LONDON'S GLOBAL UNIVERSITY
Department of Computer Science

Optimizing Validation Methods for Influenza-Like Illness Rate Prediction Models Using Web Queries

Yueling Huang

Supervisors: Vasileios Lampos

Faculty of Engineering

Department of Computer Science

University College London

A Project Report Presented in Partial Fulfillment of the Degree

Artificial Intelligence for Biomedicine and Healthcare

September 2024

Abstract

Numerous machine learning and neural network models have successfully utilized web search data to accurately forecast and provide warnings about influenza-like illness (ILI) rates across different geographic locations. Most of these models rely on manually tuned parameters or use the last segment of the training set as the validation set. However, the ILI rate trend within a single flu season (from September 1st to August 31st) typically follows a near-normal distribution, with different periods within the season potentially containing distinct information. Consequently, using different periods as validation sets can influence parameter selection, leading to varying model performance.

In this study, we will develop a three-layer feedforward neural network (FFNN) and validate five parameters using six different validation set configurations, each comprising different periods. The model will be applied to web search data from the United Kingdom for nowcasting and forecasting ILI rates 14 and 21 days ahead. Model performance will be evaluated using Mean Absolute Error (MAE) and correlation metrics across three consecutive flu seasons (2016-2017 to 2018-2019). The goal is to identify the optimal validation strategy for each of the three distinct tasks.

In all tasks, models selected using validation strategies generally outperformed the baseline that did not use validation strategies, except for the forecasting task during the 2017-2018 flu season. In the nowcasting task, the best strategy, with a validation set covering the onset, peak, and end periods of the last three years of the training set, reduced the mean absolute error (MAE) by 38.4% compared to the baseline. It also showed an 8.2% improvement over the strategy using the last 180 days of the training set for validation. For 14- and 21-day forecasts, the strategy based on peak periods from the last three years of the training set performed best, reducing MAE by 9.1% and 4.7%, respectively, compared to the 180-day validation strategy.

Keywords— Web Search Data - Influenza-Like Illness Rate - Validation Strategy

Declaration

I, Yueling Huang, I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included and referenced. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Yueling Huang

Signature

09/09/24

Date

Table of Contents

List of Figures

List of Tables

List of Abbreviations

FFNN: Feedforward Neural Network

ILI: Influenza-Like Illness

MAE: Mean Absolute Error

1 | Introduction

1.1 Research Objective

Influenza, commonly known as the flu, is a serious respiratory illness caused by influenza viruses. According to the World Health Organization (WHO), influenza remains a significant threat and a strong candidate for causing pandemics. There are around a billion cases of seasonal influenza annually, including 3 to 5 million cases of severe illness[1]. Beyond the immediate health impact, influenza outbreaks impose a substantial economic burden due to lost productivity, increased healthcare costs, and societal disruptions[2]. These figures highlight the critical impact of influenza on global public health, necessitating effective monitoring and control measures.

Given the significant threat posed by influenza, there is a growing interest in accurately predicting ILI rates[3, 4, 5]. Traditional surveillance methods, such as those employed by the Royal College of General Practitioners (RCGP) and the UK Health Security Agency (UKHSA) in the United Kingdom, provide essential data on Influenza-Like Illness (ILI) rates, representing the incidence of doctor consultations for ILI symptoms per 100,000 individuals in England. However, these methods are relatively slow to update, with weekly reporting cycles that may not capture the dynamic nature of influenza spread in real-time[6]. Furthermore, underreporting of cases-particularly those not resulting in a medical consultation-can lead to an incomplete picture of the outbreak's true scale[7].

In contrast, leveraging web search data offers a more timely, convenient, and cost-effective means of gaining insights into the population's health status[8]. This approach not only allows for near-real-time monitoring but also captures a broader spectrum of health conditions by including individuals who may not seek immediate medical attention[9]. Moreover, advancements in big data analytics and machine learning have shown promise in enhancing the accuracy of influenza predictions by analyzing vast amounts of digital data[10, 11, 12]. These technological innovations are increasingly recognized as valuable tools in complementing traditional surveillance systems and improving public health response strategies.

In summary, existing literature demonstrates the significant potential of integrating web search data with advanced machine learning models for influenza prediction. However, these studies often lack detailed explanations of parameter selection and fail to employ validation strategies to ensure the optimality of the chosen parameters. To address this gap, this project aims to develop a model that incorporates dynamic validation and optimization processes. We will systematically evaluate the impact of using different time periods as validation sets to determine the optimal validation period, thereby achieving the best-performing model.

