day (Dollars), Air Pollution, Median Age, Growth Rate, and International Passengers (Millions) are critical for accurate predictions. In contrast, features like Tourist Trips, Positive Flu Tests Min, GDP (Billions), Public Healthcare Expenditure (Percentage), Population (millions), and Water Quality Deaths (Percentage) have a lesser impact. These visualizations provide a comprehensive understanding of the neural network model's performance and feature importance.

E. RNN

The Recurrent Neural Network (RNN) model was analyzed to understand its performance in predicting respiratory disease outbreaks. The results from various graphs and plots highlight the model's efficacy and areas for improvement.

The "RMSE Values over 100 Iterations" graph showcases the Root Mean Square Error (RMSE) of the RNN model over 100 iterations. The graph indicates variability in the test RMSE, suggesting inconsistent predictions across different iterations, pointing to potential overfitting or sensitivity to specific data subsets.

The "RNN Model Predictions (Original)" graph compares the actual flu cases with the predicted cases using non-standardized data. The model shows a reasonable fit, capturing some peaks and troughs, but underpredicts actual cases at the peaks.

Similarly, the "RNN Model Predictions (Normalized)" graph displays the comparison using normalized data. The normalized predictions are closer to the actual data but still show deviations, especially at higher values.

Feature importance analysis is depicted in two graphs: "RNN Feature Importance with Higher RMSE Scores (Test RMSE)" and "RNN Feature Importance with Lower RMSE Scores (Test RMSE)." Features with higher RMSE scores upon exclusion, such as "Positive Flu Tests Max," "International Passengers (Millions)," and "Mean Income Per Day (Dollars)," are critical for predictions. Conversely, features with lower RMSE scores, like "Tourist Trips" and "Water Quality Deaths (Percentage)," have less impact.

In conclusion, the RNN model shows potential in predicting respiratory disease outbreaks, with certain features being pivotal. However, variability in RMSE and occasional prediction inaccuracies suggest further tuning and possibly a hybrid approach with other models to enhance reliability and accuracy.

F. LSTM

The LSTM model predictions are depicted in two forms: standardized and original. The "LSTM Model Predictions (Normalized)" graph illustrates the model's performance on normalized flu case data, successfully capturing the actual data trend with some noticeable deviations. The "LSTM Model Predictions (Original)" graph shows predictions in their original scale, highlighting the model's ability to approximate actual flu case counts, despite some discrepancies.

The feature importance analysis for the LSTM model is divided into two categories: higher RMSE scores and lower RMSE scores. The graph "LSTM Feature Importance with Higher RMSE Scores (Test RMSE)" presents features that resulted in higher RMSE scores when excluded, indicating their significance for accurate prediction. Notable features include "Median Age," "Growth Rate," and "Mean Income Per Day (Dollars)." The graph "LSTM Feature Importance with Lower RMSE Scores (Test RMSE)" highlights features that led to lower RMSE scores when omitted, suggesting their lesser impact on predictive accuracy. Features like "Tourist

Trips," "Positive Flu Tests Min," and "Positive Flu Tests Max" are less critical.

The "RMSE Values over 100 Iterations" graph provides insight into the model's consistency. The training RMSE remains low and stable, indicating good performance on training data. However, the test RMSE shows more variability, reflecting challenges in generalizing to unseen data. Despite this, overall RMSE values indicate that the LSTM model maintains reasonable predictive accuracy across iterations.

In summary, the LSTM model demonstrates strong predictive capabilities for flu cases, with certain features being critical for accuracy. Variability in RMSE values underscores the importance of robust model evaluation, but the LSTM model's overall performance is promising for flu prediction.

G. Summary of Feature Importance

Feature	LR	GB	RF	FNN	RNN	LSTM
Public Healthcare Exp.	X			X	X	X
Air Pollution	X			X		
Population		X	X			
Mean Income/Day		X	X	X	X	X
International Pass.			X	X	X	
GDP			X	X		
Tourist Trips			X	X	X	X
Positive Flu Tests Max				X	X	X
Water Quality Deaths				X	X	
Median Age				X		X
Growth Rate				X		X

TABLE V
SUMMARY OF FEATURE IMPORTANCE ACROSS MODELS

Table V summarizes the importance of various features across different models used in the analysis. An 'X' indicates that the feature was deemed important for that particular model.

- **Linear Regression (LR)**: Public Healthcare Expenditure and Air Pollution were significant predictors in the OLS regression model.
- **Gradient Boosting (GB)**: Population and Mean Income Per Day were identified as the most crucial features.
- **Random Forest (RF)**: Mean Income Per Day and Population were the most important, with additional contributions from International Passengers, GDP, and Tourist Trips.
- **Feed Forward Neural Network (FNN)**: Positive Flu Tests Max, Mean Income Per Day, Air Pollution, Median Age, Growth Rate, and International Passengers were critical.
- **Recurrent Neural Network (RNN)**: Positive Flu Tests Max, International Passengers, and Mean Income Per Day were pivotal for predictions.
- **Long Short-Term Memory (LSTM)**: Median Age, Growth Rate, and Mean Income Per Day were significant, with less importance attributed to Positive Flu Tests Max and Tourist Trips.

This table helps in understanding which features have the most influence across different models, highlighting their relative importance in predicting the dependent variable.

VII. CONCLUSIONS

This research aimed to determine the most relevant features for predicting potential flu outbreaks, assess the suitability of recurrent neural networks (RNNs) for respiratory disease prediction, and evaluate whether ensemble models can enhance prediction accuracy compared to neural networks.

