Machine Learning based H1N1 and Seasonal Flu Prediction

An extensive data-driven examination of the correlation between predictive models and pupil's background

by

Syed Bukhari

Abstract

Health is a fundamental necessity. This research focuses on predicting H1N1 and seasonal flu vaccination uptake, stressing the role of demographic factors such as age, education, and income, using machine learning. The central research questions are: (i) How do demographic factors (age, education, income) influence perceptions of H1N1 and seasonal flu vaccine effectiveness and risks? The findings for RQ1 show higher education, income, and age correlate with favourable vaccine perceptions and lower risks, as confirmed by multivariate logistic regression, which shows significant associations between these demographic factors and vaccination perceptions; (ii) Is there a correlation between engaging in preventative health behaviours (e.g. hand-washing, face masks) and the likelihood of getting vaccinated? The findings for RQ2 show preventative behaviours strongly correlate with vaccination uptake, with multivariate logistic regression again revealing that these behaviours, combined with demographic factors, significantly predict vaccination likelihood; (iii) How do dual methodological approaches compare at different stages of the research process? The findings for RQ3 show Method A boosts accuracy but risks overfitting. Method B shows consistent generalisability. Previous studies show demographic factors are known to influence health behaviours, but previous predictive models have often underutilised this data, leading to lesser accuracy scores. This study addresses a gap by incorporating demographic factors into predictive models, which significantly improves accuracy. This enhancement is critical for better understanding disease spread and vaccination likelihood. Furthermore, this study approaches the data from a different perspective by framing it through the aforementioned research questions. The study used a dual approach, combining traditional statistical analysis with machine learning models such as Random Forest and XGBoost. Data from the National 2009 H1N1 Flu Survey, which includes comprehensive demographic and health variables, was processed to train and validate these models. Key findings show education and income correlate with more favourable vaccine perceptions and higher vaccination rates. Preventive behaviours also significantly enhance vaccination likelihood. Machine learning models outperformed other studies, achieving 85-90% accuracy. These findings inform targeted public health strategies, which boosts vaccination rates and improves overall health outcomes.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background and Motivation

Health is a fundamental human right. "The enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition" was quoted by the 1948 Universal Declaration of Human Rights. [27] It was also mentioned health as part of the rights to an adequate standard of living (art. 12). [28] This right was then recognised again as a human right in the 1966 International Covenant on Economic, Social and Cultural Rights. [29]

The H1N1 pandemic of 2009, also known as swine flu, had a profound impact on a global scale, resulting in thousands of morbidity and mortality. It is estimated the H1N1 virus led to approximately 60.8 million cases in the U.S. alone. This resulted in 274,304 pupils being hospitalisaed and 12,469 deaths, with a large portion of the severe outcomes affecting younger populations compared to seasonal flu, which had a higher burden on older adults. [30] Furthermore, the World Health Organisation (WHO) declared H1N1 a pandemic in June 2009. This led to an estimated 151,700 to 575,400 deaths worldwide within the first year. [30] The United States mounted a complex, multi-faceted and long term response to the pandemic. On August 10, 2010, WHO announced the official end of the 2009 H1N1 influenza pandemic. Despite this declaration, the H1N1 virus remained in circulation as a seasonal influenza strain. This strain continues to evolve and cause numerous cases of illness, hospitalisation and deaths worldwide yearly. [31]. Figure **??** depicts CDC estimates of 2009 H1N1 hospitalisations in the U.S. from April 2009 to March 2010, with ranges indicated.

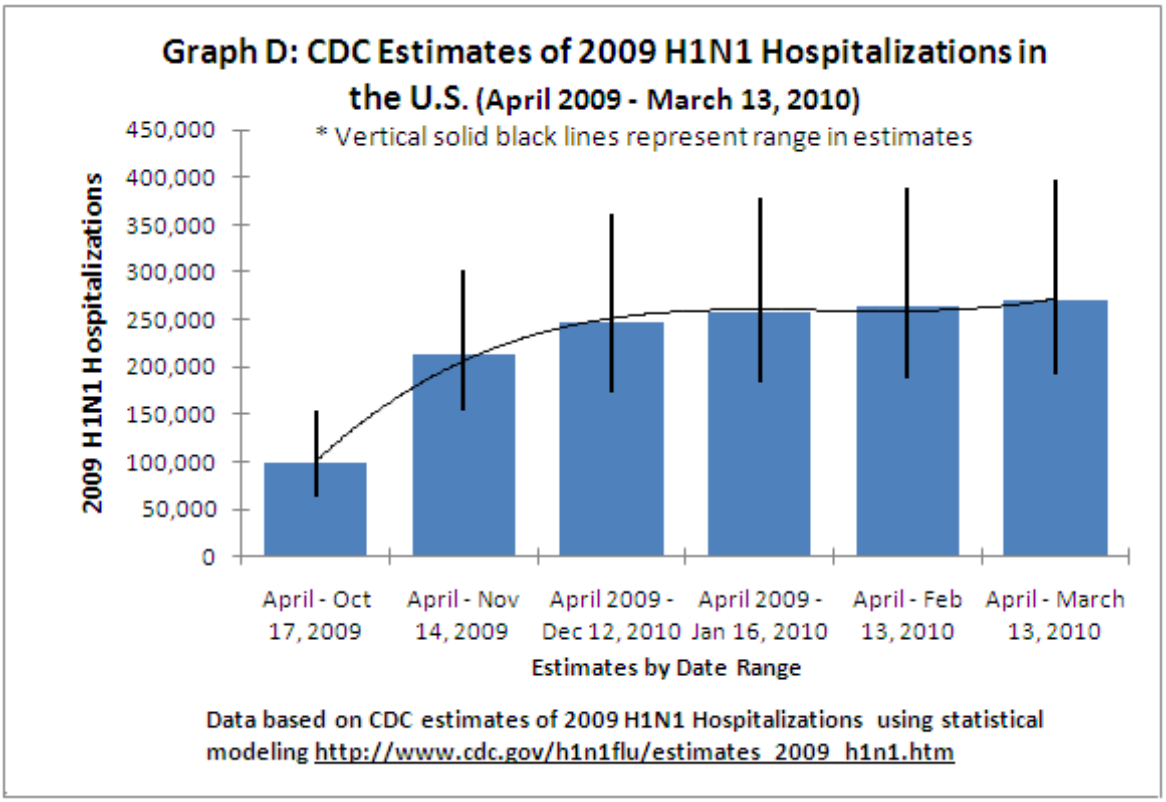

Figure 1: CDC estimates of 2009 H1N1 hospitalisations in the U.S. from April 2009 to March 2010, with ranges indicated [90]

Therefore, although the pandemic has been official concluded, newly evolved influenza viruses still continue to result in global fatalities every year where importance to demographic, social, and environmental factors of a strain need to be determined to stop the wide spread of viruses decreasing fatalities. Additionally, season flu, although generally less deadly, consistently affects millions of pupils each year, causing severe illness and death: "There are around a billion cases of seasonal influenza annually, including 3–5 million cases of severe illness; It causes 290,000 to 650,000 respiratory deaths annually; Ninety-nine percent of deaths in children under 5 years of age with influenza-related lower respiratory tract infections are in developing countries; Symptoms begin 1–4 days after infection and usually last around a week". [34] These deaths particularly lie amongst the high-risk populations such as young children, the elderly, and individuals with underlying health conditions (i.e. myocardial infarction). [35]

Understanding the data and patterns from the 2009 H1N1 pandemic is crucial for data science. It allows for the analysis of epidemiological trends, the effectiveness of health care interventions, and development of predictive models for future outbreaks. While a pandemic-level outbreak might have seemed improbable, just 15 years later, in 2020, the WHO declared the COVID-19 pandemic, which profoundly disrupted lives worldwide for over two years. [32] The significance of this project is underscored by the substantial impact of the Coronavirus pandemic, illustrating devastating effects a virus can have. 704,753,890 cases were confirmed worldwide, resulting in 7,010,681 deaths and 675,619,811 recoveries. [33] This translates to 8.81% of the global population being affected, a 0.088% death rate, and an 8.45% recovery rate. Gratefully, the detailed records of infection rates, hospitalisation, and mortality provide a rich dataset for studying the dynamics of pandemic spread and healthcare system responses.

This study addresses a significant gap in predicting flu shot uptake among various demographics using historical data. Identifying key determinants of low flu shot uptake is essential for targeting interventions to increase the vaccination coverage. With improved vaccination rates, the public can significantly reduce illness and healthcare costs demonstrating the importance of this study. The rapid evolution of viruses in general, as seen with H1N1's shift to a seasonal strain, highlights the important need for timely and effective predictive models to aid in vaccination strategies. As a comparative note, COVID-19, caused by the SARS-CoV-2 virus, has evolved and undergone numerous mutations since it has been first identified. A wide variety of viruses were identified: Alpha (B.1.1.7) identified in the United Kingdom showing increased transmissibility compared to the original strain. Beta (B.1.351) identified in South Africa, known for having mutations that affect vaccine efficacy. Gamma (P.1) identified in Brazil, showed increased transmissibility and impacts on immunity. Delta (B.1.617.2) identified in India, became a dominant strain globally due to high transmissibility and severe diseases. Omicron (B.1.1.529) identified in South Africa again, noted for large number of mutations, particularly in the spike protein, affecting immune escape. [36] This is being mentioned to illustrate how fast viruses can evolve and how immediate action should be taken. Identifying factors, patterns, and trends in the topic of data science help mitigate such issues preventing deaths.

In conclusion, the objectives of this study are to (i) develop a machine learning model to predict the likelihood of individuals receiving vaccinations for H1N1 and seasonal flu; (ii) find key factors influencing vaccination uptake; in order to meet this objective, it is necessary to (iii) find how dual methodological approaches compare at different stages of the research process. The aforementioned objectives are clearly outlined and reiterated in Section 1.2. By predicting vaccination uptake and identifying key influencing factors, this study addresses future challenges and disruptions in public health. It aids in identifying and mitigating risks associated with low vaccination rates and potential flu outbreaks.

The thesis is organised as follows: Chapter 1.2 clearly outlines the projects aims and objectives. The literature review, Chapter 2, critically reviews existing literature on ML methodologies for predicting H1N1 and seasonal flu, identifying key gaps and position this study within the broader research context. The methodology, Chapter 3 outlines the dual methodological approach, including data collecting, preprocessing techniques, feature selection and the selection and justification of machine learning models. The implementation, Chapter 4 describes the implementation process, focusing on practical application of selected methodologies, including steps mentioned in the methodology being applied. The results, Chapter 5 presents the results, providing rigorous evaluation of predictive models, including performance and comparison with dual approaches. The discussion, Chapter 6, offers comprehensive discussion of findings, linking them back to the research objectives and literature, while highlighting the study's contributions, limitations, implications for future research, LSEPI, and the project management.

1.2 Aims & Objectives

The objective of this study is to develop a machine learning model to predict the likelihood of individuals receiving vaccinations for H1N1 and seasonal flu. In order to meet this objective, it is necessary to find the key factors influencing vaccine up take.

The aim is to:

- Find the key factors influencing vaccination uptake.
- Finding how dual methodological approaches compare at different stages of the research process.

The research question is:

RQ1: How do demographic factors (age, education, income) influence perceptions of H1N1 and seasonal flu vaccine effectiveness and risks? – This comparative question investigates how background factors affect beliefs about vaccine safety and efficacy, potentially impacting vaccination rates.

RQ2: Is there a correlation between engaging in preventative health behaviours (hand- washing, face masks) and the likelihood of getting vaccinated? – This associative question explores whether individuals practising preventative measures are more likely to get vaccinated, indicating a broader health-conscious attitude.

RQ3: How do dual methodological approaches compare at different stages of the research process? – This comparative question examines how different quantitative methods at different stages of the research process provide insights, focusing on their strengths and weakness to improve model accuracy and generlisability.

2 Literature Review

This literature chapter explores ML methodologies for predicting various diseases. Its goal is to review existing literature on predictive models, evaluate their effectiveness and limitations, and justify the methods chosen for this research. The chapter is organised as follows: Subchapter 2.1 outlines the focus areas of the literature review, including ML models, preprocessing techniques, and evaluation metrics; Subchapter 2.2 describes the systematic approach used to gather and analyse relevant studies; Subchapter 2.3 provides a critical discussion of key literature, categorised by methodologies (e.g. data-driven methods and neural networks); Subchapter 2.4 identifies gaps in current literature and discusses how this research addresses them. Understanding the strengths and limitations of existing methodologies is vital for developing accurate models and informing the steps for creating disease predictive models. Insights from this review guide the methodological choices in Chapter 3, implementation in Chapter 4, and evaluation in Chapter 5.

Research into ML applications for predicting influenza and other infectious diseases has gained significant tractions over the past years, especially due to the recent COVID-19 pandemic. [37] [38]Initial efforts were primarily focused on statistical models (i.e. ARIMA). Some example of these works can be found in Subchapter 2.3. However, recent years have seen a surge in using sophisticated ML techniques. The key RQ and hypotheses relate to accuracy of prediction models to determine if a machine learning model can predict influenza outbreaks compared to traditional statistical models, and if research can do it well enough to rely on. Moreover, can the integration of climatic, spatial, and demographic data improve the predictive accuracy of these models? This research heavily focuses on demographic data in hopes to improve predictive accuracy of ML models. Moreover, how adaptable are these models to different diseases and geographic data? and lastly, can early detection models effectively aid in preventing large-scale outbreaks?

2.1 Survey Scope

The scope of this survey focuses on ML-based prediction models for H1N1 and seasonal influenza vaccination rates, with a particular emphasis on the correlation between these models and individuals' background factors. The survey includes:

- ML models/techniques & algorithms used for predicting vaccination likelihood.
- Factors related to an individual's background that may influence vaccination rates, demographic data is crucial.
- Preprocessing and feature engineering techniques for handling vaccination and background data.
- Evaluation metrics and methodologies for assessing the performance of predictive models in this domain.
- Addressing challenges such as class imbalance, interpretability, and generalization to different populations and time periods.

However, the survey does not cover:

- Detailed biological or epidemiological aspects of H1N1 and seasonal influenza viruses, unless directly relevant to the predictive modeling process.
- General ML techniques or algorithms not specifically applied to the context of influenza vaccination prediction or similar viruses or health conditions.
- Public health policies, vaccination campaigns, or resource allocation strategies, unless directly informed by the predictive models under consideration.

2.2 Literature Review Scope and Method

A structured approach was followed to identify relevant research papers. This methodology was guided by the advice of my supervisor, Dr. Helena Webb, and utilized a combination of keyword searches, database exploration, and citation analysis. Below is a detailed description of the steps taken:

Based on the topic and Dr. Webb's advice, I selected a set of keywords, "machine learning", "H1N1 prediction", "seasonal flu prediction", "vaccination rates", "predictive models", "health data analytics", "influenza vaccination" that are crucial for finding relevant literature.

Furthermore, ensuring a comprehensive search, Google Scholar, PubMed, IEEE Xplore was utilised for its extensive database of academic papers and connected papers was utilised to explore a visual citation networks of key publications (see Figure 2).

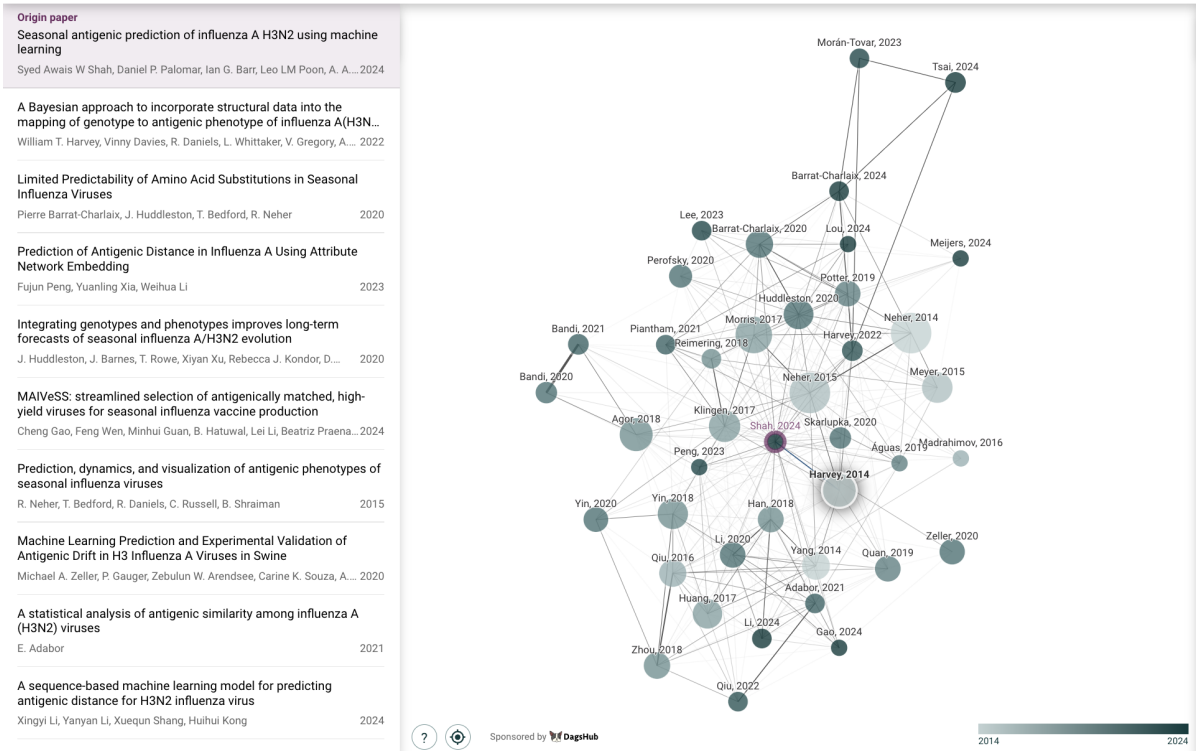

Figure 2: Influenza research network visualisation (from connectpapers.com)

Dr Helena Webb noted that when viewing scholary articles, keen attention should be paid to the date of publication as given the rapidly evolving nature of ML applications in healthcare, papers that were published in last five years was prioritised. Secondly, preference was given to papers published in high-impact journals, and with significant number of citations that indicated influence in the field.

I initiated the search with the selected keywords in Google Scholar, applying filters to include only recent publications (from 2018 onwards). The initial search results were screened for relevance based on titles and abstracts. Papers that directly addressed ML applications in predicting H1N1 and seasonal flu vaccination rates were prioritized. Moreover, through connected papers, research papers were found that were not identified in initial google scholar search.

A list was compiled a list of papers, prioritizing them based on their relevance to the topic, the impact factor of the publishing journal, and the number of citations.

Furthermore, regarding the academic disciplines and fields, the literature reviewed mainly focuses on four topic areas: it focuses on computer science, applying machine learning algorithms, data preprocessing techniques, and ensembled learning methods; epidemiology emphasising disease spread, public health interventions, and statistical models for outbreak prediction; addressing health resource management, vaccination strategies and early detection system in the field of public health; and data science where they utilise mathematical models and simulations to predict disease dynamics and outbreaks. This study research field overlapses with computer science and data science.

Category **Perception**	**Data Enhancement & Transformation**	**Visual Mapping & Structure**	**Exploration & Rendering**	**Interaction & Analysis**
Data-Centric (data-types)	**Venkatramanan et al. (2021)**: Uses anonymized mobility data to predict influenza spread.			
Multivariate & Hierarchical (hierarchical)	**Yakovyna et al. (2024)**: Combines supervised and unsupervised learning for COVID-19 predictions.	**Xue et al. (2018)**: Utilizes multiple regression and neural networks for flu activity prediction.		**Du et al. (2021)**: Hybrid ensemble model for early influenza outbreak prediction.
Graphs & Networks				
Geospace + Time (temporal)			**Venna et al. (2019)**: LTSM model capturing temporal dynamics for flu forecasting. **Watmaha et al. (2021)**: CNN model for real-time influenza forecasting incorporating spatial-temporal data.	
Co-ordinated Multiple Views				
Real-World & Applications (healthcare)	**Shah et al. (2024)**: Predicts antigenic properties of influenza A H3N2 viruses.		**Xi et al. (2018)**: Uses deep residual networks to predict influenza trends. **Watmaha et al. (2021)**: CNN model for real-time influenza forecasting incorporating spatial-temporal data.	**Yanamala et al. (2021)**: ML models to differentiate between COVID-19 and influenza. **Kim et al. (2023)**: Evaluates ML models to predict hepatitis B and C in diabetic patients.
Overview				

Table 1: Classification of Literature and Organization

2.3 Critical Discussion of Literature

2.3.1 Data-driven Methodologies

Du et al. [4] propose a data-driven methodology for detecting and predicting influenza outbreaks using diagnostic data from over 3,000 clinics in Malaysia. They develop a new region index (RI) to capture trends and apply an ensemble learning method for outbreak prediction, building on models like GARMA and Bayesian regression, and citing Dugas' research on forecasting influenza cases. The dataset includes approximately two million ILI case records from January 2016 to July 2019. The model, evaluated using recall, precision, and accuracy scores, achieves 75% recall, 74% precision, and 83% accuracy, validated with WHO and GFTs data. The RI addresses biases from clinic size and number variations, enhancing sensitivity and prediction reliability, and enables earlier outbreak prediction than traditional methods (GFT). However, the model may need adaptation for other diseases or regions, as it can be extended to the city level or any granularity by grouping clinics geographically. This methodology supports

early detection and prediction of influenza, aligning with the goal of using predictive models for effective public health resource management. In summary, this work presents a reliable model for early influenza detection, applicable to the 2009 H1N1 dataset to improve predictive accuracy and resource planning. The tools and methods, like the RI and ensemble learning, can be adapted for other diseases. Du et al. [4] and Venna et al. [19] both used advanced machine learning for influenza prediction—Du et al. [4] focused on the region index and ensemble learning, while Venna et al. [19] integrated climatic and spatio-temporal factors, highlighting the benefits of external data sources for improved prediction accuracy.