The impact of using different time periods as validation sets on the model should not be overlooked. Each time period encompasses distinct epidemiological characteristics; for example, data from the early stages of an influenza outbreak differ significantly in information density and patterns compared to data from the peak of the outbreak. Excluding data from a specific time period from the training set and using it as a validation set could significantly affect the model's training outcomes. Therefore, comparing the effects of using different time periods as validation sets is a topic worth exploring in depth. For instance, using data from the peak of the flu season as the validation set might lead to a model that is more focused on predicting extreme cases, whereas using data from the early stages of the outbreak might enhance the model's sensitivity to early warning signs.

By systematically evaluating how the choice of validation sets affects model performance, this project will gain a deeper understanding of the role different validation strategies play in model optimization. This will contribute to the development of more robust influenza prediction models, providing more reliable tools for public health decision-making and resource allocation, and ultimately making a significant contribution to the field of digital epidemiology.

In this project, we found that models using validation strategies outperformed the baseline model in all tasks except for the 17-18 flu season forecasting task. In the nowcasting task, the baseline modelâs MAE loss was 38.4% higher than that of the best-performing Strategy 6, which incorporated the onset, peak, and end periods from the last three years of the training set. Strategy 1, which used the last 180 days of the training set, serves as the baseline validation strategy. In the nowcasting task, Strategy 6's MAE was 8.2% lower than that of Strategy 1. In the forecasting tasks for 14 and 21 days ahead, Strategy 3 (used peak periods from the last three years) performed the best, achieving the lowest average MAE, with reductions of 9.1% and 4.7%, respectively, compared to Strategy 1.

1.2 Project Aims and Goals

Aims:

1. Advance epidemiological surveillance in England: Explore and identify the optimal validation strategies for feedforward neural networks in nowcasting and forecasting tasks by using web search data as input.
2. Enhance proficiency in machine learning algorithms: Deepen understanding and mastery of machine learning algorithms, and further improve practical skills in model training.

Goals:

1. Construct a Feedforward Neural Network (FFNN) Model: Build an FFNN model for three tasks: nowcasting, 14-day forecasting, and 21-day forecasting. Systematically apply and evaluate different validation strategies across these tasks.
2. Compare and Validate Different Validation Strategies: Analyze and compare the effects of different validation strategies on the FFNN model's performance in each

task, identifying the optimal validation strategy for each task.

1.3 Project Overview

Before training our machine learning models, we pre-processed the dataset to ensure its suitability for experimentation. This process included data interpolation, smoothing, and retaining only the queries most relevant to influenza.

The primary objective of this experiment was to identify the optimal validation strategy for a model that uses web search data as input to nowcast or forecast ILI rates 14 or 21 days ahead. We constructed a three-layer Feedforward Neural Network (FFNN) and optimized the model by selecting the best combination of parameters based on validation loss. The parameters considered included the top N most relevant query search frequencies ($N \in \{100, 200, 300, 400\}$), the input window length (7 or 14 days of top N query search frequency), the number of units in each neural network layer, learning rate, batch size, and other hyperparameters.

By comparing the performance of these various parameter combinations across different validation sets, we identified the best-performing model for each validation strategy. To ensure the robustness of our results, we repeated the experiments across 10 different random seeds, allowing us to calculate the average performance and ultimately determine the optimal validation strategy for this task.

1.4 Report Overview

In the subsequent chapters, the Background section will be introduced first. In this chapter, I will explain in detail the definitions of web search data and Influenza-like Illness (ILI), and review the related research in this field. Furthermore, I will explore existing methods and findings on predicting ILI rates using web search data, thereby clarifying the background and motivation of this study.

The Methodology section will describe the data sources of this study and their specific contents, detailing the data preprocessing steps, as well as the design and construction of the neural network architecture. Additionally, I will discuss the validation strategies set in this experiment.

In the Results and Analysis section, I will present and analyze the results of the nowcasting experiments and the 14-day and 21-day ILI rate forecasting experiments conducted in this study. By comparing the performance of different validation strategies in these tasks, I will analyze the computational resources required and the performance across different test seasons to identify the optimal validation strategy for each task.

In the Discussion section, I will first summarize the key results. Following this, I will discuss any limitations encountered during the project and suggest possible future directions for further research or improvement.

In the Conclusion section, I will provide an overview of the entire project.