A. Relevant Preconditions for Predicting a Possible Outbreak

The analysis revealed that certain preconditions are particularly significant for predicting flu outbreaks. The feature importance analysis across different models consistently highlighted several key features:

- **Mean Income Per Day (Dollars)** - This was a critical factor, especially evident in both Gradient Boosting and Random Forest models.
- **Population (millions)** - This feature also consistently ranked highly across models.
- **Public Healthcare Expenditure (Percentage)** - Featured prominently, particularly impacting the accuracy of the RNN and LSTM models.
- **International Passengers (Millions)** and **Tourist Trips** - Both were important in the context of global movement and potential exposure to different flu strains.
- **Positive Flu Tests (Min and Max)** - These were crucial, as they directly reflect the incidence of flu cases, thus directly influencing predictive accuracy.

B. Suitability of Recurrent Neural Networks

The results indicate that Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, are indeed suited for predicting respiratory disease outbreaks. The RNN models demonstrated a strong ability to capture temporal patterns in flu cases, which is crucial for predicting outbreaks. The comparison of actual vs. predicted data in the graphs for both normalized and original data indicated that RNNs could closely follow the trends in the actual flu cases. However, the variability in RMSE values over multiple iterations suggests that while RNNs are effective, their predictions can be influenced by the inherent variability in the data, necessitating robust evaluation and potentially ensemble approaches to stabilize predictions.

C. Effectiveness of Ensemble Models

Ensemble models, particularly Gradient Boosting and Random Forests, were evaluated to determine their effectiveness compared to neural networks. The ensemble models generally showed improved accuracy over the neural network models. This was evidenced by the lower RMSE scores and higher R-squared values, indicating better fit and predictive performance. For example, the Gradient Boosting model's learning curve demonstrated consistent improvement and convergence, while the feature importance plots underscored the relevance of the identified features. The Random Forest model, with its high R-squared value and lower mean squared error, further validated the efficacy of ensemble approaches in enhancing predictive accuracy.

D. Societal Implications

The findings from this research have significant societal implications. Accurate prediction of flu outbreaks can enable better preparedness and response from public health authorities, potentially reducing the impact of outbreaks on communities. By identifying key preconditions, public health policies can be better targeted to mitigate risk factors associated with higher flu incidence. Moreover, the use of advanced predictive models, such as RNNs and ensemble models, can provide timely and actionable insights, allowing for more efficient allocation of healthcare resources and implementation of preventive measures.

REFERENCES

[1] National Academies Press, *Under the Weather*. (2001). Retrieved from https://doi.org/10.17226/10025

[2] Loo, B. P. Y., Tsoi, K. H., Axhausen, K. W., Cao, M., Lee, Y., & Koh, K. P. (2023). Spatial risk for a superspreading environment: Insights from six urban facilities in six global cities across four continents. *Frontiers in Public Health*, 11, 1128889. doi: https://doi.org/10.3389/fpubh.2023.112888910.3389/fpubh.2023.1128889. PMID: 37089495; PMCID: PMC10113652.

[3] Microsoft Power BI. (n.d.). Retrieved from https://app.powerbi.com/view?r=eyJrIjoiZTkyODcyOTEtZjA5YS00ZmI0LWFkZGUtODIxNGI5OTE3YjM0IiwidCI6ImY2MTBjMGI3LWJkMjQtNGIzOS04MTBiLTNkYzI4MGFmYjU5MCIsImMiOjh9

[4] Dattani, S., Spooner, F., Mathieu, E., Ritchie, H., & Roser, M. (2023). *Influenza*. Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/influenza

[5] S. D. and H. W., "Risk factors associated with respiratory infectious disease-related presenteeism: a rapid review," *BMC Public Health*, 2023. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-15285-6. [Accessed: May 14, 2024].

[6] Centers for Disease Control and Prevention, "Investigating Unexplained Respiratory Outbreaks (URDO)," 2023. [Online]. Available: https://www.cdc.gov/urdo/investigation-guidelines.html. [Accessed: May 14, 2024].

[7] World Health Organization, "Chronic respiratory diseases," 2023. [Online]. Available: https://www.who.int/health-topics/chronic-respiratory-diseases. [Accessed: May 14, 2024].

[8] Various Authors, "Mass Gatherings and Respiratory Disease Outbreaks in the United States – Should We Be Worried? Results from a Systematic Literature Review and Analysis of the National Outbreak Reporting System," *PLOS ONE*, 2023. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249678. [Accessed: May 14, 2024].

[9] Various Authors, "Investigating Unexplained Respiratory Outbreaks," *Oxford Academic Open Forum Infectious Diseases*, 2023. [Online]. Available: https://academic.oup.com/ofid/article/8/3/ofab078/6133719. [Accessed: May 14, 2024].

[10] Our World in Data, *Tourism*. (2024). Retrieved from https://ourworldindata.org/tourism

[11] World Population Review, *Countries Population*. (2024). Retrieved from https://worldpopulationreview.com/countries

[12] Our World in Data, *Public Healthcare Spending as a Share of GDP*. (2024). Retrieved from https://ourworldindata.org/grapher/public-healthcare-spending-share-gdp?tab=table

[13] Our World in Data, *Daily Median Income*. (2024). Retrieved from https://ourworldindata.org/grapher/daily-median-income?tab=table&time=2000..latest

[14] Our World in Data, *Air Pollution*. (2024). Retrieved from https://ourworldindata.org/air-pollution

[15] Our World in Data, *Age Structure*. (2024). Retrieved from https://ourworldindata.org/age-structure

[16] Our World in Data, *Population Growth Rates (2009-2010)*. (2024). Retrieved from https://ourworldindata.org/grapher/population-growth-rates?tab=table&time=2009..2010

[17] Our World in Data, *Air Passengers Carried*. (2024). Retrieved from https://ourworldindata.org/grapher/air-passengers-carried