2.3.2 Neural Networks and Deep Learning

Venna et al. [19] demonstrated the effectiveness of an LSTM-based deep-learning method, surpassing existing time series forecasting models. Their two-stage approach involved LSTM for initial forecasting, followed by error adjustment using climatic and spatio-temporal factors. Building on ARIMA and EAKF models, they enhanced predictions with CDC-reported ILI data (2002-2006) and GFTs (2003-2014), evaluated using MAPE, RMSE, and RMSPE. This model outperformed ARIMA and EAKF by integrating environmental data, though it relies on large datasets for training. The research improves predictive accuracy for H1N1 and seasonal flu datasets, though challenges remain in data reliability and model scalability. Additionally, Venna et al. [19] and Watmaha et al. [21] both used neural network-based models to enhance influenza forecasting with environmental data. Venna used LSTM, while Watmaha employed CNNs, which outperformed RNN and LSTM in handling complex patterns. Watmaha proposed a CNN model integrating climate and spatio-temporal data, using 1D CNNs for climate and 2D CNNs for flu count data, achieving better accuracy than ARIMA and LSTM models with U.S. data from 1997-2016. Although successful, this CNN model isn't applicable to 2009 H1N1 and seasonal flu datasets, which lack the necessary features for this study.

Compared to Watmaha et al. [21], who used CNNs for broader geographic regions, Xi et al. [22] focused on predicting intra-urban influenza trends by integrating spatial-temporal properties with a deep residual network. Xi et al. used a CNN with deep residual units to capture spatial dependencies and prevent model degradation, improving accuracy over models like linear regression, ANN, LSTM, and spatiotemporal LSTM, which ignored intra-city spatial correlations. Their data included historical ILI percentages from Shenzhen, China (Jan 2015 - Aug 2017) across 10 districts. The model, evaluated using mean absolute error and mean absolute percentage, showed better accuracy, especially for one- and two-week forecasts. The strength is in spatial-temporal integration, though it demands significant computational resources, enhancing influenza prediction by addressing intra-city spatial correlations.

Taj et al. [17] developed an LSTM-based RNN model to predict COVID-19 cases in Morocco using daily regional data, building on prior LSTM models for influenza prediction. Data was collected daily from Morocco's official Coronavirus Portal, including confirmed, recovered, and death cases. The model, evaluated using mean absolute error, effectively aided pandemic response decisions. It emphasizes the importance of demographic factors like population density and the need for region-specific approaches due to varying pandemic impacts. The model's strength is in incorporating demographic and temporal factors, though it is limited by data accuracy and availability. This validated LSTM model is adaptable to H1N1 and seasonal flu datasets. In contrast, Marquez et al. [12] employed a different methodological approach for diagnosing seasonal flu.

Marquez et al. [12] applied machine learning to diagnose seasonal influenza using clinical and demographic data. They used Random Forest and Bagging classifiers on 15,480 patient records from Mexico City (2010-2020), employing random oversampling (ROS) to balance the

training set. Building on prior ML studies for influenza prediction, the models were evaluated using accuracy, specificity, sensitivity, precision, F1-measure, and AUC. Random Forest performed best with an AUC of 0.94 and accuracy of 0.86. The study highlights the effectiveness of ensemble methods like Random Forest and Bagging, particularly with ROS, though the binary representation of clinical features may limit the model's representativeness. The research demonstrates ML's potential to provide alternative diagnostic tools in regions with limited RT-qPCR access, reducing reliance on costly molecular tests and increasing diagnostic accessibility in low-resource settings. Both Taj et al. [17] and Marquez et al. [12] illustrate the effectiveness of ML in epidemic prediction and management.

Elbasi et al. [5] and Marquez et al. [12] both used the Random Forest classifier for disease classification. Marquez et al. [12] focused on diagnosing seasonal flu, while Elbasi et al. [5] applied this method to classify patient data into H1N1 and COVID-19 categories, showcasing machine learning's adaptability. Elbasi compared various algorithms, including Bayes network (86.57%), naive Bayes (82.34%), multilayer perceptron (99.31%), locally-weighted learning (88.89%), and Random Forest (83.16%). Using a dataset of 1,467 patient records (70% H1N1, 30% COVID-19) with 42 features, the study evaluated accuracy, true positive rate, false positive rate, precision, and recall. The multilayer perceptron achieved the highest accuracy (99.31%), outperforming other algorithms. Despite its comprehensive algorithm comparison, the study is limited by its small dataset and fixed features, affecting generalisability. The 2009 H1N1 and seasonal flu datasets, with 27,000 patient records and 39 features, offer better potential for broader application and improved predictive accuracy.

2.3.3 Time Series and Forecasting Models

Sultana and Sharma [15] predicted future swine flu cases in India using various time series forecasting models to help the Indian Government prevent the disease's spread. They employed models like Box-Cox transformation, Exponential Smoothing, Seasonal Naive, and Neural Networks. The study builds on methodologies such as the Ljung Box test and Augmented Dickey-Fuller test, using data from 2010 to 2017 from India's Integrated Disease Surveillance Programme. Evaluation metrics included mean error, mean absolute error, root mean square error, and mean absolute scaled error. The neural network model achieved the highest accuracy at 98.4%, outperforming other models, demonstrating its effectiveness in handling complex patterns. Traditional methods like Box-Cox and Exponential Smoothing were less effective. The study's strength lies in the neural network's low error values, but traditional models like Seasonal Naive showed high errors. Sultana and Sharma [15] also highlighted the need for ensemble models to improve prediction accuracy, noting this as a direction for future research. This work can enhance the thesis by offering robust approaches for predicting flu cases, applicable to the 2009 H1N1 U.S. dataset.

2.3.4 Combined Data Sources for Disease Prediction

Similarly, Sultana and Sharma [15] and Xue et al. [23] used time series forecasting models to predict influenza activity. Sultana and Sharma employed traditional models like Box-Cox and Neural Networks, while Xue et al. developed multiple regression and neural network models using GFT and CDC data. Xue et al. utilised least squares and backpropagation neural networks, optimised by genetic algorithms, to fit model parameters, comparing error and sample fitting accuracy. Their data included 604 weeks of GFT and CDC data from 2003 to 2015 for training, with 2015 data for testing. Models were evaluated using mean square error, mean absolute percentage error, and relative mean square error, with the combined GFT + CDC regression model, especially with seasonal adjustments, offering superior prediction accuracy. This aligns with Ayachit et al. [1], who also combined GFT and CDC data for regression. Xue et al.'s

study emphasises the value of integrating historical and real-time data and tailoring models to seasonal flu variations, noting that the combined GFT + CDC model enhances accuracy. However, GFT data alone may cause errors and overestimations in some flu seasons. The literature supports using combined data sources for better flu prediction and highlights the limitations of single-source models, providing a comprehensive overview of influenza prediction methods.

Lober et al. [11] developed a machine learning model using Random Forests and XGBoost to predict disease cases across 1,804 cities for Dengue, 211 for Zika, 274 for Influenza, and 5,564 for COVID-19 in Brazil, incorporating socioeconomic and geographic data. The study used weekly case data from 2014 to 2023, sourced from Brazilian government databases, and evaluated model performance using mean absolute error against a seasonal naive baseline. Geographic data significantly improved predictions for COVID-19 and Zika but not for Dengue and Influenza, showing the varying impact of geographic and socioeconomic factors on different diseases. The study highlights the importance of these factors in predictive models for certain diseases but notes limitations for others. This work provides a framework for enhancing predictive models with geographic and socioeconomic data, particularly for diseases like H1N1. From Lober et al. [11] and Xue et al. [23], it is evident that diverse datasets can enhance model performance, though there are gaps in improving predictions uniformly across all diseases. Given the absence of geographic data in the H1N1 dataset, it suggests the need to explore the impact of socioeconomic factors on model accuracy.

Viana et al. [20] and Venkatramanan et al. both emphasise the importance of real-time data in disease prediction. Viana et al. developed and compared machine learning models to optimise test prioritisation for COVID-19 and influenza during concurrent outbreaks. They trained eight algorithms—decision tree, multilayer perceptron, gradient boosting machine, Random Forest, extreme gradient boosting, k-nearest neighbours, support vector machine, and logistic regression—using Brazilian demographic and symptom data. The models were validated with 10-fold cross-validation, involving 55,676 individuals tested for COVID-19 and 14,570 for influenza. Evaluation metrics included precision, recall, accuracy, Brier score, ROC, PR curve, and statistical validation with Friedman and Nemenyi tests. Tree-based models like GBM, RF, and XGBoost performed best, offering high interpretability, aiding healthcare decision-making, and optimising resource allocation during coexisting outbreaks. However, the study did not consider simultaneous COVID-19 and influenza infections. This work addresses the gap in ML models for concurrent outbreaks in low-resource settings, contributing to optimising test prioritisation and enhancing clinical decision support systems. While Viana et al. focus on test prioritisation during concurrent outbreaks, Venkatramanan et al. use mobility data for real-time influenza forecasting, each addressing different aspects of outbreak management.

2.3.5 Convolution Theory and Rapid Decision-Making

Interestingly, to combat resource usage, Nieto-Chaupis [13] creates an application of the Wiener series-based machine learning for rapid decision-making in flu pandemic scenarios. The methodology employs the Wiener series with a machine learning framework to model and predict flu outbreaks, using convolutional integrals to estimate interventions and optimise resource usage.

$$E(t) = \int P(\tau_1)T(t-\tau_1)\,d\tau_1$$

The above formula was followed, treating experience as an ordinary output from the input-output theory perspective. This paper builds on convolution theory and machine learning algorithms, particularly those by Mitchell and Wiener. The data focuses on the 2009 A(H1N1) flu outbreak in Peru, specifically the case velocity in Lima. Simulations compared the machine learning model's predictions with actual outbreak data, evaluating metrics like the number of

cases over time. The model includes interventions like rapid decision-making and resource optimisation tailored to user needs, highlighting the necessity of adaptable approaches in different outbreak scenarios. Venkatramann et al. also addressed predictive models in pandemic management, with Nieto-Chaupis using Wiener series for rapid decision-making and Venkatramann et al. using mobility data to improve disease prediction. The strength of Nieto-Chaupis's model is its effective mimicry of real-world data, showing that Wiener series can significantly aid in managing flu outbreaks by optimising public health responses and resource allocations, making it a reliable tool for public health management. However, its reliance on historical data may limit predictive accuracy in novel outbreaks without similar data patterns. The literature aligns with this research by emphasizing the need for advanced computational methods in outbreak management but highlights the gap in real-time adaptability. This work contributes to understanding machine learning applications in pandemic responses, offering tools that enhance predictive accuracy and decision-making, with insights applicable to developing predictive models.

2.3.6 Mobility Data and Metapopulation Models

Venkatramanan et al. [18] used anonymised mobility app data from smartphones to forecast influenza activity, implementing a metapopulation model with Bayesian calibration. This study builds on existing mobility models in epidemiology, such as gravity and radiation models, using data on anonymised mobility flows, emergency department visits in NYC, and flu-positive counts for New Jersey and Australia. The AMM model's performance was compared to traditional commuter surveys and gravity/radiation models using MAPE. Results showed the AMM model performs comparably to traditional methods, predicting disease spread across state boundaries and highlighting the utility of machine-learned mobility data for real-time disease forecasting. While the study's strength is in forecasting disease spread globally with high-quality mobility data, it does not distinguish between residents and transients. Venkatramanan et al. [18] underscore the importance of real-time mobility data in improving disease forecasting, addressing gaps that the AMM model aims to fill, and exploring methods to combat influenza and other diseases.

2.3.7 Vaccine Uptake and Predictors

Byrne et al. [3] and Busse et al. [2] both studied vaccine uptake factors but with different focuses. Byrne et al. [3] investigated behavioural predictors in university students, while Busse et al. [2] assessed immunogenicity and safety in asthma patients, showing the varied determinants influencing vaccination strategies. Byrne et al. [3] focused on university students' intentions to receive the H1N1 vaccine during the 2009 pandemic. They conducted a cross-sectional survey with 200 students, using the Theory of Planned Behaviour and the Health Belief Model to predict vaccination intentions. Key predictors included health status, trust in authorities, optimism, conscientiousness, self-efficacy, and Health Belief Model components. Logistic regression and Kruskal-Wallis tests linked non-intention to vaccinate with disbelief in vaccine efficacy, negative attitudes, and low perceived threat, underscoring the role of psychoeducation. However, low internal consistency in perceived benefits and social influence measures limited their explanatory power. This study is essential for understanding psychological predictors of vaccination intent and can inform interventions for H1N1 and seasonal flu vaccination uptake.

Busse et al. [2] approached the topic from a different angle, assessing the immunogenicity and safety of an unadjuvanted 2009 H1N1 vaccine in asthma patients. The open-label trial involved 290 participants aged 12-79, who were randomised to receive either 15mg or 30mg doses, administered twice, 21 days apart. The study measured seroprotection, seroconversion rates, reactogenicity, asthma exacerbations, and pulmonary function. The 30mg dose provided higher

seroprotection (94.1%) than the 15mg dose (77.9%), particularly in severe asthma patients, highlighting the need for tailored vaccination strategies based on age and asthma severity. Despite its strengths, the study lacked a healthy control group. These findings emphasise the importance of demographic and health factors in vaccine efficacy, informing flu vaccination strategies for at-risk populations.

2.3.8 Hybrid and Ensemble Learning Models

Yakovyna et al. [24] introduced hybrid machine learning ensembles combining supervised and unsupervised learning for COVID-19 classification and regression. The approach used clustering, Boruta feature selection, decision trees, and Random Forest, followed by a stacking ensemble with a Random Forest meta-model. The study builds on prior COVID-19 prediction techniques, including hybrid CNNs and LSTM algorithms. Data included daily COVID-19 cases and deaths across 3,142 U.S. counties from January 2020 to June 2021, with 46 features covering demographics, geography, climate, and social factors. Evaluation used metrics like accuracy, F1-score, ROC-AUC for classification, and MSE, ROC-AUC, and MPP for regression, with nested five-fold cross-validation. The best classification performance reached 0.912 accuracy, 0.916 ROC-AUC, and 0.916 F1-score, with an 11% increase in regression accuracy. The hybrid classifiers outperformed single algorithms, demonstrating the effectiveness of combining supervised and unsupervised learning. The study's strength is its significant improvement in prediction accuracy, though it requires extensive computational resources. This novel ensemble approach could enhance predictive models for H1N1 and seasonal flu.

However, Singh and Mittal [14] took a different approach by optimising machine learning parameters using Ant Colony Optimisation (ACO). They introduced a model for pandemic outbreaks that optimises machine learning parameters in Polynomial Regression, SVM, and Linear Regression using ACO. Building on previous disease prediction models, they applied climate factors for malaria outbreak analysis to identify the best-suited algorithms. The dataset included daily COVID-19 reports from Johns Hopkins University, covering 90 days for training and testing. Evaluation metrics included accuracy and RMSE score, with the PR-ACO model showing the highest accuracy and lowest RMSE, proving its effectiveness. Singh and Mittal's study focuses on improving outbreak prediction accuracy through ACO, contrasting with non-optimised approaches. The PR-ACO model's strength lies in its higher accuracy and reduced error rates, though it was limited to the COVID-19 dataset. This research shows how optimisation techniques can enhance ML models for pandemic predictions, with potential applications for H1N1 and seasonal flu data. Both Yakovyna et al. [24] and Singh and Mittal [14] explore hybrid and optimised machine learning models to improve pandemic predictions in their own ways.

Indhumathi et al. [7] contributed by using a boosted Random Forest algorithm to achieve 95% accuracy in forecasting seasonal infections based on electronic health records. This approach combines boosting techniques with multiple decision trees to enhance prediction accuracy. The study builds on previous work using geographically weighted regression models for tracking influenza spread. The data includes health records from a hospital, covering symptoms, patient area, age, gender, and month of visit. The boosted Random Forest algorithm was validated using accuracy metrics, achieving a 95% accuracy score, significantly improving predictive accuracy over existing models and demonstrating practical utility in forecasting disease outbreaks. Indhumathi et al. also employed numerical and categorical imputation for missing data and used the Boruta algorithm for feature selection, efficiently extracting 10 relevant attributes from 50. The study's strength lies in the improved accuracy of the boosted Random Forest compared to other machine learning algorithms, though the time complexity (1.98 seconds) is higher than that of the Boruta algorithm. The use of machine learning techniques proved effective in predicting future seasonal infectious diseases by combining climate

and disease data. The methods used in this study, particularly the boosted Random Forest algorithm, could be applied to the 2009 H1N1 and seasonal flu dataset to enhance predictive accuracy and identify significant seasonal patterns. Indhumathi et al.'s study aligns with hybrid methodologies and the use of machine learning models for disease prediction.

2.3.9 Machine Learning Models for Disease Prediction

While Indhumathi et al. [7] achieved high accuracy in forecasting seasonal infections using a boosted Random Forest, Kim et al. [9] took a different approach by predicting hepatitis in diabetic patients using various ML models. Kim et al.'s contribution lies in evaluating machine learning models to predict hepatitis B and C virus infections in diabetic patients. They used four models—RF, SVM, XGBoost, and LASSO—with hyperparameter tuning for improved performance, and applied SMOTE for preprocessing and balancing. The study utilised NHANES data from 2013 to 2018, including demographics, body measurements, lipids, and questionnaire data from 1,396 diabetic patients. Evaluation metrics included accuracy, sensitivity, specificity, precision, F1 score, and AUC-ROC, with LASSO achieving the highest accuracy (0.978) and specificity (0.993). The study identified illegal drug injection and poverty as the most significant predictors for hepatitis in diabetic patients, highlighting variability in predictors across populations. The study effectively identified key predictors using robust ML models, though it was limited by the cross-sectional nature of the NHANES data and lack of serological data. The research demonstrates the potential of ML models in identifying significant predictors and aiding early detection of hepatitis in diabetic patients, addressing gaps in traditional screening methods. Kim et al.'s study provides a foundation for applying these methods to other datasets, such as the U.S. 2009 H1N1 and seasonal flu datasets, enhancing predictive accuracy and model reliability. Each study demonstrates the effectiveness of tailored models for specific diseases.

Syed et al. [16] developed a machine learning model using AdaBoost with an ensemble of decision trees to predict the antigenic properties of influenza A H3N2 based on HA1 sequences and metadata. The model, trained on past season data, used genetic differences encoded with a mutation matrix and metadata features like virus avidity and antiserum potency, adopting a seasonal prediction approach. The study utilised 36,709 virus-antiserum pairs across 37 influenza seasons (2003NH to 2021SH), with genetic data from GISAID and IVR databases. The model's performance over 14 test seasons achieved a mean absolute error (MAE) of 0.702 antigenic units, effectively distinguishing antigenic variants and identifying key HA1 sites influencing changes. Incorporating partial antigenic data from circulating isolates significantly improved prediction accuracy, although performance varied due to the choice of the amino acid mutation matrix. This work highlights the importance of comprehensive metadata in improving model accuracy, relevant to predictive modeling for the 2009 H1N1 and seasonal flu datasets.

Moreover, Syed et al.'s [16] and Yanamala et al.'s [25] studies focus on different disease predictions, demonstrating varying approaches to enhancing predictive accuracy for viral infections. Yanamala et al. [25] developed and validated a supervised machine learning pipeline using the XGBoost model to distinguish between COVID-19 and influenza based on vital signs and demographic data, achieving high accuracy and AUC. The WVU patient cohort was divided into training (80%) and testing (20%) sets to develop four context-specific XGBoost models. The study used data from 3,883 patients, with external validation from 15,597 encounters in 3,125 patients, and achieved high accuracy (98.8%) and AUC (92.8%) in distinguishing COVID-19 from influenza. The study emphasises the utility of machine learning models in clinical settings for rapid and accurate diagnoses, though it is limited by reliance on specific vital signs and demographic data, which may not be available in all settings. Yanamala et al.'s findings are relevant to differentiating viral infections using machine learning, highlighting gaps in model applicability across diverse populations. The methods and models in this study can inform the

development of similar models for H1N1 and seasonal flu, potentially improving early diagnosis.

Li et al. [10] developed the ELPPJD method for multilabel disease risk prediction using physical examination records. They transformed multilabel classification into multiclass classification and used pruned datasets with joint decomposition to address imbalance learning. This study builds on multilabel classification methods like RAkEL and HOMER. The dataset included 110,300 anonymous physical examination records, focusing on 62 examination items and 6 chronic diseases. Evaluation metrics such as average accuracy, precision, recall, and F1 score were validated through 10-fold cross-validation. The ELPPJD_LS model achieved an average accuracy of 88.59%, outperforming RAkEL and HOMER, and significantly improving disease risk prediction based on physical exams. The ELPPJD method effectively tackles imbalance in multilabel data by introducing pruning and joint decomposition strategies. However, its complexity may increase with more labels, leading to higher time complexity. Li et al.'s research addresses the challenge of predicting multiple disease risks simultaneously, filling a significant gap in medical data analysis. While the study contributes to predicting multiple chronic diseases, its complexity makes it less likely to be applicable to predicting H1N1 and seasonal flu.

Ayachit et al. [1] and Inampudi et al. [8] worked on the same dataset and achieved similar accuracy scores, but with different approaches. Ayachit et al. [1] applied various algorithms to predict flu vaccination likelihood, while Inampudi et al. [8] used ANN and other models to predict H1N1 and seasonal flu vaccination status, demonstrating how different methods inform vaccination strategies. Ayachit et al. proposed a machine learning model using nine algorithms: MlBox, TPOT, Random Forest, MLP, Linear Regression, Decision Trees, Polynomial Feature, XGBoost, and CatBoost, with CatBoost performing the best, achieving an accuracy of 0.8617. The study builds on previous work using machine learning techniques and external factors for influenza prediction, showing that the GFT + CDC regression model is better for monitoring influenza activity than GFT and CDC models alone. The dataset used is from the National 2009 H1N1 Flu Survey (NHFS) provided by the U.S. CDC, processed with a simple imputer and label encoder. CatBoost achieved the highest probability score of 0.86 using ROC AUC as the metric. The study advanced flu vaccine prediction by effectively utilising ensemble learning and emphasised the importance of handling categorical features natively, noting CatBoost's advantage in this area. A strength of the study is the use of diverse machine learning models, ensuring robustness and comprehensive analysis. However, it did not comprehensively explore the impact of additional environmental factors. The study provides a robust predictive model, aligning with research on using machine learning techniques to predict flu vaccine uptake and highlighting the potential of ensemble learning methods. This ensemble approach can enhance predictive models for the 2009 H1N1 and seasonal flu datasets, contributing to more accurate predictions and better public health strategies.

Both Inampudi et al. [8] and Hussain and Fatima [6] address the same issue, using the same dataset and similar methodologies, which are highly relevant to this research. Both applied machine learning models to predict influenza outcomes. Inampudi et al. [8] focused on vaccination status using ANN, while Hussain and Fatima [6] evaluated various machine learning techniques to predict influenza outbreaks. Inampudi et al. [8] introduced a model to predict H1N1 and seasonal flu vaccination status using data from the National 2009 H1N1 Flu Survey, employing multiple machine learning algorithms, including Multiple Linear Regression, Support Vector Regression, Random Forest Regression, Logistic Regression, and Artificial Neural Networks with optimisers like Adam, RMSprop, and SGD. This literature builds on studies applying machine learning and statistical techniques to detect and analyse influenza, such as Mabrouk et al., who used methods like correlation dimension and largest Lyapunov exponent to detect H1N1 and distinguish between pandemic and classical influenza viruses. The dataset includes

over 53,000 responses, with 26,000 observations in the training set and the remainder used for testing. Inampudi et al. [8] utilised traditional evaluation techniques like accuracy and ROC curve analysis, finding that ANN was their best model, achieving 82.57% accuracy for H1N1 and 86.01% for seasonal flu vaccination prediction, outperforming other models such as SVM, Random Forest, and Logistic Regression. The study highlights the need for different predictive approaches for different populations, contrasting with other cited literature. Inampudi et al. [8] also emphasise the role of social media and large-scale surveys in understanding vaccination behaviour, citing literature on human behavioural patterns related to virus spread, including Lober, Roster, and Rodrigues [11], and Busse et al. [2], who provide insights closer to 2009. The strength of Inampudi et al.'s [8] research lies in the high accuracy of the ANN model in predicting vaccination status, demonstrating deep learning techniques' effectiveness. However, a noted weakness was the reliance on Twitter data, which introduces inconsistencies due to non-uniform data collection across different times and regions. This literature review situates the research within the context of machine learning applications for flu prediction, identifying gaps in data consistency and regional behavioural analysis relevant to this study. Inampudi et al.'s [8] research directly impacts this study's research and decision-making by enhancing the understanding of vaccination behaviour using machine learning models, potentially improving statistical accuracy, and exploring tools and methodologies not used by researchers in this literature review. Additionally, it is interesting to observe various technologies being used to address similar issues from the 2009 influenza and swine flu pandemic. This study does not avoid any specific dataset manipulation if other literature demonstrates its efficacy. The aim is to achieve the highest possible statistical accuracy score, adhering to proven methodologies and techniques where applicable. The study will prioritise implementing best practices identified in existing research to ensure the robustness and reliability of the predictive models.

Similarly, Hussain and Fatima's [6] study offers a different perspective. Their contribution lies in evaluating various machine learning techniques for predicting influenza outbreaks like H1N1 and B-Victorica through regression and classification. Specifically, they employed logistic regression, support vector machines (with polynomial and RBF kernels), naive Bayes, and Random Forest algorithms. The paper highlights that pandemic prediction is a key research area in mathematical biology, aimed at achieving optimal solutions quickly. Hussain and Fatima [6] explored related topics, citing a wide range of literature and tools used by other authors. The dataset includes monthly case numbers for various influenza types in Pakistan and China from 2009 to mid-2023, combining sentinel and non-sentinel surveillance data. Evaluation methods include accuracy, precision, recall, and F1 scores. Using a logistic classifier, they achieved an accuracy of 0.84, precision of 0.79, recall of 0.84, and F1 score of 0.79. The study found Random Forest to be the most accurate model, with 87% accuracy in predicting influenza outbreaks in China, concluding that the dataset is best suited for classification using Random Forest. The study also details common methods like logistic regression and SVM, while emphasising Random Forest's optimal performance. The strength of the study lies in Random Forest's high accuracy in predicting influenza outbreaks, with an average accuracy of 74% across all models, and a maximum of 87% with Random Forest. However, a weakness noted was the ineffectiveness of the SVM with RBF kernel. The literature relates to the research question by showing that machine learning models, particularly Random Forest, can accurately predict influenza outbreaks, highlighting gaps in SVM methods. This work contributes to the 2009 H1N1 and seasonal flu dataset by validating Random Forest's effectiveness, which can be applied to future influenza prediction models. Random Forest will be used in this study to compare accuracy scores and methods with Hussain and Fatima's [6] research. Both Inampudi et al.'s [8] and Hussain and Fatima's [6] studies demonstrate the applicability of machine learning in addressing influenza-related public health challenges.

2.4 Literature Gaps and Methodology Challenges

The literature gaps and methodology challenges were assessed in relation to known challenges within the research fields, based on pain points identified in Chapter 2 and the methodology outlined in Chapters 3 to Subchapter 3.10.

Obtaining comprehensive high-quality data on individuals' backgrounds, health conditions, and vaccination status is difficult. Identifying the most relevant features from the large set of background factors influencing vaccination likelihood can be complex. Choosing appropriate ML algorithms and techniques for specific task prediction requires careful consideration, and selecting model hyperparameters to achieve the best predictive results is crucial. The number of vaccinated and unvaccinated individuals may be skewed, leading to class imbalance, which can impact model performance. Ensuring models are interpretable is essential, especially when health-care data is needed to inform health interventions. Insights into underlying factors and patterns may not be easily noticed through viewing raw data alone. Additionally, ensuring the model generalises to different populations, geographic regions, and time periods presents a significant challenge.

Moreover, current influenza prediction models often rely on the single methodological approaches, limiting robustness and comparative insights. These studies focus on one type of methodology, which can introduce potential biases and skew findings. For example, studies such as Viana et al [20], Venna et al [19], and Yakovyna et al. [24] have predominantly employed single-method approaches, missing opportunities to explore alternative techniques in preprocessing, feature selection, model selection and evaluation.

3 Methodology

This methodology chapter provides a detailed overview of the data resources, preprocessing techniques, feature selection methods, and model evaluation strategies considered in this research. Building on theoretical frameworks and insights from the literature review in Chapter 2, it lays the groundwork for Chapter 4 and the subsequent analysis in Chapter 5. The objective is to detail dual systematic approaches, justifying the methodological choices critical for the research's success. The chapter structure is as follows: Subchapter 3.3 discusses the methodology for answering RQ1 and RQ2 in 5; Subchapter 3.2 provides an overview of dataset variables and their significance; Subchapter 2.4 addresses challenges highlighted in Chapter 2; Subchapter 3.5 outlines data preparation; Subchapter 3.6 explains feature selection and engineering techniques; Subchapter 3.7 details the preprocessing pipeline creation, ensuring consistency across training and testing, crucial for valid results in Chapter 5; Subchapter 3.8 discusses hyperparameter tuning; Subchapter 3.9 covers robustness through ensemble methods; Subchapter 3.10 covers model training and evaluation metrics, setting the stage for the performance analysis in Chapter 5. This methodology is essential for ensuring the research's reliability and accuracy, systematically addressing challenges identified in Chapter 2 and employing robust techniques in Chapter 4. The methodology chapter provides a strong foundation for the findings and interpretations in Chapter 5. Dual methodologies will be utilised at different stages of the methodological and implementation process, as stated in Subchapter 1.2 and Subchapter 1.1.

3.1 Data Resources

Data is vital for this research, offering key insights and forming analysis foundation. This study utilised data from the National 2009 H1N1 Flu Survey (NHFS) by the CDC, which aimed to assess H1N1 vaccination efforts among U.S. adults and children. Moreover, the survey, conducted through electronic media covered all 50 states and D.C. participants answered questions on: H1N1 and seasonal flu vaccination status, flu-related behaviours, vaccine safety perceptions, and medical history, including recent respiratory illness and pneumococcal vaccination. This data was crucial for understanding public awareness of flu vaccine safety and effectiveness, and reasons some avoided H1N1 and seasonal flu vaccinations. The results chapter covers the reasoning in Chapter 5. Furthermore, refer to Subchapter 6.4 for more information regarding ethical concerns and legal, social, economic, and professional issues.

3.2 Data Characteristics

Variable	**Description**	**Data Type**
Binary Variables: 0 = No; 1 = Yes		
behavioral_antiviral_meds	Has taken antiviral medications.	float64
behavioral_avoidance	Has avoided close contact with others with flu-like symptoms.	float64
behavioral_face_mask	Has bought a face mask.	float64
behavioral_wash_hands	Has frequently washed hands or used hand sanitizer.	float64
behavioral_large_gatherings	Has reduced time at large gatherings.	float64
behavioral_outside_home	Has reduced contact with people outside of own household.	float64
behavioral_touch_face	Has avoided touching eyes, nose, or mouth.	float64

doctor_recc_h1n1	H1N1 flu vaccine was recommended by doctor.	float64
doctor_recc_seasonal	Seasonal flu vaccine was recommended by doctor.	float64
chronic_med_condition	Has any chronic medical conditions (asthma, diabetes, etc.).	float64
child_under_6_months	Has regular close contact with a child under six months.	float64
health_worker	Is a healthcare worker.	float64
health_insurance	Has health insurance.	float64
Ordinal Variables		
h1n1_concern	Level of concern about the H1N1 flu. 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.	float64
h1n1_knowledge	Level of knowledge about H1N1 flu. 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.	float64
opinion_h1n1_vacc_effective	Respondent's opinion about H1N1 vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.	float64
opinion_h1n1_risk	Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.	float64
opinion_h1n1_sick_from_vacc	Respondent's worry of getting sick from taking H1N1 vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.	float64
opinion_seas_vacc_effective	Respondent's opinion about seasonal flu vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.	float64
opinion_seas_risk	Respondent's opinion about risk of getting sick with seasonal flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.	float64
opinion_seas_sick_from_vacc	Respondent's worry of getting sick from taking seasonal flu vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.	float64
Other Variables		
age_group	Age group of respondent.	object
education	Self-reported education level.	object

race	Race of respondent.	object
sex	Sex of respondent.	object
income_poverty	Household annual income of respondent with respect to 2008 Census poverty thresholds.	object
marital_status	Marital status of respondent.	object
rent_or_own	Housing situation of respondent.	object
employment_status	Employment status of respondent.	object
hhs_geo_region	Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services.	object
census_msa	Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.	object
household_adults	Number of other adults in household, top-coded to 3.	float64
household_children	Number of children in household, top-coded to 3.	float64
employment_industry	Type of industry respondent is employed in.	object
employment_occupation	Type of occupation of respondent.	object

Table 2: Summary of Variables Used in H1N1 & SF Dataset [26]

3.3 Multivariate Logistic Regression

Research Question 1 and 2 both utilised Multivariate logistic regression. MLR analyses relationships between multiple independent variables and binary outcome, providing separate coefficients for each variable. Interactions terms capture how the effect of one variable changes in the presence of another. Moreover, MLR is necessary when variables may influence each other, and interaction terms help reveal combined effects that are not apparent when variables are considered independently. Moreover, the implementation involves fitting logistic model where each variable's coefficient represents its effect on the outcome while controlling for other variables. Interaction terms are included to understand how variables jointly influence the outcome. The strength of this methodology is it provides a nuanced understanding of multiple variables interact and affect the outcome. However, it can become complex and difficult to interpret with many variables and interaction terms. Compared to confusion matrix, summarising prediction accuracy, MLR gives insights to underlying relationships between variables and outcome. The confusion matrix is simpler, but less informative regarding how different variables contribute to predictions.

3.4 Methodology Comparison

Stage	Method A	Justification	Method B	Justification
Pre-processing	Duplicate Removal: ‘pandas‘ (‘duplicated()‘, ‘drop_duplicates(	Prevented skewed predictions by ensuring the model trained only on unique and valid data points, crucial for accuracy.	Standardisation: StandardScaler in ‘Pipeline‘	Standardised numerical features to ensure they contributed equally to the model, addressing issues related to differing scales across the dataset.
	Missing Data: Iterative Imputation with Label Encoding/Decoding	Handled missing data while preserving complex interrelationships between variables, essential for accurate downstream analysis.	Missing Data: Mean/Mode Imputation, Downsampling for prototyping	Used median for numerical data to mitigate the impact of outliers, and mode for categorical data to retain the most common category, ensuring statistical integrity. Downsampling performed for faster prototyping.
	Outliers: Retained after initial removal lowered accuracy	Retaining outliers maintained the model’s predictive accuracy, as removing them degraded generalisability and performance.		
Feature Selection	EDA - Aggregated behavioural (”cleanliness”) and opinion metrics	Aggregated related features to simplify the model and reduce redundancy, improving clarity in the interpretation of factors influencing vaccination intent.	SelectKBest (chi-squared, top 10 features)	Applied SelectKBest to tackle high-dimensionality, specifically choosing the chi-squared test to identify and retain the most statistically significant features.

Stage	Method A	Justification	Method B	Justification
Pipeline	No formal pipeline, direct preprocessing	A direct approach was sufficient for this method due to the specific preprocessing steps applied.	'Pipeline' with StandardScaler and RandomForestClassifier	Integrated preprocessing and model training into a unified workflow, ensuring consistency in data transformation and preventing data leakage between training and testing.
Hyperparameter Tuning	GridSearchCV, RandomizedSearchCV for RandomForest, XGBoost; BinaryRelevance for multi-label classification	Used GridSearchCV and RandomizedSearchCV to exhaustively search for the optimal hyperparameters, ensuring the model's performance was finely tuned for multi-label classification tasks. BinaryRelevance was used to handle the complexity of multi-label problems effectively.	HalvingGridSear for RandomForest, XGBoost, Gradient Boosting	Chose HalvingGridSearchCV for its ability to efficiently handle large hyperparameter spaces, focusing computational resources on the most promising configurations.
Ensemble	AdaBoost, Bagging, Extra Trees, Gradient Boosting, Random Forest; ShuffleSplit cross-validation (10 splits)	Combined various ensemble methods to enhance robustness and reduce the likelihood of overfitting, ensuring the model could generalise well across different data subsets.	Combined RandomForest, XGBoost, Gradient Boosting in preprocessing pipeline	Used an ensemble of diverse models within the pipeline to capture complex patterns in the data, improving the overall predictive accuracy and robustness of the model.

Stage	Method A	Justification	Method B	Justification
Model Training & Evaluation	GridSearchCV; evaluated with ROC-AUC, macro average, cross-validation accuracy	Employed GridSearchCV to optimise key model parameters, ensuring high performance. Evaluation metrics like ROC-AUC and cross-validation accuracy provided a comprehensive assessment of the model's predictive capabilities.	Stratified k-fold cross-validation, ColumnTransformer for scaling/encoding, accuracy-focused evaluation	Utilised stratified k-fold cross-validation to maintain class balance across folds, crucial for reliable model validation. Prioritised accuracy for precise classification outcomes.

Table 3: Comparison of Methods A and B Across Different Stages. Chapter 4 provides a detailed implementation and justification of Methodologies A and B

3.5 Preprocessing

3.5.1 Removing Duplicated Entries

Handling and removal of duplicates is a core fundamental of preprocessing the data. [39] It involves identifying and eliminating repeated records within the data to ensure quality and integrity. This methodology aims to enhance the accuracy and reliability of subsequent analysis by removing redundant information. The primary contribution is the reduction of noise that can skew results and lead to incorrect conclusions. The process of handling duplicates begins with entries based on specific criteria, such as unique identifiers or a combination of attributes. For example, techniques such as the 'drop_duplicates' function in the pandas library for Python are common. The parameters can be set to specify which columns to consider and whether to keep first, last, or none of the duplicates. This approach ensures a clean dataset ready for analysis.

This approach is supported by Du et al. [4] and Venna et al. [19], who emphasise the importance of data integrity for accurate influenza predictions. Removing duplicates is crucial for reducing noise and enhancing the reliability of predictive models, as noise can significantly skew results, leading to inaccurate conclusions. [40] However, the process may inadvertently remove non-exact duplicates that contain variations but are still relevant. Furthermore, compared to other complex methods (e.g. fuzzy matching), the straightforward approach of using unique identifiers is faster and simpler. While fuzzy is able to identify near-duplicates, it is computationally intensive and may not be necessary for well-structured data sets. [41] Ensuring data accuracy and consistency remains an ongoing challenge, where it must be meticulously handled. The proceeding steps involve addressing outliers and handling missing data.

3.5.2 Handling Missing Data

Iterative Imputer

The iterative imputer is a sophisticated imputation technique for handling missing data by modelling each feature with missing values as a function of other features. [48] This methodology iteratively eliminates missing values by fitting regressors in a round-robin fashion, treating each feature as a target. [49] It then refines these estimates through multiple iterations which improves a model's accuracy. Moreover, this methodology is a robust way to handle complex data structures with interdependencies among features, offering better imputation performance compared to simpler methods such as median or mode imputation. Marquez et al. [12] demonstrated the effectiveness of advanced imputation techniques in managing complex data sets, underscoring importance of this approach in data integrity.

The imputer in scikit-learn operates by designating one feature column as the target, and the others as predictors, fitting regressor on these predictors it then estimates missing values. This imputation technique typically uses BayesianRidge as the default estimator. [50] The parameters included are 'max_iter' which controls the number of iterations, and 'initial_stategy' defines how initial missing values are estimated. The imputer can also use a specified number of nearest features to expedite computation which makes it adaptable to various complexities.

One of the key strengths of utilising the iterative imputation is the ability to handle complex relationships between features and its flexibility in using various regression models for imputation. This leads to more accurate and reliable estimates of missing data. However, iterative processes can be a lengthy process, especially with large datasets. Additionally, the method is considered experimental, with potential changes in its implementation and behaviour. [51] Compared to a simple imputer, which fills missing values with a single statistic (e.g. mean or median), the iterative imputer offers a more sophisticated approach by modelling each feature iteratively. While the simple imputer is faster and less complex, it may not capture the underlying data structures as effectively as the iterative imputer. Thus, the iterative imputer is preferable for datasets with intricate relationships among features.

3.5.3 Addressing Outliers

The primary concept of addressing outliers is to further enhance the data quality from Subchapter 3.5.1, by identifying and handling outliers to prevent them from skewing analytical results which is crucial for this study's research questions. Outliers significantly impact the performance of machine learning models, leading to potential biased or inaccurate outcomes. [42] Through effectively managing outliers, the data becomes more representative of underlying trends. Addressing outliers involve statistical techniques to detect anomalies and apply corrective measures such as removal or transformation. Outlier detection and fixation can be implemented through several steps. Firstly, identifying outliers using statistical methods such as Z-score, which measures how far a data point is from the mean in terms of standard deviations. [43] Secondly, visualising the data distributions using box plots or scatter plots to pin point said anomalies. [44] Lastly, applying techniques such as data transformation (e.g. log or square root) or imputation to adjust the outliers, or removal if they are deemed incorrect inputs. [45]

A strength is handling outliers does improve data quality and model performance by ensuring the data accurately reflects the populations, enabling it to achieving more robustness and reliable statistical analysis. [46] However, the outlier detection methods may sometimes mistakenly classify valid data points as outliers, leading to potential data loss. Additionally, it may be computationally intensive for larger data sets. Furthermore, compared to robust scaling, which

reduces the impact of outliers by using interquartile range (IQR), traditional outlier handling through removal or transformation is more straightforward. [47] However, robust scaling preserves all the data points, which might be beneficial for smaller data sets.

This approach aligns with the findings of Sultana and Sharma [15], who demonstrated that unaddressed outliers could significantly skew model predictions, leading to unreliable results. Managing outliers through statistics techniques like Z-score helps ensure the data reflecting true trends. However, this method also risks misclassifying significant anomalies as outliers. A discussion of how the trade-off between removing outliers and preservation of critical data points was managed is crucial.

Median/Mode Imputation

The primary contribution of the mean/mode simple imputer methodology is to address missing values in the data by substituting them with the mean (for numerical features) or mode (for categorical features). [52] This technique is specifically beneficial for maintaining the integrity of the data. This also avoids the exclusion of incomplete input records which is crucial for data that contains moderate amount levels of missing values (or NaN values). The mean/mode imputation works by firstly calculating the mean of all available values in a numerical feature or mode for a categorical feature. The missing values in the respective feature are then replaced with these calculated values. This can be efficiently implemented by using Python's Scikit-learn's SimpleImputer, where the imputation strategy is set as 'mean' for numerical features' and 'most_frequent' for categorical features. [53] For example, within a data set that contains missing values in a 'income' column, the mean income is computed and used in order to replace missing entries.

The strength of utilising the mean/mode imputation is that its simplicity, but mostly being computationally efficient, is suitable for large-scale data sets. It maintains all records in the data sets, which is crucial for preserving the data sets completeness. However, as Indhumathi et al. [7] highlights, while simple imputation methods such as mean/mode are efficient, they can distort the original data distribution by artificially reducing variance. This leads to potential bias in statistical estimates. It also fails to capture any underlying relationships between features, overlooking complex data patterns. In comparison to MICE, the mean/mode imputation is significantly faster and easier to implement. While MICE can model complex relationships between missing and observed data, it is still computationally intensive and complex to apply. [54] Furthermore, in contrast to the KNN imputation, KNN can model more complex relationships by considering similarity between data points, but again, is also computationally more intensive and does not scale well with large data sets.

3.6 Feature Selection & Engineering

3.6.1 Exploratory Data Analysis

Exploratory data analysis responsibility is to identify and create most relevant features that significantly improve a model's performance. [55] EDA provides insights into which features are most predictive and helps in transforming raw data into informative features. This involves both visual and statistical methods to highlight importance and relationships of features. [56] EDA emphasises the creation of features through techniques such as combining existing features or creating interaction terms in order to enhance predictive power. EDA involves several steps such as the initial data exploration in order to understand the data set, identification of key features through correlation matrices, and statistical tests. Visualisations such as scatter and pair plots are utilised to explore relationships between features. The feature engineering aspect

is involved by creating new features by combining existing ones, such as interaction terms or aggregating related features. Advanced techniques such as PCA are employed to reduce dimensionality while retaining essential information. [57]

The strength of utilising EDA lies in its ability to identify most relevant features and creation of new, more informative features, that aid in improving the model's accuracy and interpretability. It helps in simplifying complex data sets by reducing dimensionality through the use of feature selection and engineering. However, as noted by Du et al. [4] and Venna et al. [19] its manual nature can introduce subjective bias, affecting reproducibility. Moreover, EDA is very resource-intensive, requiring computational power and time in order to understand the data characteristics and testing various manipulations of the raw data may be manipulated, especially for large data sets. Feature engineering might introduce bias if not done carefully and can then lead to overfitting if too many features are created without proper validation. In comparison to automated feature selection such as Recursive Feature Elimination, EDA provides a more intuitive understanding of the data and allows for manual feature creation based on domain knowledge. [58] Furthermore, while RFE systematically selects features by eliminating the least important ones, EDA's visual and statistical approach offers a deeper insight into feature relationships. EDA is preferred when domain knowledge and intuition plays a significant part in feature engineering and selection.

3.6.2 SelectKBest

SelectKBest feature selection method is its ability to identify most relevant features for improving model performance by selecting the top K features based on statistical criterion. [59] This method enhances a model's accuracy by reducing the dimensionality of the data, thus focusing on the very most informative variables. SelectKBest introduces an effective way to handle large datasets by filtering out less significant features. [60] It leverages statistical tests, such as chi-square (and/or ANOVA), to rank features according to their relevance. [61] This feature selection method works by evaluating each feature individually and selecting the top K features with the highest scores where K is the number of features selected as a parameter. Parameters such as number of features (K) and scoring functions (e.g. chi-square, f-classif) are pivotal in its implementation. This method ranks features based on their scores and retains those with highest scores for model training. This process is straight forward and can easily be integrated into machine learning pipelines.

The strength of this methodology is that it is simple to implement and computationally efficient, making it suitable for high-dimension datasets such as the 2009 H1N1 and seasonal flu data set. It provides a clear and interpretable ranking of features based on its statistical significance. However, as Sultana and Sharma [15] demonstrated, while feature selection methods such as SelectKBest effectively filter out less informative features, they may also overlook feature interactions that would potentially be important to the model, missing out on synergistic effects between features. It also relies heavily on the choice of scoring function, where the function itself does not always capture the most predictive features. Compared to PCA, SelectKBest directly identifies most relevant features based on their individual statistical significance. [61] Meanwhile, PCA reduced dimensionality by transforming features into principal components that capture maximum variance, it can be computationally intensive and may result in components that are harder to interpret. SelectKBest is simpler and faster to implement, making it more practical for large data sets where interpretability of features is essential such as the data set this study is using.

3.7 Pipeline with StandardScaler

Pipeline with StandardScaler in preprocessing is to streamline the workflow by chaining multiple processing steps into a single, cohesive unit. [62] This ensures each step, including scaling features to zero mean and unit variance using StandardScaler is applied consistency during both training and testing phases. [63] It is beneficial as it prevents data leakage and improves a model's robustness. The pipeline in scikit-learn is implemented by defining a sequence of steps where each step is a tuple containing an name and a transformer or an estimator. For instance, a pipeline with StandardScaler, followed by a regression model could be defined as Pipeline([('scaler', StandardScaler()), ('regressor', Ridge())]). This ensures during fitting, the scaler is applied to the training data passing the scales data to the regressor. Parameters such as with_mean=True and with_std=True in StandardScaler are chosen to ensure the data is centred and scaled correctly.

This approach aligns with the practises observed in Venkatramanan et al. [18], who utilised standardised pipelines to maintain consistency and ensure process integrity. The strength of using Pipeline with StandardScaler, along other steps include code readability and reduced risk of data leakage by ensuring consistency application of processing steps across both training and testing sets. Moreover, it simplifies hyper parameter tuning by allowing the use of tools such as GridSearchCV to optimise parameters across the entire pipeline. However, a notable weakness is that StandardScaler can be sensitive to outliers, which distorts the scaling process. This sensitivity to outliers was also noted by Venkatramanan et al. [18] suggesting that the choice of scaling method should be carefully handled. Additionally, constructing the pipeline itself requires careful handling to ensure data transformations do not inadvertently introduce biases or information leakage. In comparison to the MinMaxScaler, the StandardScaler is more robust in ensuring features have zero mean and unit variance. [64] This is particularly beneficial for algorithms assuming normally distributed data. However, MinMaxScaler scales features to a fixed range, which can be advantageous when dealing with bounded data but might not handle outliers as well. StandardScaler is preferred in scenarios requiring normalisation for linear models, while MinMaxScaler can be better for ML algorithms such as neural networks that benefit from bounded input ranges.

3.8 Hyper-parameter Tuning

The HalvingGridSearchCV method is to efficiently tune hyperparameters by progressively reducing the number of candidate hyperparameters as the search progresses, based on performance. [65] This method significantly reduces computational cost compared to exhaustive grid searches such as HalvingGridSearchCV. It introduces an innovative approach to balance exploration and exploitation of hyperparameter space. HalvingGridSearchCV provides more accurate and faster results for hyperparameter optimization in large data sets such as the data set this study is using. [66] HalvingGridSearchCV starts by evaluating a large pool of hyperparameter combinations on a small portion of the data. It iteratively increases the amount of data used for evaluation while halving the number of candidate hyperparameters based on their performance. Parameters such as 'factor', 'resource', and 'aggressiveness' are vital for tuning this method. Lastly, it involves a detailed evaluation of the remaining candidates on the entire data set to determine optimal hyperparameters.

Its key strength is that it is highly efficient. It drastically reduces computational time and resources significantly compared to traditional grid search. It ensures thorough exploration of the hyper parameter space while maintaining computational feasibility and time. However, HalvingGridSearchCV may prematurely eliminate a set of hyperparameters that could perform well with more data. It also requires careful tuning of the initial settings to avoid biassed

output results. Compared to RandomizedSearchCV, HalvingGridSearchCV systematically narrows down hyperparameter choices, making it more efficient in all areas (i.e. computationally, cost). [65] While RandomizedSearchCV explores hyperparameter space randomly, it can miss optimal regions, which in turn makes HalvingGridSearchCV more reliable for large datasets. However, RandomizedSearchCV is simpler to implement but is less structured in its search approach. This approach aligns with findings of Singh and Mittal, who optimised models using advanced hyper parameter tuning techniques similar to HalvingGridSearchCV, demonstrating its efficiency in exploring hyper parameter space. While HalvingGridSearchCV balances thoroughness and computationally efficiency, the potential risk of prematurely eliminating promising hyper parameters can be mitigated by careful tuning of initial grids and continuous monitoring of intermediate results. [67]

3.9 Ensemble

The contribution of ensemble methods is to enhance predictive performance by combining multiple models to reduce: variance, bias, or improve predictions. Ensembles (bagging, boosting, and stacking) leverage the strengths of multiple algorithms to produce better results compared to individual models. [68]These methods have been shown to improve accuracy throughout studies, and have several other benefits (i.e. stability and robustness of the model). [69] Numerical outcomes often demonstrate significant performance improvements such as lower error and higher classification accuracy. Ensemble works by aggregating predictions from multiple models, either through averaging (bagging), weighted voting (boosting), or combining diverse model types (stacking). [70] For instance, in bagging, models are trained on bootstrapped subsets of the data, while boosting adjusts models weights iteratively to focus on misclassified instances. Moreover, the parameters such as the number of base learners and learning rate in boosting are crucial for optimisation. The architecture typically involves parallel or sequential training of models followed by aggregation of their outputs.

The strength of ensemble is that it increases model accuracy and generalisability by reducing overfitting and leveraging models. It also provides more robustness against noisy data and outliers that may have been introduced. However, the methodology itself is computationally intensive due to the need to train on multiple models. Additionally, they can be more complex to interpret compared to single methods. Compared to a single decision tree, ensemble methods such as random forests provide higher accuracy and better generalisation. While a decision tree is simpler and faster to train, it is prone to overfitting, making ensemble methods an ideal choice for complex data sets such as the one being handled in this study. This approach aligns with findings by Marquez et al. [12] and Elbasi et al. [5] who demonstrated that ensemble methods significantly enhance predictive accuracy by combining multiple models.

3.10 Model Training & Evaluation

The purpose of model training and evaluation is to systematically develop and assess machine learning models to predict outcomes accurately. [71] This methodology introduces cross-validation techniques and performance metrics in order to ensure the robustness discussed earlier in previous chapters. For example, the use of k-fold-cross-validation provides comprehensive measures of model performance across different subsets of data. [72] This approach with the emphasis on cross-validation by Indhumathi et al. [7] and Ayachit et al. [1], who highlighted its effectiveness in preventing overfitting and ensuring robust model evaluation. These methods yield both numerical and qualitative insights into the model effectiveness and generalisability. Model training involves splitting the data into training and testing sets, then using the training set to the model. K-fold-cross-validation is implemented to mitigate overfitting by training and validating the model on different data folds. [74] Parameters such as the number of folds (k)

and the choice of evaluation metrics (e.g. accuracy, precision, recall) are critical in this process. [75]The architecture of this method typically includes iterative training and validation steps to fine tune model parameters for optimal performance.

A strength of utilising cross-validation for model training and evaluation is that it ensures a model's performance is evaluated comprehensively. [73] Additionally, this methodology is effective in preventing overfitting and provides a realistic estimate of the model performance on unseen data. However, the methodology can be quite computationally intensive for large datasets due to the repeated training and validation processes. Furthermore, it may not capture temporal dependencies if the data is not shuffled appropriately.

Methodology Discussion

The detailed methodology outlined in this chapter supports the implementation of predictive models discussed in the following chapters, where results and their implementation will be thoroughly analysed.

4 Implementation

The implementation chapter describes practical execution of preprocessing, feature selection, and model training strategies defined in Chapter 3. The objective is to apply dual methodological approach to most relevant features, and prepare the model for training and tuning for subsequent analysis of result chapter. The purpose of this chapter is to document specific steps taken to prepare data, select features, and train models, ensuring accuracy in predicting vaccination behaviours. The chapter structure is as follows: Subchapter 3.3 discusses implementation from subchapter 3.3 to answer RQ1 and RQ2; Subchapter 4.2 details preprocessing; Subchapter 4.3 entails description of feature selection & engineering; Subchapter 4.4 explains the use of standarised pipeline ensuring consistency; Subchapter 4.5 outlines methods used for tuning model parameters; Subchapter 4.6 discusses implementation of ensemble techniques; Subchapter 4.7 provides an overview of training, cross-validation, and performance metrics. The implementation steps outlined here are critical for achieving reliable results presented in the following chapter.

4.1 Multivariate Logistic Regression

Research Question 1 was answered by using MLR and interaction terms. MLR was employed to examine influence of demographic factors (i.e. age, education, income) on perceptinos of vaccine effectiveness and risk for H1N1 and seasonal flu. Dependent variables representing perceptions of vaccine effectiveness and risk were converted into binary outcomes. Responses were classified responses of 4 or higher as "effective" or "risky", and all others as "not effective" or "not risky". Furthermore, using one-hot encoding, categorical demographic variables were used to prepare them for regression analysis. This process transformed demographic categories into set of binary variables, allowing model to assess impact of each demographic factor independently. The logistic regression models for each of the outcome variables were fitted: H1N1 vaccine effectiveness, H1N1 vaccine risk, seasonal flu vaccine effectiveness, and seasonal flu vaccine risk. The independent variables in these models were encoded demographic factors. After fitting the models, the coefficients were extracted, which represent the effective size of each demographic factor on the likelihood of perceiving vaccine as effective or risky. These coefficients were organised into dataframes for clarity. Chapter 6 outlines the advantages and disadvantages. Results are in Subchapter 5.1.

Research Question 2 was answered by utilising MLR and interaction terms. MLR was employed to assess impact of specific preventative behaviours (hand washing, mask wearing, and antiviral medication use), on likelihood of receiving H1N1 and seasonal flu vaccines. Initially, a correlation matrix (before EDA), Figure 3 was generated to examine linear relationships between behaviours and vaccination outcomes. The observed correlations were weak, indicating individual behaviours had limited predictive power when considered in isolation. To address interaction effects, interaction terms were utilised by calculating the product of relevant behaviours (i.e. interaction between hand washing and mask wearing, and between hand washing and antiviral medication use). These interactions terms were incorporated to capture combined effects of behaviours that might not be apparent when evaluated separately. Subsequently, MLR was applied to model the probability of vaccination as a function of these behaviours and their interactions. Each behaviour and interaction term was assigned a coefficient within the model. These coefficient quantify effect size and direction of each behaviour on vaccination likelihood, controlling for the influence of other variables. This approach enabled detailed quantification of independent and interactions effects of preventative behaviours on vaccination outcomes, providing a nuanced understanding of the factors influencing vaccination decisions. Chapter 6 outlines the advantages and disadvantages. Results are in Subchapter 5.2.

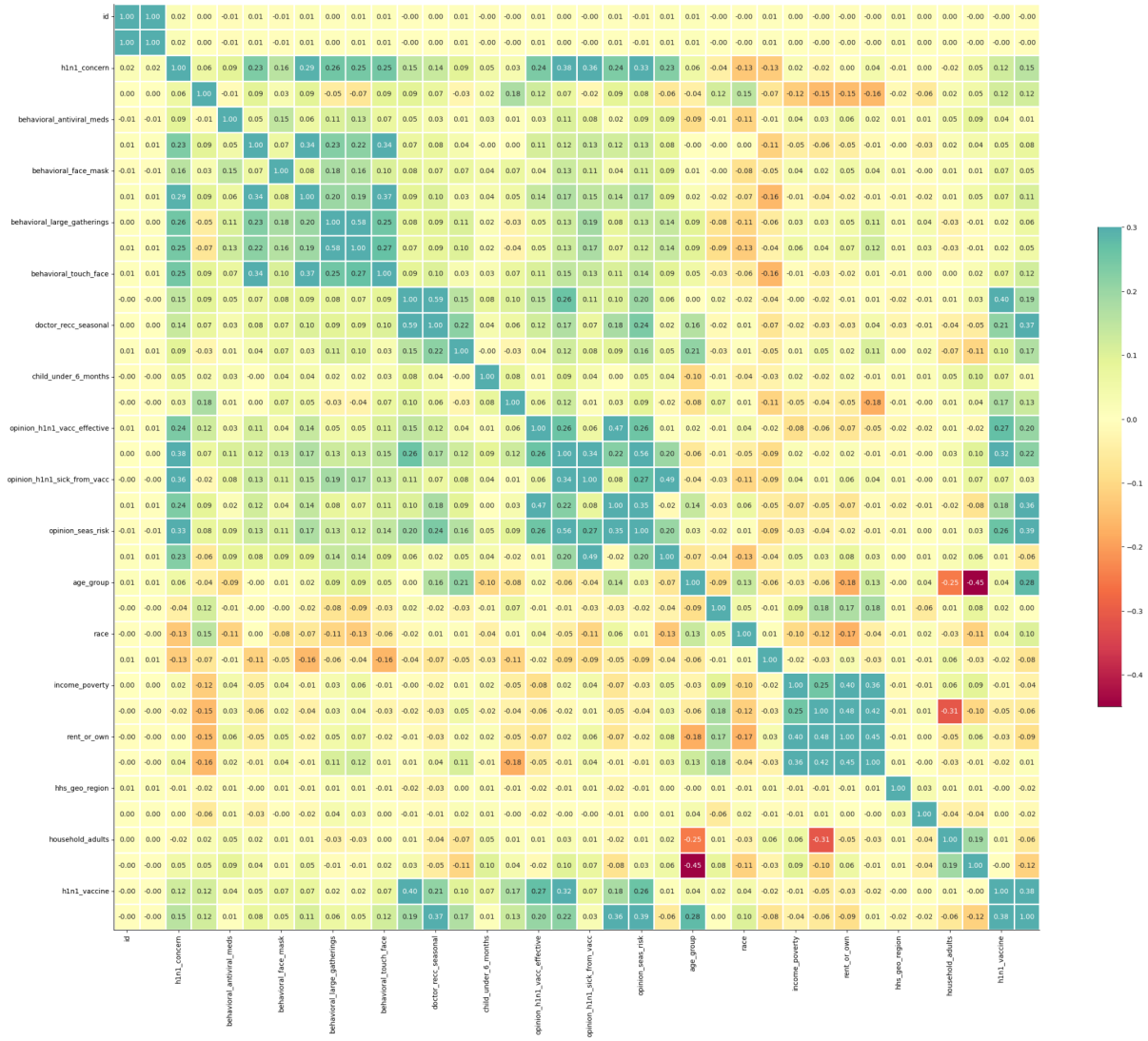

Figure 3: Before EDA: Correlation matrix of individual variables on H1N1 & Seasonal Flu (concerns, vaccination behaviours, demographics, and vaccine opinions)

4.2 Preprocessing

Research Question 3 was answered by:

4.2.1 Removing Duplicate Entries

The research began by first identifying and removing duplicated entries for method A. This step was necessary after observing the potential for redundant data which skews model predictions. Removing these duplicated entries ensures that the model learns from unique and valid data points, improving its accuracy. The implementation uses pandas' duplicated() function to detect duplicates, followed by drop_duplicates to remove them.

This step was executed without any additional data manipulation or complex techniques due to the reliability of the pandas library. Conducting a similar search for duplicates in the second data set version, data_m1, would have been redundant and unnecessary. Consequently, the study confirmed that there were no duplicated entries in the data set. Chapter 6 outlines

the advantages and disadvantages.

4.2.2 Handling Missing Data

Iterative Imputation

The decision to use iterative imputation stemmed from the need to handle missing data in a manner that preserves the intricate relationship between variables, and the need to handle categorical data without losing the inherent variability and patterns present in the data. Mode imputation would simply replace missing values with the most frequent category introducing bias within the data. This would not be a representative of the overall data distribution. The initial decision to encode categorical variables using LabelEncoder was based on the necessity to transform non-numeric data into numeric data so that Iterative Imputer could process. Categorical variables were converted into integer values. This was to ensure the unique categories within each variable were preserved for subsequent decoding. Chapter 6 outlines the advantages and disadvantages.

Unlike mode imputation which oversimplifies data by repeatedly imputing the same value, Iterative Imputation uses a sophisticated approach that reflects the complexity of the data. This was important for variables that may be highly interrelated. Following the imputation, the encoded categorical variables were decoded back to their original form. This decision was made to ensure the interoperability of the data, which was crucial to understanding the results of subsequent analysis for the results. Chapter 6 outlines the advantages and disadvantages.

Mean & Mode Imputation

This approach uses median and mode imputation to handle missing data maintaining the dataset's statistical integrity. Median is chosen for numerical to mitigate outlier effects, whilst mode is chosen for categorical data to preserve the most common category. The function "fill_missing_values" iterates over each column, imputing missing values with the median for numerical and mode for categorical ones. Downsampling was done for computational efficiency. This allows tasks to be performed at a much quicker rate. This is referred to as prototyping. Chapter 6 outlines the advantages and disadvantages.

4.2.3 Addressing Outliers

In this study, the decision to retain outliers was based on their critical role in maintaining the accuracy score. The research applied removing outliers through the use of box-plots, but noticed a poor generalisability and overall accuracy score when outliers were removed. Extreme cases above 95th percentile, and below the 5th percentile were removed. This decision did not go on. Subchapter 6.3 explains the reasoning. Chapter 6 outlines the advantages and disadvantages.

4.3 Feature Selection Engineering

4.3.1 Exploratory Data Analysis

Method A deployed an EDA approach. This approach involved aggregating related features to create comprehensive metrics to simplify the analysis. Specifically, features relating to behavioural conditions were combined to form a "cleanliness" metric, and opinion-related features were aggregated into "opinion" metrics for both H1N1 and seasonal flu vaccines. This aggregation was needed in order to interpret the influence of behavioural and opinion factors on vaccination intent, and to improve models prediction ability. The implementation utilised python's pandas library to aggregate several behavioural and opinion-based features. Behavioural features

such as "behavioral_antiviral_meds", "behavioral_avoidance", "behavioral_face_mask", "behavioral_wash_hands", "behavioral_large_gatherings", "behavioral_outside_home", and also "behavioral_touch_face" were summed to form a new "cleanliness" feature. Similarly, features relating to opinions on vaccine effectiveness, risk, and sickness (for both H1N1 and seasonal flu) were aggregated into opinion and opinion_h1n1 and opinion_seasonal metrics. These new features provided a holistic view of individual attitudes and behaviours, directly addressing complexities of factors influencing vaccination behaviour. The decision to aggregate these features were driven by observations that individual behavioural and opinion features were highly correlated and could be grouped together to reduce redundancy and improve clarity of the model's predictions. This was determined by analysing correlation matrix and identifying clusters of features that represented similar concepts - represented in Figure 3 (before EDA), and Figure 4 (after EDA). Chapter 6 outlines the advantages and disadvantages.

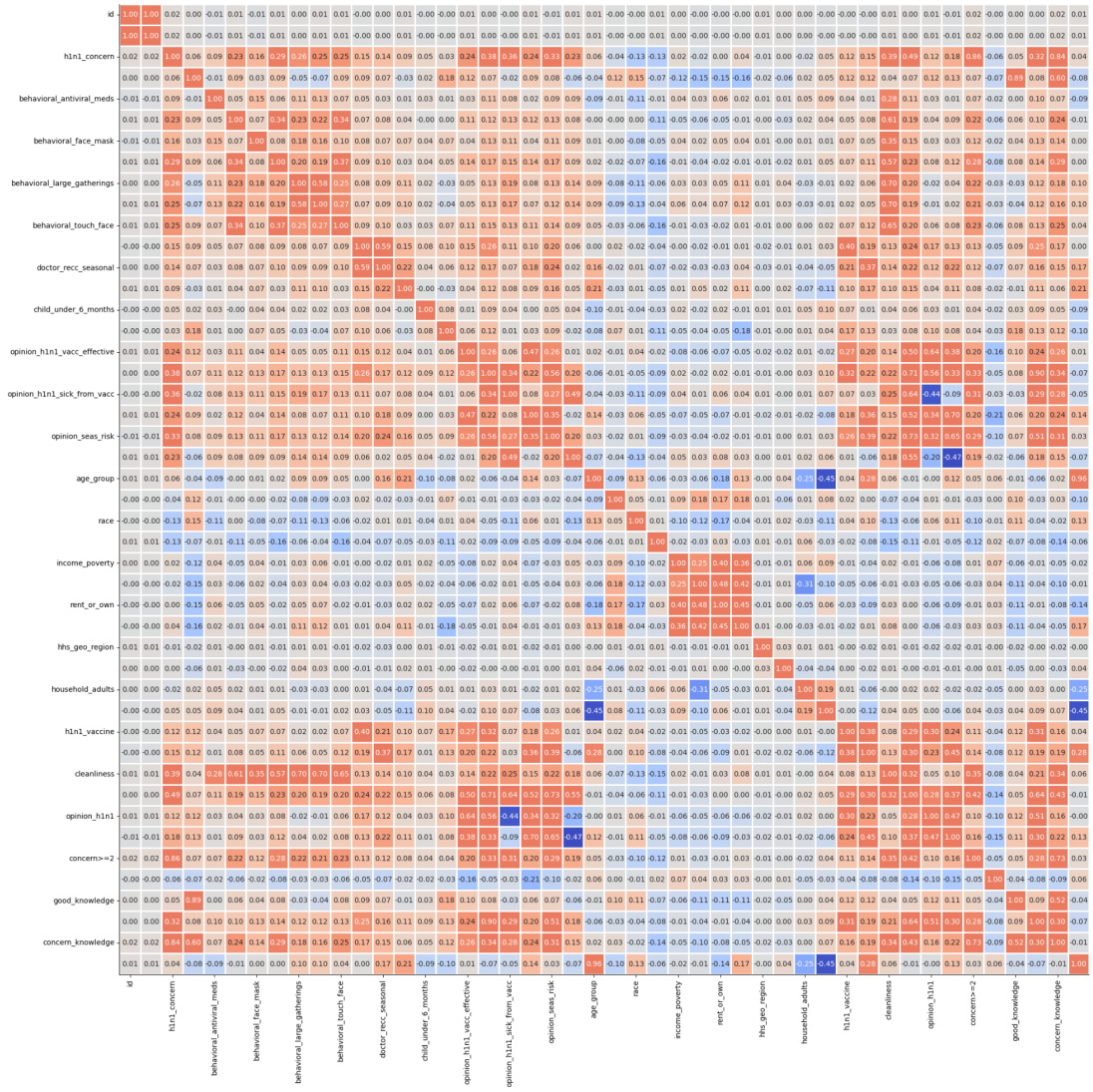

Figure 4: After EDA: Correlation matrix of individual variables on H1N1 & Seasonal Flu (concerns, vaccination behaviours, demographics, and vaccine opinions)

4.3.2 SelectKBest

Method B deployed a SelectKBest algorithmic approach. This unique approach of using SelectKBest for separate feature selection for h1n1 and seasonal flu prediction ensures the most statistically significant features are retained. This effectively addresses the high dimensionality problem within the data set. This method was vital for improving model accuracy by narrowing down feature sets to those relevant for predicting each type of vaccination, aligning with the research questions in Subchapter 1.2. SelectKBest was implemented as the feature selection method after evaluating the nature of the data set, which included a mix of categorical and numerical features. The chi-squared (chi2) test was selected because it is well suited for categorical data, which formed a significant portion of the data set. This specific choice to separate the feature selection process was based on the hypothesis that different factors might influence the decision to receive each vaccine. This led to independent application of SelectKBest for each target variable, ensuring the most 10 features were identified and retained for further modelling. The decision to use k=10 was based on the need to maintain model simplicity while retaining enough features to capture complex patterns. Chapter 6 outlines the advantages and disadvantages.

4.4 Pipeline with StandardScaler

Additionally, Method B utilised an additional approach. This approach focuses on utilising a pipeline with standardscaler to preprocess the data before applying a machine learning model. The concept is to ensure all numerical features are standardised, which is crucial because the data contained features with varying scales. Standardisation ensures that each feature contributes equally to the model's learning process, addressing the challenge of handling diverse data types. It was clear that consistency in preprocessing and model application was critical to avoid issues with data leakage or inconsistent feature transformation between training and testing phases. To address this, the decision was made to implement a pipeline using sklearn.pipeline.Pipeline, which allows for a seamless combination of preprocessing steps and model training with a single, unified workflow. The pipeline starts with the standardscaler to standardise all numerical features. This ensures each feature's mean is set to 0 and the standard deviation to 1, making them comparable across the data set. Following the scaling step, the standardised data is passed to a RandomForestClassifier, alongside other models, due to its robustness and ability to handle a mix of categorical and numerical data. The pipeline structure was specifically chosen because it automatically applies the same scaling to any data that passes through it, ensuring the models receive the data in a consistent format. This is important to models such as randomforest, which can be sensitive to variations in input scales. The decision was influenced by the need to ensure all features were treated equally during model training, preventing any feature from disproportionately influencing the model due to its scale. Chapter 6 outlines the advantages and disadvantages.

4.5 Hyper parameter Tuning

Method A involves hyper parameter tuning and experimenting with different methods. This approach is built around hyperparameter tuning using multiple methodologies - GridSearchCV, and RandomisedSearchCV, applied to both randomforest and XGBoost classifiers, wrapped within a BinaryRelevance multi-label classification framework. This decision was motivated by the need to explore a diverse range of hyperparameters across different classifiers to optimise the performance of models in predicting multiple vaccination labels, which is central to the research questions in Subchapter 1.2. the goal was to determine the most effective classifier and hyperparameter combination for accurately classifying vaccination intents. The implementation consisted of several steps:

1. RandomForestClassifier with GridSearchCV: the decision to use GridSearchCV with randomforest was driven by its ability to exhaustively search across a specified hyperparameter grid, ensuring the optimal combination of hyperparameters is identified. Given the ensemble nature of randomforest, which mitigates overfitting by averaging multiple decision trees. It was a natural choice for the base model. The grid included parameters such as the number of trees, n_estimators, splitting criterion, criterion, the number of features considered at each split max_features, tree depth max_depth, and the minimum samples required to split a node "min_samples_split", the inclusion of these parameters in the grid was based on their critical role in influencing the model's performance.

2. XGBoost with GridSearchCV: XGBoost was selected for its advanced capabilities in handling sparse data and preventing overfitting through regularisation techniques. Greidsearchcv was employed to thoroughly explore different hyperparameter combinations, including learning rate , learning_rate, tree depth max_depth, regularisation parameters lambda, subsample ratios subsample, and column sampling colsample_bytree. The choice to focus on these hyperparameters was based on XGBoost's sensitivity to these settings, which can significantly impact model accuracy and training speed.

3. RandomForestClassifier with RandomisedSearchCV: RandomSearchCV was chosen for its efficiency in exploring a broader hyperparameter space with fewer computational resources compared to GridSearchCV. This method was particularly useful for sampling from large hyperparameter grids that included variations in the number of trees, n_estimators, features considered max_features, tree depth max_depth, and other important parameters such as min_samples_split,, min_samples_leaf, and bootstrap. The rationale for using RandomisedCV was to strike a balance between search comprehensiveness and computational efficiency, especially when the full grid was computationally prohibitive.

4. BinaryRelevance for multi-label classification: BinaryRelevance was used to adapt the classifiers to a multi-label classification task. This method treats each label independently, which simplifies the problem but requires robust base classifiers. The decision to use BinaryRelevance was influenced by its simplicity and effectiveness in transforming multi-label problems into multiple binary classification tasks.

The outcomes, defined in Subchapter 5.3 confirmed the efficacy of the hyperparameter tuning methods used. Method B involves meticulously identifying optimal configuration for each model in predicting vaccination outcomes. This process involved exploring a wide range of hyperparameters for multiple machine learning models to enhance prediction accuracy. The selection of hyperparameters for each model was driven by both theoretical understanding and empirical evidence. For each model, hyperparameters that significantly impact the model's learning process were selected. For example, in RandomForestClassifier, parameters such as n_estimators (controls number of trees), and max_depth (which limits the depth of the tree) were tuned to control the model's complexity and prevent overfitting. For the XGBClassifier, parameters such as learning_rate (which affects how much each tree contributes to the final prediction) and max_depth (which controls the complexity of the model) were chosen to balance between learning speed and accuracy. These hyperparameters were selected because they directly influence the models performance and its ability to capture patterns in the data.

The execution of the HalvingGridSearchCV was employed to perform hyperparameter tuning. This method was specifically chosen for its ability to efficiently handle large hyperparameter spaces by progressively narrowing down parameter combinations. During this process, the method starts with a large grid of hyper parameters and iteratively eliminates less promising combinations based on model performance, concentrating computational resources on the most promising candidates. This iterative approach allowed for a more exhaustive

search compared to traditional search, while being computationally efficient. After completing the HalvingGridSearchCV, the best hyperparameters for each model were identified based on their performance metrics (i.e. accuracy). For instance, the RandomForestClassifier with 200 trees (n_estimators = 200), a maximum depth of 6 (max_depth=6), and minimum sample split of 2 (min_samples_split=2) was selected as the optimal configuration. These parameters were chosen because they consistently provided highest accuracy during cross-validation, indicating their suitability for prediction tasks at hand. These tuned models demonstrated high accuracy, reflecting their consistency across different data subsets. The results were recorded in a dataframe capturing the best parameters and accuracy metrics for each model. The various types of parameters utilised can be found in Table 4.

This allowed a more efficient exploration of hyperparameter space without compromising the quality of search. Unlike the weaknesses mentioned in Subchapter 4.5 stating that it is computationally intensive, this research found that it was the opposite while being able to get through the parameters at a fast pace compared to other methodologies such as GridSearchCV. Chapter 6 outlines the advantages and disadvantages.

Table 4: Method B Hyperparameters for Each Classifier

Classifier	**Hyperparameter**	**Tuned Values**
RandomForestClassifier	n_estimators	100, 200
	criterion	gini
	max_features	'sqrt', 0.5
	max_depth	4, 6
	min_samples_split	2, 5
XGBClassifier	n_estimators	100, 200
	max_depth	3, 6
	learning_rate	0.1, 0.2
	subsample	0.8
	colsample_bytree	0.8
	use_label_encoder	False
GradientBoostingClassifier	n_estimators	100, 200
	learning_rate	0.1, 0.2
	max_depth	3, 5
	subsample	0.8, 1.0
AdaBoostClassifier	n_estimators	50, 100
	learning_rate	0.1, 1.0
LogisticRegression	C	0.01, 0.1, 1, 10
	penalty	l2
RidgeClassifierCV	No specific hyperparameter tuning	
GaussianNB	No specific hyperparameter tuning	
KNeighborsClassifier	n_neighbors	5, 10
	weights	'uniform', 'distance'
SVC (Support Vector Classifier)	C	0.1, 1, 10
	gamma	0.01, 0.1, 1
DecisionTreeClassifier	criterion	'gini', 'entropy'
	max_depth	4, 6, None
LinearDiscriminantAnalysis	No specific hyperparameter tuning	

4.6 Ensemble

Method A combined the outputs of multiple models to reduce errors and increase robustness by utilising ensemble methods. Ensemble methods were chosen after determining that a single model may not capture complex relationships in the data. The ensemble methods included AdaBoost, Bagging, Extra Trees, Gradient Boosting, and Random Forest, each selected for its ability to reduce overfitting and variance. Syed et al. [16] also used AdaBoost with decision trees to predict the antigenic properties of influenza A H3N2. ShuffleSplit cross-validation with 10 splits was used to evaluate model performance, ensuring the models were tested on various

data subsets to prevent overfitting and assess generalisation.

Method B first split the training and testing sets using "training_test_split" to ensure model evaluation could be performed on unseen data. Feature transformation was then applied using a preprocessing pipeline that handled tasks such as scaling and polynomial feature generalisation. These preprocessing steps were informed by the need to standardise and enhance the feature set for better model performance. Subsequently, a range of models were selected based on their performance in prior studies and their suitability for classification tasks. Hyperparameter grids were defined for each model to optimise their performance referenced in Subchapter 4.5. The models consistently outperform individual models. The decision to proceed with this approach was based on a trade-off between achieving higher accuracy and managing the increased complexity, with the former being prioritised due to the research questions in Subchapter 1.2. Chapter 6 outlines the advantages and disadvantages.

4.7 Model Training & Evaluation

Method A employed GridSearchCV to optimise parameters like n_estimators and max_depth, using a predefined grid and two-fold cross-validation for robust model performance. The model was trained on the subset (10%), with performance evaluated using ROC-AUC scores that were computed for each label. An average macro score was calculated to provide a single evaluation metric representing the model's overall performance across all labels. The training accuracy was computed first to check for any signs of underfitting or overfitting. Subsequently, cross-validation accuracy was calculated to validate the model's ability to generalise to new data.

Model 2 employed an ensemble of machine classifiers (e.g. random forest, XGBoost, gradient boosting, Adaboosting, logistic regression, decision trees, and linear discriminant analysis). This method addresses the research problem of predicting vaccination behaviour by integrating diverse algorithms to improve predictive accuracy which was needed to capture the complex interplay of demographic, behavioural, and opinion-based factors influencing vaccine uptake. The first decision was to split the data into training and testing sets using an 80/20 split to ensure the model's performance could be evaluated on unseen data. This split ratio was chosen to balance the need for sufficient training data with the requirement of having a robust test set for evaluation. Next, feature transformation was conducted using a "columntransformer" that applied scaling and polynomial feature generation to numerical features and one-hot encoding to categorical features. This step was essential to standardise the numerical features to create interaction terms that might capture non-linear relationships, which was hypothesised to improve performance. The decision to use one-hot encoding for categorical features was based on the categorical nature of the data which could not be handled directly by models used. The evaluation of the models focused on accuracy as the primary metric, with stratified k-fold-cross-validation employed to ensure generalisability across different data splits. The choice of accuracy as the main evaluation metric was made because it directly relates to the correct classification of vaccination behaviour, which was critical for public health interventions. Stratified k-fold-cross-validation was selected to preserve class distribution within each fold, addressing potential imbalance in the data. Chapter 6 outlines the advantages and disadvantages.

5 Results & Evaluation

This results chapter presents findings that address the three research questions outlined in 1.2, building on the methodology chapter 3 and implementation chapter 4. It evaluates models performance in predicting vaccination behaviours. The chapter structure is as follows: Subchapter 5.1 discusses results for RQ1, focusing on coefficient analysis and demographic evaluation; Subchapter 5.2 presents findings for RQ2, analysing behavioural variables; Subchapter 5.3 details results for RQ3, comparing Method A and Method B in preprocessing, feature selection, hyperparameter tuning, and model training. These results are critical in meeting the objectives and validating the approaches defined in methodology chapter 3 and implementation chapter 4.

5.1 Research Question 1

Coefficient Results

Coefficient Category	**Coefficient Value**
age_group_18 - 34 Years	0.247422
age_group_35 - 44 Years	0.083541
age_group_45 - 54 Years	0.081367
age_group_55 - 64 Years	0.122413
age_group_65+ Years	-0.005328
education_12 Years	0.133523
education_¡ 12 Years	-0.017701
education_College Graduate	0.269278
education_Some College	0.144314
income_poverty_¡= $75,000, Above Poverty	0.048851
income_poverty_¿ $75,000	0.418512
income_poverty_Below Poverty	0.062051

Table 5: H1N1 Vaccine Effectiveness Model Coefficients

Coefficient Category	Coefficient Value
age_group_18 - 34 Years	-0.028717
age_group_35 - 44 Years	0.109059
age_group_45 - 54 Years	-0.062101
age_group_55 - 64 Years	-0.113067
age_group_65+ Years	-0.361747
education_12 Years	-0.158388
education_¡ 12 Years	0.116832
education_College Graduate	-0.224115
education_Some College	-0.190903
income_poverty_¡= $75,000, Above Poverty	-0.368154
income_poverty_¿ $75,000	-0.179947
income_poverty_Below Poverty	0.091527

Table 6: H1N1 Vaccine Risk Model Coefficients

Coefficient Category	Coefficient Value
age_group_18 - 34 Years	0.053597
age_group_35 - 44 Years	-0.023344
age_group_45 - 54 Years	0.026904
age_group_55 - 64 Years	0.287686
age_group_65+ Years	0.483608
education_12 Years	0.232674
education_¡ 12 Years	0.064788
education_College Graduate	0.330223
education_Some College	0.200766
income_poverty_¡= $75,000, Above Poverty	0.244556
income_poverty_¿ $75,000	0.53422
income_poverty_Below Poverty	0.049676

Table 7: Seasonal Flu Vaccine Effectiveness Model Coefficients

Coefficient Category	Coefficient Value
age_group_18 - 34 Years	-0.129029
age_group_35 - 44 Years	-0.073439
age_group_45 - 54 Years	-0.094083
age_group_55 - 64 Years	0.062335
age_group_65+ Years	0.057103
education_12 Years	-0.046202
education_¡ 12 Years	0.058317
education_College Graduate	-0.115457
education_Some College	-0.073772
income_poverty_¡= $75,000, Above Poverty	-0.226313
income_poverty_¿ $75,000	-0.019067
income_poverty_Below Poverty	0.068267

Table 8: Seasonal Flu Vaccine Risk Model Coefficients

Demographic Evaluation

Age

In regards to H1N1 vaccine effectiveness, young adults, aged 18-34 have a strong positive perception, with a coefficient of 0.247. This suggests younger adults are more likely to believe in the efficacy/effectiveness of the H1N1 vaccine compared to older individuals. Furthermore, the coefficient for older adults, aged 65+, is -0.005, indicating a near-neutral or slightly negative perception of the vaccine's effectiveness. This suggests older adults are less convinced of the H1N1 vaccine's benefits.

Furthermore, in regards to H1N1 vaccine risks - younger adults aged 18-34 have a coefficient of -0.029 indicating relatively neutral perception of risk among younger adults, suggesting they do not see the H1N1 vaccine as particularly risky. Meanwhile, 65+ years of age have a coefficient of -0.362, perceiving the H1N1 vaccine as more risky. This might be due to concerns about side effects or lowered perceived need for the vaccine.

In regards to seasonal flu vaccine effectiveness, the coefficient is 0.054 shows moderately positive perception of the seasonal flu vaccine's effectiveness among younger adults. While the coefficient for 65+ years is 0.484, the highest among all age groups, indicating very strong belief in the effectiveness of the seasonal flu vaccine. This is likely due to public health recommendations and awareness targeted specifically for this age group.

Regards to seasonal flu vaccine risks - the negative coefficient of -0.129 suggests younger individuals perceive a higher risk associated with seasonal flu compared to older individuals having a coefficient of 0.057, indicating low perception of risk among older adults. This is likely due to familiarity and trust in the vaccine.

Education

College graduates with a coefficient of 0.269 in H1N1 vaccine effectiveness have found to have strong positive perception of the H1N1 vaccine's effectiveness. This suggests higher education is associated with greater confidence in the vaccine's efficacy. While less than 12 years

of education, with a coefficient of -0.018, indicates slightly negative perception of the vaccine's effectiveness among individuals, reflecting possible scepticism or lack of information.

Relating to H1N1 vaccine risk, college graduates have a coefficient of -0.224 for college graduates indicating lower perceived risk, suggesting higher education correlates with greater trust in vaccine's safety. While less than 12 years of education show a positive coefficient of 0.117 suggesting individuals with less education perceive H1N1 vaccine as risker, possibly due to less exposure to accurate information.

Similarly, college graduates have a coefficient of 0.330 indicating a strong positive perception of the seasonal flu vaccine's effectiveness, similar to their perception of the H1N1 vaccine. Meanwhile, less than 12 years of education have a coefficient of 0.065, reflecting moderately positive perception of seasonal flu vaccine's effectiveness, though less strong compared to those with higher education.

In regards to seasonal flu vaccine risk, college graduates have negative coefficient of -0.115 suggesting college graduates perceive less risk associated with seasonal flu vaccine. While less than 12 years of education have a coefficient of 0.058, indicating higher perceived risk among those with less education, consistent with the pattern observed for the H1N1 vaccine.

Income

Higher income (>$75,000 USD) have a coefficient of 0.419 for H1N1 vaccine effectiveness. Individuals in higher income brackets have a very strong positive perception of the H1N1 vaccine's effectiveness, likely reflecting greater access to healthcare information and services. While below poverty line have a coefficient of 0.062 suggesting a less positive perception of effectiveness among those below the poverty line, which may reflect limited access to information or healthcare.

Moreover, regarding H1N1 vaccine risk, the higher income have a negative coefficient of -0.180 indicating lower perceived risk among higher-income individuals, suggesting they may feel more secure in the safety of the vaccine. In contrast, the below poverty line have a positive coefficient of 0.092 suggesting higher perceived risk among those below poverty line, which could be due to economic and social factors influencing trust in vaccines.

Regarding seasonal flu vaccine effectiveness, similarly to the H1N1 vaccine, higher income have a coefficient of 0.534. This indicates a very strong positive perception of seasonal flu vaccine's effectiveness among higher-income individuals. In contrast, the coefficient of 0.050 reflects a modestly positive perception of effectiveness among those below poverty line, though it is less strong compared to higher-income groups.

Regarding seasonal flu vaccine risk, the negative coefficient of -0.019 suggests low perception of risk among higher-income individuals. However, below poverty line has a coefficient of 0.068 indicating individuals below poverty line perceive higher risk associated with seasonal flu vaccine.

5.2 Research Question 2

The data set contains key variables:

- Behavioral_wash_hands: frequency with which respondents wash their hands.
- Behavioral_face_mask: usage of face masks by respondents
- Behavioral_antiviral_meds: respondent use of antiviral medications

- Behavioral_avoidance: respondent avoidance of large gatherings
- Vaccination status: indicator of whether the respondent received the h1n1 vaccine
- Seasonal_vaccine: indicator of whether the respondent received the seasonal flu vaccine.

Table 9: H1N1 and Seasonal Flu Vaccine Models

Model	Behavioural Variable	Description
H1N1 Vacc.	Wash Hands (Coefficient: 0.4401)	Frequent hand-washing increases likelihood of vaccination.
	Face Mask (Coefficient: 0.7440)	Strong positive effect on vaccination likelihood.
	Antiviral Meds (Coefficient: 0.4341)	Positive association between antiviral medication use and vaccination.
	Interaction Terms	Hand-washing and mask-wearing together significantly amplify the likelihood of vaccination.
Seasonal Flu Vacc.	Wash Hands (Coefficient: 0.5394)	Positive relationship with vaccination.
	Face Mask (Coefficient: 0.4034)	Mask usage associated with a higher likelihood of vaccination.
	Antiviral Meds (Coefficient: 0.3768)	Positive relationship with vaccination.
	Interaction Terms	Hand-washing and mask-wearing together significantly increase vaccination likelihood.

The analysis utilised logistic regression to model relationships between individual preventative behaviours - such as hand washing, mask wearing, and antiviral medication use - and the likelihood of getting vaccinated. This statistical method was chosen because it is well suited for predicting binary outcomes. In this case, whether an individual was likely to get vaccinated or not, based on their engagement in aforementioned behaviours.

Moreover, within the logistic regression framework, coefficients were calculated for each behaviour. For the H1N1 vaccine model, the coefficient for hand washing was 0.4401, indicating that an increase in the frequency of hand-washing corresponds to a 0.4401 increase in the log-odds of vaccination. The coefficient for wearing face masks was 0.7440, implying that face mask have a stronger positive effect on the likelihood of getting vaccinated. Similarly, the coefficient for antiviral medication use was 0.4341, reflecting positive association with vaccination likelihood.

Interaction terms were incorporated into a logistic expression model to capture the combined effects of multiple behaviours. Specifically the interaction between hand-washing and mask-wearing was found to be significant. This suggests these behaviours together have a greater impact on the likelihood of getting vaccinated than when considered separately. This interaction term indicates that the combined practice of hand-washing and wearing face masks amplifies the likelihood of vaccination beyond the additive effects of a behaviour individually.

Furthermore, the seasonal vaccine model showed similar patterns. The coefficients for hand-washing was 0.5394, face mask usage was 0.4034, and antiviral medication was 0.3768 - all

indicating positive relationships with vaccination likelihood. This interaction term between hand-washing and mask-wearing in this model was also significant. This reinforces the finding that these behaviours have a synergistic effect on vaccination likelihood.

5.3 Research Question 3

Preprocessing

Method A was very thorough in cleaning the data. It involved checking for counting NaN values, resulting in 60,762 NaN values within the data set. Specific columns such as "health_insurance", "employment_industry", and "employment_occupation" were dropped due to nearly half of their values being NaN. Additionally, iterative imputation was used to handle remaining missing values. Outliers were retained after noticing they aided towards the prediction of the model. This approach ensured only relevant and high-quality data was used for model training, laying a strong foundation.

Method B was more efficient, filling all NaN values using mean and mode imputation, avoiding manually cleaning the data. The data set was processed quickly and was balanced adequately for model training. This led to faster preparation and better generalisation. This led to overlooking subtle data patterns, potentially limiting model accuracy.

Method A's pre-processing involves checking and dropping large numbers of NaN values and specific columns, which aim for maximum data quality. Method A was also computationally intensive. This led to strong initial model performance but at cost of overfitting. Method B streamlined approach facilitated preprocessed and produced more generalisable models, though it may have missed finer details.

Feature Selection & Engineering

Method A involved detailed EDA reducing the number of features from 36 to 26 by aggregating related features and conducting correlation analysis using correlation matrix (before EDA) in Figure 3. This created a rich feature set that supported high model accuracy. This captured complex relationships within the data, contributing to strong cross-validation performance (i.e. CatBoostClassifier with 85.26% cross-validation accuracy). However, it was time-consuming and domain knowledge of the topic was needed before conducting EDA.

Method B utilised SelectKBest to reduce features from 36 to 10, focussing on top statistically significant features for each target variable. This method was simpler, and quicker, but still effective. The selection made it quicker to train and less likely to overfit, while still maintaining predictive power. However, the reliance on statistical significance alone might have caused exclusion of less obvious, but potentially important features, limiting the model's depth.

Method A's feature selection reduced features from 36 to 26, contributing to high accuracy and interpretability but requiring more time and risking overfitting. Method B's simpler approach reduced features more drastically, from 36 to 10. This led to a faster, more resource-efficient solution, but potentially at the expense of some depth and detail in the model.

Hyperparameter Tuning

Method A utilised different settings of the model that were tested to find the best one. The result is GridSearchCV and RandomisedSearchCV (and other models) were automatically applied to the hyperparameter space. This led to highly optimised models. However, this approach also

increases the risk of overfitting, as shown by the drop to 83.70% cross-validation. Additionally, it was time-consuming and required lots of CPU computational power. The accuracy and cross-validation results are in Subchapter 5.3.

Method B used HalvingGridSearchCV, a more efficient approach that quickly narrowed down the best parameters. These parameters were automatically applied, such as Method A. this resulted in well-balanced models with minimal overfitting. This method was computationally efficient and focused on maintaining generalisability, reflected in Subchapter 5.3. Note that there is a close alignment between training accuracy (83.80%) and cross validation accuracy (82.44%). While avoiding overfitting, this method did not achieve the same peak accuracy result as Method A, which is a potential limitation.

Method A's exhaustive tuning led to higher accuracy but also increased overfitting and more resources. Method B's more efficient tuning produced models that were more generalizable and less resource intensive, though with slightly lower peak performance.

Model Training and Evaluation

Method A used ensemble methods and cross-validation to optimise model accuracy. For example, the RandomForestClassifier achieved 91.76% training accuracy but dropped to 83.92% in cross-validation, indicating overfitting. The ensemble methods and thorough evaluation helped achieve high in-sample accuracy, effectively capturing complex data. The drop in performance means the model will not perform well on new data.

Method B focused on consistent performance across training, testing, and stratified cross-validation. The SVC model achieved training accuracy of 83.57% and cross-validation accuracy of 82.63% for H1N1 predictions, showing strong generalisability. Method B pipeline is robust and generalizable, making it suited for real-world applications (such as the 2009 H1N1 and seasonal flu data set). This method requires fewer sources, and allows for quicker deployment.

Model A's model training achieved higher accuracy during training, however suffered from overfitting - this makes it less practical for generalisation to new data. Method B, emphasis on generalisability and efficiency, produced models that were more reliable across different data sets, though it had slightly lower accuracy. The study identified Random Forest as the best-performing model, consistent with Hussain and Fatima's [6] findings, which demonstrated Random Forest's effectiveness in predicting influenza outbreaks.

Classifier	**Tuning Method**	**Training Accuracy**	**Cross-Validation Accuracy**
RandomForestClassifier	Grid Search	91.76%	83.92%
	Randomized Search	98.29%	84.24%
XGBClassifier	Grid Search Hyperparameter Tuning	87.18%	84.15%
	Randomized Search	94.79%	83.70%
CatBoostClassifier	Randomized Search	92.02%	85.00%
	Default Parameters	94.34%	85.26%

Table 10: Performance of classifiers with different hyperparameter tuning methods.

Rank	Model	Training H1N1 Accuracy	Test H1N1 Accuracy	Training Seasonal Accuracy	Test Seasonal Accuracy	Stratified CV H1N1 Accuracy	Stratified CV Seasonal Accuracy
4	LogisticRegression	83.24%	84.86%	77.53%	74.21%	82.96%	77.53%
5	RidgeClassifierCV	83.33%	84.30%	77.11%	74.02%	82.96%	76.73%
10	LDA	83.10%	84.30%	77.29%	73.83%	82.91%	76.78%
1	XGBClassifier	90.03%	84.30%	86.66%	75.14%	82.63%	77.99%
9	Decision Tree	83.66%	83.55%	78.09%	73.46%	82.63%	77.62%
8	SVC	83.57%	84.49%	80.15%	75.33%	82.63%	77.48%
0	Random Forest	83.80%	83.93%	78.14%	73.64%	82.44%	77.20%
3	AdaBoost	83.19%	84.67%	77.06%	74.02%	82.40%	77.53%
2	GradientBoosting	84.74%	84.30%	80.99%	74.58%	81.88%	77.90%
7	KNN	83.99%	83.18%	82.30%	74.21%	81.60%	76.73%
6	GaussianNB	80.15%	82.99%	74.72%	72.90%	79.77%	73.97%

Table 11: Performance of different models on H1N1 and Seasonal flu datasets.

Results & Evaluation Discussion

For RQ1, Subchapter 5.1, this study found demographic factors such as age, education, and income significantly influence how individuals perceive effectiveness and risks of H1N1 and seasonal flu vaccines:

- Age: younger adults tend to perceive H1N1 vaccine as more effective but are less confident in seasonal flu vaccine, while older have a strong belief in effectiveness of seasonal flu vaccine but are more cautious about the H1N1 vaccine's risks.
- Education: higher education levels correlate with greater confidence in vaccine effectiveness and lower perceived risks, indicating education contributes to improving vaccine perceptions.
- Income: higher-income individuals consistency view vaccines as more effective and less risky, while those below poverty lines are more cautious and less convinced of their benefits.

These findings heavily inform public health strategies to tailor messaging and interventions according to demographic factors, thereby improving vaccine uptake and addressing specific concerns with different population groups. The importance can heavily be seen in this study's findings.

For RQ2, Subchapter 5.2, this study found the employment of logistic regression analysis demonstrated there is statistically significant positive correlation between engaging in preventative health behaviours - specifically hand-washing and face mask usage - and likelihood of getting vaccination. The inclusion of interaction terms allowed identification of combined effects. This shows these behaviours, especially when practised together, substantially increases probability of vaccination. These findings underscore the importance of promoting such preventive behaviours as part of public health strategies to enhance vaccination uptake.

For RQ3, Subchapter 5.3, found method A focused on achieving high accuracy through preprocessing, feature selection, and tuning. While this approach led to high performance in

training, it came with risks of overfitting and longer process times. This method is ideal if the result is to maximise accuracy on training data, but it will struggle to generalise well to new datasets.

Method B has a more efficient and generalisable approach, producing models that perform consistently well across different data sets, with fewer resources, and in less time. Although method B did not achieve the same peak accuracy as Method A, it practically makes it more suitable for real-world application where timelines, public health emergencies emerge, and resource allocation is limited. Its ability to adapt to new data points is very important.

6 Discussion & Reflection

6.1 Summary of Work

The primary objective was to identify factors influencing vaccination uptake, analysing impact of preventative behaviours, and compare effectiveness of different preprocessing and modeling methods. Method A focused on thorough data cleaning and detailed feature selection, leading to high accuracy but overfitting. Method B prioritised efficiency and generlisability, resulting in more consistent performance across the data, though with slightly lower peak accuracy. Literature review in Chapter 2 was instrumental in selecting methodologies that integrate theoretical models with practical applications in predictive modeling. This research demonstrates how preventative behaviours, when combined with preprocessing techniques, significantly enhance accuracy of vaccination predictors. This bridges gaps between theoretical insights and real world interventions. This study employed dual methodological approaches, listed in Chapter 4. Furthermore, Subchapter 5.1 shows demographic factors significantly influencing vaccination uptake. These insights offer holistic understanding. Subchapter 5.2 reveal preventative behaviours such as hand washing and mask wearing significantly impact vaccination uptake. Subchapter 5.3 reveals Method A achieves higher accuracy but at the cost of overfitting. Method B produced generalisable results, making it suitable for real-world applications. The research highlights compounded effects of combined preventative behaviours and critical role of methodological choices in predicting vacination uptake. These findings suggest targeted behavioural interventions, coupled with efficient generalisable modeling approaches significantly improve vaccination strategies. This work emphasises the importance of balancing accuracy with generlisability - suggesting efficient, behaviour informed models have lasting impact on policy development.

6.2 Thesis Contributions

RQ1 5.1 met by showing higher education and income levels lead to greater vaccine acceptance, directly influencing vaccination rates. This meets objective by quantifying demographic impacts on vaccination behaviour. RQ2 5.2 met by finding that preventative behaviours significantly increase vaccination likelihood, with strong effects when combined. This address objective by highlighting how specific behaviours correlate with higher vaccine uptake. RQ3 5.3 met by comparing Method A and B, showing Method A boosts accuracy but risks overfitting. Method B shows consistent generalisability. This fulfills objective by demonstrating impact of methodological choices on model performance.

In regards to contribution and new knowledge of this study, this thesis provides a novel comparison of two distinct methodological approaches, revealing how Method A (detailed and accuracy focused) and Method B (efficient and generalisable) impact predictive models for vaccine uptake. It advances understanding by quantifying how demographic factors and preventative behaviours directly influence vaccination likelihood, filling gaps in existing current research. Furthermore, the findings can guide future responses to evolving diseases such as H1N1 and seasonal flu. Applying the methodology from this study to other diseases could help determine if similar patterns are present. These insights support practitioners in targeting specific demographic groups for vaccine education, as the study demonstrates that this approach influences vaccination rates. The multi-approach enhances the robustness of findings by comparing detailed, accuracy-focused methods with more efficient, generalisable ones. This comparison provides critical insights into how methodological choices impact model performance and contributes to the field of data science by offering guidance on model selection in public health data analysis. This guides healthcare practioners, analysts, and policy makers (e.g. Government & Parliament) in selecting the most appropriate approach for specific public health interventions, such as WHO.

6.3 Evaluation of Methodological Approaches and Implications

Iterative imputation: Iterative imputation was specifically chosen for its ability to iteratively predict missing values for both numerical and categorical data in a way that maintains dependencies and correlations between variables. One of the key strengths of this approach is its ability to maintain natural variability and patterns within the data. However, a weakness is the overfitting, especially if the imputation model becomes too complex by trying to fit the noise in the data rather than just missing values. This leads to imputed values that do not generalise well to unseen data, resulting in skewed results.

Mean/Mode Imputation: The method's strength is how straightforward it is and its efficiency, preserving data with minimal computation. However, it may not capture complex patterns in missing data, a limitation in datasets with more nuanced structures.

Removing Outliers: Removing outliers showed evidence of overfitting and reduced generalisability. Outliers were preserved as it captures essential variations and correlations in the data, crucial for accurately predicting vaccination behaviour. This approach ensures that the model remains comprehensive and effective across diverse scenarios, aligning with the thesis objective.

EDA: The main strength of this implementation is its ability to condense complex, high dimensional data into a set of interpretable and meaningful features, enhancing both model accuracy and usability. The strength was anticipated based on prior studies such as Hussain and Fatima [6] in Chapter 2 that demonstrated the benefits of feature aggregation in similar contexts, where aggregated features often lead to robust models. However, the aggregation process also introduced a key weakness: the potential oversimplification of the data. Through combining, individual features may have been lost. This reduction in detail could result in the model missing out on nuanced patterns in the data that might be critical for understanding specific behaviours or opinions, particularly in edge cases where individual feature variations are significant.

SelectKBest: The primary strength of this implementation was its ability to simplify the model while retaining most relevant features for each vaccination prediction, thereby improving interpretability. A potential weakness is the chi-squared used in SelectKBest assumes independence between features, which may not hold in cases where interactions between features are significant. This limitation suggests further exploration of feature interactions could enhance model performance, especially in complex scenarios where feature dependencies play a crucial role.

Pipeline with StandardScaler: The main strength of this implementation lies in the consistent application of preprocessing and modelling steps through the pipeline, reducing potential sources of error (i.e. data leakage). However, a potential weakness is that the choice of scaling method, which is generally beneficial, may not capture non-linear relationships between features, which could limit the model's performance in certain contexts.

Hyperparameter Tuning: The GridSearchCV methods provided detailed optimisation at a higher computational cost, while RandomisedSearchCV offered a quicker, more flexible search. The selection of classifiers and hyperparameter tuning strategies was validated by the performance metrics achieved described in Chapter 4 demonstrated their suitability for the research problem at hand. The thorough approach to hyperparameter tuning ensured that the possible configurations were identified for each model. This maximises the predictive accuracy and robustness of the models. By systematically exploring a wide range of hyperparameters the methodology effectively captures the nuances of the data, leading to reliable predictions. Fur-

thermore, the use of HalvingGridSearchCV significantly reduced the computational overhead associated with traditional grid search, as stated in Subchapter 4.5. However, a potential disadvantage that may have been faced was the tuning process aiming to find the best parameters and over-fitting in the cross-validation sets, particularly since the model was too complex as a result of hyper parameter tuning. Overfitting reduces the model's ability to generalise to new, unseen data.

Ensemble: The strength was the increased accuracy due to the ensemble methods, selected to improve robustness in predictions. However, the main weakness was the computation time especially with gradient boost and KNN. The strength was identified based on performance of the ensemble using cross-validation. however, the complexity and resource demands of the ensemble approach were recognised as significant weaknesses. These issues were evident during the hyperparameter tuning process which took substantial computational resources and time.

Model Training & Evaluation: The strength of this implementation is its ability to aggregate the strength of various classifiers, leading to improved prediction accuracy. This strength was anticipated based on prior studies such as Hussain and Fatima [6] referenced in Chapter 2. However, the reliance on accuracy as the sole evaluation metric is a weakness as it may not capture the full performance nuances, especially in imbalanced data sets.

6.4 Legal, Social, Ethical, and Professional Issues

This study handled Legal, Social, Economic, and Professional Issues throughout the project. This included: UK and U.S. [76] [80] legislation (as the data was collected from U.S, ACM Code of Ethics [77], IEEE [78], British Computing Society (BSC) [79], laws [80], and GDPR [81]. From a legal standpoint, both UK and U.S. regulations, particularly the UK's GDPR and Data Protection Act 2018, had to be precisely followed. Article 5 of the GDPR mandates lawful, fair, and transparent processing of personal data, ensuring health data is managed with integrity. [82] Article 6 required a legal basis for data processing, such as consent or public interest, crucial for handling sensitive health data in this project. [83] Article 9 of the GDPR mandates strict safeguards for processing special categories of personal data, including health information, requiring robust protection measures such as anonymisation and secure handling in this project. [84] The United States, HIPAA imposes similar requirements if U.S. sourced data is used. [85] This project stored all the information on the University of Nottingham cloud servers and filled ethical forms (i.e. Full Ethics Checklist, Data Management Plan, and SOP2.2 Text Data) before utilising any data. The data had been anonymised by the United States National Center for Health Statistics, and data characteristics listed within the data set cannot be used to track down any participants. Furthermore, the HIPAA Privacy Rule mandates strict handling of identifiable health information, requiring patient consent and adherence to the U.S. privacy laws. The HIPAA Security Rule further enquires robust security measures to protect electronic health data, ensuring strong data security protocols throughout the project. [85] Subchapter 3.1 discusses the origin of the data.

Social issues are important, especially regarding the project's impact on public health and equitable healthcare distribution. The predictive models avoid reinforcing existing social inequalities, such as any biases in healthcare access or outcomes. This aligns with the ACM Code of Ethics [77], particularly Principle 1.2, which requires computing professionals to avoid causing harm, ensuring models do not introduce biases that could harm public health strategies or worsen equal healthcare outcomes across demographics. [86]

Economically, this project could optimise healthcare resource allocations and provide domain knowledge of how viruses spread, drawn from Chapter 5, specifically RQ1 5.1 and RQ2 5.2.

This would save costs and reduce strain on health care systems due to the number of reports being less. However, the costs for compliance with data protection regulations and security measures must be balanced against these economic benefits for better public health outcomes and more efficient resource use.

Professional issues are tied to ethical standards, with the ACM [77] and IEEE Codes of Ethics [78] providing key frameworks. Principle 1.3 of the ACM Code stresses honesty and trustworthiness, requiring the project team to be transparent about data sources, methodologies and model limitations. [87] Transparency is crucial for building stakeholder trust and ensuring effective implementation within the public health. Principle 1.6 of the ACM Code, which emphasises respecting user privacy, requires the project to implement strict privacy measures, such as data anonymisation and secure storage, to meet legal and ethical standards and protect personal data. [89] The aforementioned actions detail steps taken to meet every principle mentioned.

The IEEE Code of Ethics strengths the project's ethical framework. Principle 1 emphasises public safety, health, and welfare, aligning with the spirit of the project goals in Subchapter 1.2. [88] The project prioritised public health by ensuring predictive models are reliable, risk-free and beneficial to society through accurate predictions. Furthermore, Principle 5 of the IEEE Code highlights the need to improve public understanding of technology's capabilities and societal impacts. The project team must clearly communicate how the predictive models work and their potential effects. This transparency is important for maintaining the public's trust and ensuring responsible use in public health.

In conclusion, the integration of the aforementioned ethical codes, legal frameworks, and professional standard ensures this project was conducted responsibly and with integrity. Adhering to these guidelines and meeting legal requirements upheld high ethical and professional standards that ensure reliable, transparent and beneficial outcomes for public health. This approach is crucial for managing sensitive data, not breaking the law, and developing predictive models with significant public health.

6.5 Project Management

I set a personal deadline for the 30th August due to commitment immediately afterward, which led me to start the project early to gain a few days advantage. This early progress was important in managing the time loss expected later (6 days), allowing me to stay on track. Although I created an initial Gantt chart in Figure 5, I found myself relying more on a daily listing approach, focusing on specific tasks rather than a strict timeline. This flexibility was essential given the unpredictable nature of the project, where I find the Gantt chart does not allow for. This allowed for new techniques and research directions frequently being emerged, and requiring adjustments. Despite this, I still updated the Gantt Chart to look at the over-all progress of the thesis at the end, found in Figure 6 In hindsight, beginning with the literature review would have better informed my methodology and project direction, making the research process more efficient. A subsequent approach of doing the literature review chapter, then the project code, followed by doing the results chapter instantly (as it would be fresh on my mind), then documenting the methodology and implementation would have made the workload easier. Thankfully, these steps were heavily documented in the project code.

Furthermore, past modules from my undergraduate at the University of Nottingham was an enourmous help: COMP4030 (Data Science with Machine Learning) was immensely crucial , helping in my data science skills and report writing; COMP3003 (Undergraduate Dissertation) helped in structuring the overall thesis; COMP4037 (Research Methods) was a stepping stone

in order to write an informative summary of a literature review, and set the standards of what was expected; COMP3013 (Software Quality Assurance) facilitated the development of logical reasoning writing; COMP3020 (Professional Ethics in Computing) aided in writing Subchapter 6.4, detailing Legal, Social, Economic, and Professional Issues. More time should have been spent on preliminary data investigations to completely understand the dataset's characteristics fully, however, I do believe a part of working on a data analyst/science project is the more you go along, the more you understand. This step would have highlighted potential issues early, ensuring subsequent analysis was accurate and effective. The time constraints were a significant challenge, particularly when certain techniques had already been used in a methodology (A or B) which limited my options of choice. The daily task list for the upcoming days helped me navigate through any disruptions and stay focused on my daily objectives. Grouping literature review by methodologies first instead of by topic helped clarify research gaps and ensured my approach was relevant. This helped distinguish between overlapping topics and identify what methodologies would be best suited for the dual approaches. Overall, there were no specific hiccups and specific tasks from meeting to meeting with my supervisor were met - having something to show for every meeting. Of course, more time in any project often leads to new interesting finds and refinements of methodologies. However, I take that as a good sign that progression and contributions to the study and overall field has been made. Therefore, Subchapter 6.6 discusses future suggestions of this project.

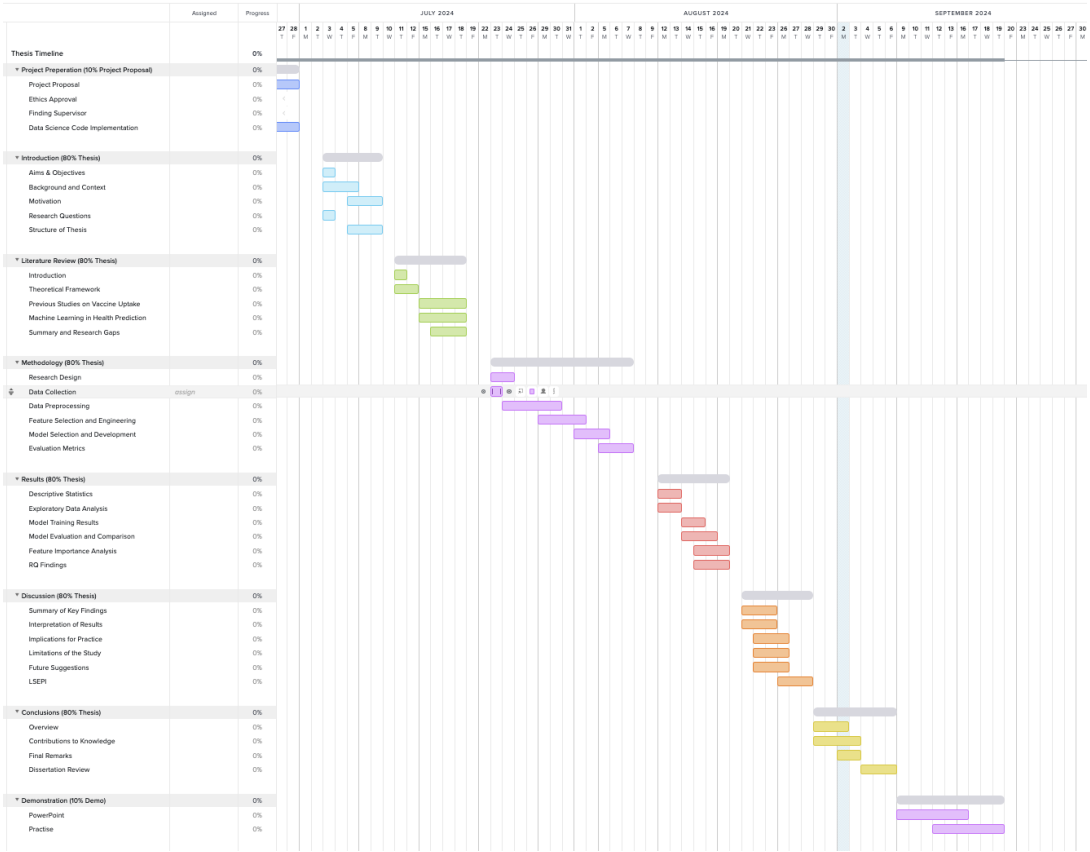

Figure 5: Original project Gantt Chart

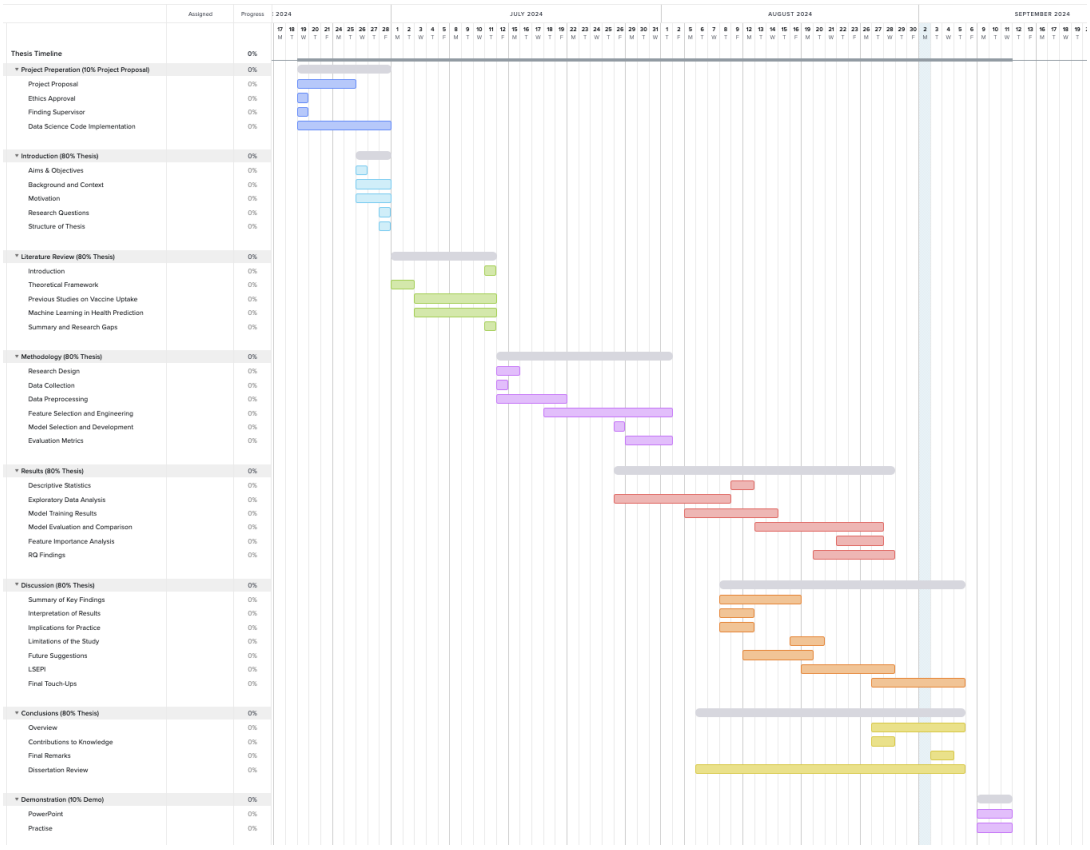

Figure 6: Revised project Gantt Chart

6.6 Future Suggestions

Future research is recommended to use metrics such as precision, recall and F1 score, helps in understanding how well a model is performing across different aspects. These metrics offer a clearer view of true positive identification (recall) and positive precision. Applying them would allow a further detailed comparison of both models, ensuring the chosen model is not just accurate, but also reliable across contexts. Moreover, combining the best elements of both methodologies, now that a better understanding has been developed would be the very first step for the extension of this project. Moreover, the data set used in this study is somewhat outdated (from 2009). Future research should use more recent datasets, especially those reflecting COVID-19 influences. For instance, incorporating variables i.e., religious and cultural hesitancy, (e.g. Jehovah's Witness and blood products, Muslims and vegetarians with gelatin in vaccines) could offer more nuanced analysis - relating it back to RQ1 in Subchapter 1.2 and Subchapter 5.1. This could address current societal challenges in vaccine distribution and uptake, providing fresh insights into how various populations might respond to flu vaccination today.

Furthermore, future work could take a narrower perspective. Now that the study is aware of age groups having an effect on vaccine uptake, manipulating the data to stratify the data into finer age brackets (e.g. "55-64 Years", "35-44 Years", "18-34 Years", "45-54 Years", 65+ Years") from age_group would help capture specific behaviours and attitudes toward flu vaccination. This could reveal how younger groups are influenced by social media, while older groups might respond to traditional media or public health campaigns. Such details could significantly improve the model's predictive accuracy and relevance. This granularity helps the model identify trends and patterns specific to each age group, leading to accurate and relevant predictions in various contexts. Lastly, using Boosted Random Forest, as Indhumathi et al. [7] did, for

high-accuracy seasonal infection accuracy would be an interesting research find.

References

[1] Ayachit, S.S., Kumar, T., Deshpande, S., Sharma, N., Chaurasia, K. and Dixit, M. (2020). Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach. *IEEE Xplore*. https://doi.org/10.1109/ICACCCN51052.2020.9362909.

[2] Busse, W.W., Peters, S.P., Fenton, M.J., Mitchell, H., Bleecker, E.R., Castro, M., Wenzel, S.E., Erzurum, S.C., Fitzpatrick, A.M., W. Gerald Teague, Jarjour, N.N., Moore, W.C., Kaharu Sumino, Simeone, S., Suphagaphan Ratanamaneechat, Madhuri Penugonda, Gaston, B., Ross, T.M., Sigelman, S.M. and Schiepan, J.R. (2011). Vaccination of patients with mild and severe asthma with a 2009 pandemic H1N1 influenza virus vaccine. *The Journal of Allergy and Clinical Immunology*, 127(1), pp.130-137.e3. https://doi.org/10.1016/j.jaci.2010.11.014.

[3] Byrne, C., Walsh, J., Kola, S. and Sarma, K.M. (2011). Predicting intention to uptake H1N1 influenza vaccine in a university sample. *British Journal of Health Psychology*, 17(3), pp.582–595. https://doi.org/10.1111/j.2044-8287.2011.02057.x.

[4] Du, L. and Pang, Y. (2021). A novel data-driven methodology for influenza outbreak detection and prediction. *Scientific Reports*, 11(1). https://doi.org/10.1038/s41598-021-92484-6.

[5] Elbasi, E., Zreikat, A., Mathew, S. and Topcu, A.E. (2021). Classification of influenza H1N1 and COVID-19 patient data using machine learning. https://doi.org/10.1109/tsp52935.2021.9522591.

[6] Hussain, S. and Fatima, U. (2024). Exploring Machine Learning Utilization on Influenza Pandemic Dataset. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-4388322/v1.

[7] Indhumathi, K. and Kumar, K.S. (2022). Seasonal Infectious Disease Prediction based on Electronic Patient Health Records using Boosted Random Forest Algorithms. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. https://doi.org/10.1109/icacite53722.2022.9823453.

[8] Inampudi, S., Johnson, G., Jhaveri, J., Niranjan, S., Chaurasia, K. and Dixit, M. (2021). Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination. *Communications in Computer and Information Science*, pp.139–150. https://doi.org/10.1007/978-981-16-0401-0_11.

[9] Kim, S.-H., Park, S.-H. and Lee, H. (2023). Machine learning for predicting hepatitis B or C virus infection in diabetic patients. *Scientific Reports*, [online] 13(1), p.21518. https://doi.org/10.1038/s41598-023-49046-9.

[10] Li, R., Liu, W., Lin, Y., Zhao, H. and Zhang, C. (2017). An Ensemble Multilabel Classification for Disease Risk Prediction. *Journal of Healthcare Engineering*, 2017, pp.1–10. https://doi.org/10.1155/2017/8051673.

[11] Lober, L., Roster, K.O. and Rodrigues, F.A. (2024). Integrating socioeconomic and geographic data to enhance infectious disease prediction in Brazilian cities. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.01422.

[12] Marquez, E., Barrón-Palma, E.V., Rodríguez, K., Savage, J. and Sanchez-Sandoval, A.L. (2023). Supervised Machine Learning Methods for Seasonal Influenza Diagnosis. *Diagnostics*, [online] 13(21), p.3352. https://doi.org/10.3390/diagnostics13213352.

[13] Nieto-Chaupis, H. (2019). Face To Face with Next Flu Pandemic with a Wiener-Series-Based Machine Learning: Fast Decisions to Tackle Rapid Spread. *IEEE Xplore*. https://doi.org/10.1109/CCWC.2019.8666474.

[14] Singh, S. and Mittal, S. (2023). Pandemic Outbreak Prediction using Optimization-based Machine Learning Model. https://doi.org/10.1109/access57397.2023.10199872.

[15] Sultana, N. and Sharma, N. (2018). Statistical Models for Predicting Swine F1u Incidences in India. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. https://doi.org/10.1109/icsccc.2018.8703300.

[16] Syed, Palomar, D.P., Barr, I., Leo, Ahmed Abdul Quadeer and McKay, M.R. (2024). Seasonal antigenic prediction of influenza A H3N2 using machine learning. *Nature Communications*, 15(1). https://doi.org/10.1038/s41467-024-47862-9.

[17] Towards Using Recurrent Neural Networks for Predicting Influenza-like Illness: Case Study of COVID-19 in Morocco. (2020). *International Journal of Advanced Trends in Computer Science and Engineering*, 9(5), pp.7945–7950. https://doi.org/10.30534/ijatcse/2020/148952020.

[18] Venkatramanan, S., Sadilek, A., Fadikar, A., Barrett, C.L., Biggerstaff, M., Chen, J., Dotiwalla, X., Eastham, P., Gipson, B., Higdon, D., Kucuktunc, O., Lieber, A., Lewis, B.L., Reynolds, Z., Vullikanti, A.K., Wang, L. and Marathe, M. (2021). Forecasting influenza activity using machine-learned mobility map. *Nature Communications*, [online] 12(1), p.726. https://doi.org/10.1038/s41467-021-21018-5.

[19] Venna, S.R., Tavanaei, A., Gottumukkala, R.N., Raghavan, V.V., Maida, A.S. and Nichols, S. (2019). A Novel Data-Driven Model for Real-Time Influenza Forecasting. *IEEE Access*, [online] 7, pp.7691–7701. https://doi.org/10.1109/ACCESS.2018.2888585.

[20] Viana, I., Álvaro Sobrinho, Dias, L. and Perkusich, A. (2023). Machine Learning for COVID-19 and Influenza Classification during Coexisting Outbreaks. *Applied Sciences*, 13(20), pp.11518–11518. https://doi.org/10.3390/app132011518.

[21] Watmaha, J., Kamonsantiroj, S. and Pipanmaekaporn, L. (2021). An Integrated Climate and Spatio-temporal Determinant for Influenza Forecasting based on Convolution Neural Network. https://doi.org/10.1145/3479162.3479178.

[22] Xi, G., Yin, L., Li, Y. and Mei, S. (2018). A Deep Residual Network Integrating Spatial-temporal Properties to Predict Influenza Trends at an Intra-urban Scale. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. https://doi.org/10.1145/3281548.3281558.

[23] Xue, H., Bai, Y., Hu, H. and Liang, H. (2018). Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network. *IEEE Access*, [online] 6, pp.563–575. https://doi.org/10.1109/ACCESS.2017.2771798.

[24] Yakovyna, V., Shakhovska, N. and Szpakowska, A. (2024). A novel hybrid supervised and unsupervised hierarchical ensemble for COVID-19 cases and mortality prediction. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-60637-y.

[25] Yanamala, N., Krishna, N.H., Hathaway, Q.A., Radhakrishnan, A., Sunkara, S., Patel, H., Farjo, P., Patel, B. and Sengupta, P.P. (2021). A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients. *npj Digital Medicine*, [online] 4(1), pp.1–10. https://doi.org/10.1038/s41746-021-00467-8.

[26] DrivenData (n.d.). Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines. [online] DrivenData. Available at: https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/.

[27] Gurney-Read, J. (n.d.). Human rights. *The British Medical Association is the trade union and professional body for doctors in the UK*. Available at: https://www.bma.org.uk/what-we-do/working-internationally/our-international-work/human-rights#:~:text=The%20preamble%20further%20states%20that.

[28] International standards on the right to physical and mental health — OHCHR. Available at: https://www.ohchr.org/en/special-procedures/sr-health/international-standards-right-physical-and-mental-health.

[29] OHCHR. (n.d.). *International Covenant on Economic, Social and Cultural Rights*. Available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights#:~:text=Article%2012.

[30] 2009 H1N1 Pandemic. *Centers for Disease Control and Prevention*. Available at: https://archive.cdc.gov/www_cdc_gov/flu/pandemic-resources/2009-h1n1-pandemic.html.

[31] World Health Organization (2010). Swine flu I WHO emergency situation overview. *World Health Organization*. Available at: https://www.who.int/emergencies/situations/influenza-a-(h1n1)-outbreak.

[32] World Health Organization (2024). Coronavirus Disease (COVID-19) Pandemic. *World Health Organization*. Available at: https://www.who.int/europe/emergencies/situations/covid-19.

[33] Worldometer (2024). Coronavirus Toll update: Cases & Deaths by Country. *Worldometer*. Available at: https://www.worldometers.info/coronavirus/.

[34] WHO (2023). Influenza (Seasonal). *World Health Organization*. Available at: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

[35] GOV.UK. (n.d.). Flu vaccination programme 2023 to 2024: information for healthcare practitioners. *GOV.UK*. Available at: https://www.gov.uk/government/publications/flu-vaccination-programme-information-for-healthcare-practitioners/flu-vaccination-programme-2023-to-2024-information-for-healthcare-practitioners#:~:text=However%2C%20more%20serious%20illness%20may [Accessed 30 Jul. 2024].

[36] Aleem, A., Akbar Samad, A.B. and Slenker, A.K. (2021). Emerging Variants of SARS-CoV-2 And Novel Therapeutics Against Coronavirus (COVID-19). *PubMed*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK570580/.

[37] Keshavamurthy, R., Dixon, S., Pazdernik, K.T. and Charles, L.E. (2022). Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches. *One Health*, 15, p.100439. Available at: https://doi.org/10.1016/j.onehlt.2022.100439.

[38] Syrowatka, A., Kuznetsova, M., Alsubai, A., Beckman, A.L., Bain, P.A., Craig, K.J.T., Hu, J., Jackson, G.P., Rhee, K. and Bates, D.W. (2021). Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. *npj Digital Medicine*, [online] 4(1), pp.1–14. Available at: https://doi.org/10.1038/s41746-021-00459-8.

[39] Buhl, N. Mastering Data Cleaning & Data Preprocessing. [online] Encord.com. Available at: https://encord.com/blog/data-cleaning-data-preprocessing/#:~:text=Removing%20Duplicates%3A%20Duplicate%20entries%20can [Accessed 10 Aug. 2024].

[40] GeeksforGeeks. Data Duplication Removal from Dataset Using Python. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/data-duplication-removal-from-dataset-using-python/ [Accessed 10 Aug. 2024].

[41] lbarrera. Fuzzy Matching 101: Cleaning and Linking Messy Data. [online] Data Ladder. Available at: https://dataladder.com/fuzzy-matching-101/.

[42] Brown, I. Ph.D. Navigating the Outlier Odyssey for Enhanced Machine Learning Models. [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/handling-outliers-ml-best-practices-robust-data-iain-brown-ph-d–mwf6e/ [Accessed 10 Aug. 2024].

[43] Prabhakaran, S. How to detect outliers with z-score. [online] Machine Learning Plus. Available at: https://www.machinelearningplus.com/machine-learning/how-to-detect-outliers-with-z-score/.

[44] Swayam. Outlier detection and removal using box plots and scatter plots. [online] Medium. Available at: https://medium.com/@swayampatil7918/outlier-detection-and-removal-using-box-plots-and-scatter-plots-aab2fa17c4dc#:~:text=Box%20plot%3A%20One%20common%20method [Accessed 10 Aug. 2024].

[45] Linkedin.com. How do you handle data outliers during cleaning and transformation? [online] Available at: https://www.linkedin.com/advice/3/how-do-you-handle-data-outliers-during-cleaning-transformation-zvgfe#:~:text=To%20minimize%20the%20impact%20of [Accessed 10 Aug. 2024].

[46] Acharya, A. Improving Data Quality Using End-to-End Data Pre-Processing Techniques in Encord Active. [online] Encord.com. Available at: https://encord.com/blog/enhancing-data-quality-in-computer-vision/#:~:text=Identifying%20and%20handling%20outliers%20is [Accessed 10 Aug. 2024].

[47] Scikit-learn.org. sklearn.preprocessing.RobustScaler — scikit-learn 0.24.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.

[48] Kamali, A. From Missing to Meaningful: A Dive into Scikit-learn's IterativeImputer for Advanced Imputation Techniques. [online] Medium. Available at: https://medium.com/full-metal-data-scientist/from-missing-to-meaningful-a-dive-into-scikit-learns-iterativeimputer-for-advanced-imputation-c175e21ffb59#:~:text=Scikit%2Dlearn [Accessed 10 Aug. 2024].

[49] Scikit-learn. Imputing missing values with variants of IterativeImputer. [online] Available at: https://scikit-learn.org/dev/auto_examples/impute/plot_iterative_imputer_variants_comparison.html [Accessed 10 Aug. 2024].

[50] Hoque, G. A Better Way to Handle Missing Values in your Dataset: Using IterativeImputer (PART I). [online] Medium. Available at: https://towardsdatascience.com/a-better-way-to-handle-missing-values-in-your-dataset-using-iterativeimputer-9e6e84857d98.

[51] Scikit-learn. sklearn.impute.IterativeImputer. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html.

[52] Data, T. in. Imputing missing data with Scikit-learn's simple imputer. [online] Train in Data's Blog. Available at: https://www.blog.trainindata.com/imputing-missing-data-with-scikit-learns-simple-imputer/.

[53] Scikit-learn.org. sklearn.impute.SimpleImputer — scikit-learn 0.24.1 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html.

[54] Ahmad Najmi Ariffin. In the world of data analysis, the decision between using a 'MICE' (Multiple Imputation by Chained Equations) approach or embracing the prowess of machine learning (ML) can make all the difference in accurately predicting or imputing continuous data. [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/mice-ml-purrfect-solution-data-imputation-ahmad-najmi-ariffin-pmrvc/ [Accessed 10 Aug. 2024].

[55] Segun (jr), O. DESCRIPTIVE ,PREDICTIVE AND FEATURE SELECTION ON TIME SERIES DATA. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/descriptive-predictive-and-feature-selection-on-time-series-data-813a202312b1.

[56] Analytics Vidhya. Step-by-step exploratory data analysis (EDA) using Python. [online] Available at: https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is, formulating%20hypotheses%20for%20further%20analysis.

[57] Developer.ibm.com. IBM Developer. [online] Available at: https://developer.ibm.com/tutorials/awb-reducing-dimensionality-with-principal-component-analysis/.

[58] AVContentTeam. Recursive Feature Elimination: Working, Advantages & Examples. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/.

[59] AK, A. Clever Cuts: Uncovering the Power of SelectKBest for Feature Selection in Machine Learning. [online] Medium. Available at: https://medium.com/@abelkuriakose/clever-cuts-uncovering-the-power-of-selectkbest-for-feature-selection-in-machine-learning-c8d20d75c82f#:~:text=The%20selection%20of%20the%20%E2%80%9Cbest [Accessed 10 Aug. 2024].

[60] D, K. Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning. [online] Medium. Available at: https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48#:~:text=SelectKBest%20is%20a%20type%20of [Accessed 10 Aug. 2024].

[61] AK, A. Unlocking the Power of Dimensionality Reduction Techniques: SelectKBest vs. PCA. [online] Medium. Available at: https://medium.com/@abelkuriakose/unlocking-the-power-of-dimensionality-reduction-techniques-selectkbest-vs-pca-10c7fff8cb60#:~:text=%2D%20SelectKBest%20is%20used%20for%20feature [Accessed 10 Aug. 2024].

[62] Prasanna, C. Streamlining Your Machine Learning Workflow with Scikit-Learn Pipelines — Pipeline Explained. [online] Medium. Available at: https://medium.com/@chanakapinfo/streamlining-your-machine-learning-workflow-with-scikit-learn-pipelines-pipeline-explained-ee63347b01d8 [Accessed 10 Aug. 2024].

[63] Scikit-Learn. sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[64] Google.com. StandardScaler vs MinMaxScaler zero mean unit variance - Google Search. [online] Available at: https://www.google.com/search?q=StandardScaler+vs+MinMaxScaler+zero+mean+unit+variance&oq=StandardScaler+vs+MinMaxScaler+zero+mean+unit+variance&gs_lcrp=EgZjaHJvbWUyBggAEEUYOdIBBjk0ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8 [Accessed 10 Aug. 2024].

[65] Scikit-learn.org. 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/grid_search.html.

[66] Roepke, B. 5-10x Faster Hyperparameter Tuning with HalvingGridSearch. [online] Dataknowsall.com. Available at: https://dataknowsall.com/blog/hyperparameter.html [Accessed 10 Aug. 2024].

[67] Scikit-learn. sklearn.model_selection.HalvingGridSearchCV. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html.

[68] Kalirane, M. Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/#:~:text=Ensemble%20learning%20combines%20multiple%20machine.

[69] Jadama, A.F. Ensemble Learning: Methods, Techniques, Application. [online] doi:http://dx.doi.org/10.13140/RG.2.2.28017.08802.

[70] Bhatnagar, S. Ensemble Methods in Machine Learning - Shashank Bhatnagar - Medium. [online] Medium. Available at: https://medium.com/@shashank25.it/ensemble-methods-in-machine-learning-2d4cc7513c77#:~:text=In%20conclusion%2C%20ensemble%20methods%20are [Accessed 10 Aug. 2024].

[71] Anon. A Comprehensive Guide to Model Evaluation in Machine Learning. [online] Available at: https://graphite-note.com/a-comprehensive-guide-to-model-evaluation-in-machine-learning/#:~:text=In%20machine%20learning%2C%20model%20evaluation.

[72] Chugani, V. A Comprehensive Guide to K-Fold Cross Validation. [online] Available at: https://www.datacamp.com/tutorial/k-fold-cross-validation.

[73] Komarraju, N. Model Evaluation Guide: Cross-Validation Techniques. [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/model-evaluation-guide-cross-validation-techniques-navadeep-komarraju-y4a7c/ [Accessed 10 Aug. 2024].

[74] Pandian, S. K-Fold Cross Validation Technique and its Essentials. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/#:~:text=Employing%20K%20fold%20cross%20validation [Accessed 10 Aug. 2024].

[75] Pandian, S. K-Fold Cross Validation Technique and its Essentials. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/#:~:text=It%20involves%20splitting%20the%20dataset [Accessed 10 Aug. 2024].

[76] EPIC - Electronic Privacy Information Center. (n.d.). U.S. Privacy Laws. *EPIC*. Available at: https://epic.org/issues/privacy-laws/united-states/#:~:text=The%20Privacy%20Act%20of%201974.

[77] Association for Computing Machinery. (2018). ACM Code of Ethics and Professional Conduct. *Association for Computing Machinery*. Available at: https://www.acm.org/code-of-ethics.

[78] IEEE. (2020). IEEE Code of Ethics. *IEEE*. Available at: https://www.ieee.org/about/corporate/governance/p7-8.html.

[79] BCS. (2022). BCS, The Chartered Institute for IT Code of Conduct for BCS Members. *BCS*. Available at: https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf.

[80] GOV.UK. (2023). Data protection. *GOV.UK*. Available at: https://www.gov.uk/data-protection/#:~:text=The%20Data%20Protection%20Act%202018.

[81] UK Government. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). *Legislation.gov.uk*. Available at: https://www.legislation.gov.uk/eur/2016/679/contents.

[82] General Data Protection Regulation (GDPR). (n.d.). Art. 5 GDPR – Principles relating to processing of personal data. *GDPR Info*. Available at: https://gdpr-info.eu/art-5-gdpr/#:~:text=5%20GDPR%20Principles%20relating%20to.

[83] Intersoft Consulting. (2013). General Data Protection Regulation (GDPR). *GDPR Info*. Available at: https://gdpr-info.eu/art-6-gdpr/.

[84] Intersoft Consulting. (2013). General Data Protection Regulation (GDPR) – Final text neatly arranged. *GDPR Info*. Available at: https://gdpr-info.eu/art-9-gdpr/.

[85] U.S. Department of Health & Human Services. (2019). Health Information Privacy. *HHS.gov*. Available at: https://www.hhs.gov/hipaa/index.html.

[86] www.acm.org. (n.d.). The Code affirms an obligation of computing professionals to use their skills for the benefit of society. *ACM*. Available at: https://www.acm.org/code-of-ethics/#:~:text=1.2%20Avoid%20harm.&text=Well%2Dintended%20actions%2C%20including%20those.

[87] Acm.org. (2023). The Code affirms an obligation of computing professionals to use their skills for the benefit of society. *ACM*. Available at: https://www.acm.org/code-of-ethics/#:~:text=as%20public%20resources.- [Accessed 3 Sep. 2024].

[88] DX Editor. (2018). Code of Ethics. *IEEE Computer Society*. Available at: https://www.computer.org/education/code-of-ethics/#:~:text=Principle%201%20%E2%80%93%20PUBLIC [Accessed 3 Sep. 2024].

[89] Acm.org. (2023). The Code affirms an obligation of computing professionals to use their skills for the benefit of society. *ACM*. Available at: https://www.acm.org/code-of-ethics/#:~:text=1.6%20Respect%20privacy.&text=Therefore%2C%20a%20computing%20professional%20should [Accessed 3 Sep. 2024].

[90] Centers for Disease Control and Prevention. (n.d.). CDC estimates of 2009 H1N1 influenza cases, hospitalizations and deaths in the United States, April 2009 – March 13, 2010. *Centers for Disease Control and Prevention*. Available at: https://archive.cdc.gov/www_cdc_gov/h1n1flu/estimates/April_March_13.htm [Accessed: 06 Sep. 2024].

7 Appendices

SurVis literature collection

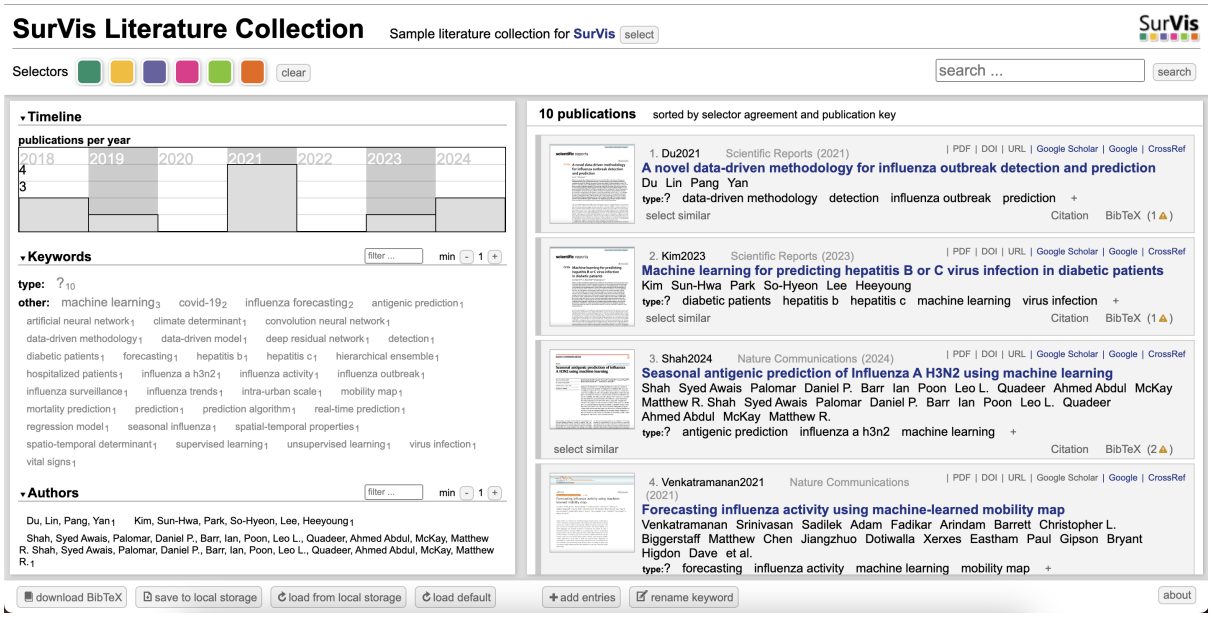

Figure 7: SurVis literature collection, `https://coderab23.github.io/survis-demo/src/index.html`

Shah et al. (2024) has been categorized as real-world & applications (healthcare) in data enhancement & transformation because it focuses on predicting antigenic properties of influenza A H3N2 viruses through data transformation.

Venkatramanan et al. (2021) has been categorized as data-centric (data-types) in data Enhancement & transformation because it relies on anonymized mobility data to enhance and transform data for influenza forecasting.

Yakovyna et al. (2024) has been categorized as multivariate & hierarchical (hierarchical) in data enhancement & transformation because it uses complex hierarchical data structures combining supervised and unsupervised learning for COVID-19 predictions.

Xue et al. (2018) has been categorized as multivariate & hierarchical (high dimensional overview) in Visual mapping & structure because it utilizes multiple regression and neural networks, emphasizing the visual mapping of flu activity predictions.

Venna et al. (2019) has been categorized as geospace + time (temporal) in interaction & analysis because it integrates temporal dynamics in its LTSM model for influenza forecasting, focusing on interaction and analysis of temporal data.

Xi et al. (2018) has been categorized as real-world & applications (healthcare) in exploration & rendering because it applies deep residual networks for influenza trend prediction, focusing on the exploration and rendering of healthcare data.

Watmaha et al. (2021) is categorized under geospace + time (temporal) in exploration & rendering for its CNN model that uses spatial-temporal and climate data for real-time flu

forecasting, and under real-world & applications (healthcare) in interaction & analysis for its application to healthcare data for interactive flu forecasting.

Yanamala et al. (2021) has been categorized as real-world & applications (healthcare) in Interaction & Analysis because it develops ML models to differentiate between COVID-19 and influenza, emphasizing interactive analysis in healthcare settings.

kim et al. (2023) has been categorized as real-world & applications (healthcare) in interaction & analysis because it evaluates ML models for predicting hepatitis B and C in diabetic patients, emphasizing the interaction and analysis of health data for prediction and screening.

Du et al. (2021) has been categorized as multivariate & hierarchical (hierarchical) in interaction & analysis because it uses a hybrid model combining multiple data sources for early influenza outbreak prediction, focusing on interactive analysis.

Word count: 19,810[1] / 20,000

[1]This word count is an approximate total, including all words in the Overleaf document, excluding appendices, footers, and tables. It may also include Overleaf formatting elements.