2 | Background

In this chapter, I will introduce the definitions and research related to Web search data and Influenza-like Illness (ILI). Additionally, I will discuss the models and validation methods used for nowcasting and forecasting ILI rates with Web search data, thereby clarifying the background, direction, and knowledge gaps addressed by this study.

2.1 Web Search Data

Web search data can be succinctly defined as the digital footprints generated by users' queries on search engines. Beyond merely transmitting information, web search data serves as a powerful tool for predicting future trends across various domains, including economics, politics, and health. In this section, I will provide a detailed analysis of the impact of web search data in these fields.

In the financial markets, web search data has shown significant and multifaceted implications. Research by Szczygielski et al.[13] demonstrates the influence of Google Search Trends (GST) on stock markets. As a form of web search data, GST acts as an indicator of market sentiment, attention, and uncertainty. It is notably correlated with stock market returns and volatility. Typically, an increase in search volume signals impending significant price movements, potentially leading to decreased returns and increased volatility. Thus, GST can be used to predict systematic market drivers and their impact on market volatility. Furthermore, Aoki's research[14] highlights that web search data not only reflects public interest and behavior towards the economy but also provides more timely forecasts of economic indicators compared to traditional metrics like GDP and unemployment rates, which often lag by several months. The real-time nature of search data allows it to quickly capture sudden changes in the economic environment, offering timely economic forecasts and supporting decision-making processes.

In the political realm, web search data has also demonstrated substantial impact. Camilo's research [15] reveals that since 2004, Google Trends has successfully predicted all election outcomes in Canada and the United States. Additionally, Kassraie's study [16] indicates that when Google Trends data is analyzed in conjunction with Twitter data and validated against actual election results, it can accurately reflect public interest and sentiment.

Web search data has further proven its significance in the health sector. For example, Agarwal's study [17] demonstrated that by analyzing data from the Baidu search engine, researchers could predict whether users would visit medical facilities in the future based on their search behavior. This finding underscores the potential of web search data in forecasting healthcare service demand.

Across various domains, web search data has proven to be a valuable resource for forecasting and understanding trends. Its ability to provide real-time insights and reflect public sentiment makes it a powerful tool in contemporary research. Whether in finance, politics, or health, web search data enhances predictive models, supports decision-making, and offers a deeper understanding of complex dynamics in the modern world.

2.2 Influenza-Like Illness

Influenza is a respiratory infectious disease caused by the influenza virus, typically characterized by fever, cough, sore throat, muscle aches, headaches, and fatigue. The influenza virus can spread through droplets and can easily propagate among people, especially during the winter and early spring seasons. According to the WHO [1], it causes 290,000 to 650,000 respiratory deaths annually. Influenza can exacerbate symptoms of other chronic diseases, and in severe cases, it can lead to pneumonia and sepsis. Individuals with underlying medical conditions or severe symptoms should seek medical attention.

Influenza is closely related to Influenza-Like Illness, a group of respiratory infections that exhibit similar symptoms to influenza. Although ILI symptoms such as fever, cough, and sore throat resemble those of influenza, they may be caused by other viruses, such as rhinoviruses or parainfluenza viruses. The definition and monitoring of ILI are commonly used in public health to help track the spread of influenza and similar illnesses.

Traditional methods of influenza surveillance primarily rely on data collected from healthcare facilities such as clinics, hospitals, and emergency rooms. For example, the syndromic surveillance network coordinated by the U.S. Centers for Disease Control and Prevention (CDC) largely depends on data gathered from various healthcare institutions across the country. In the United Kingdom, data on ILI rates are primarily collected and analyzed by two agencies: the Royal College of General Practitioners (RCGP) and the UK Health Security Agency (UKHSA), based on reports of ILI symptoms from general practitioner (GP) clinics.

However, these data have certain limitations. Firstly, some patients may not actively seek medical attention when they contract influenza, leading to a potential underestimation of the actual ILI rate. Additionally, in developing countries, limited healthcare resources mean that only a small proportion of patients visit hospitals, resulting in less accurate ILI rate data obtained from medical institutions. Furthermore, government-released official ILI rate data often experiences delays, as these figures are compiled and published on a weekly basis, typically 1 to 2 weeks after data collection. Therefore, analyzing web search data related to ILI provides a more real-time understanding of ILI spread compared to official governmental methods, and it also covers a broader population.

2.3 ILI Rate Prediction Using Web Search Queries

In this section, I will introduce nowcasting and forecasting models that utilize web search data to predict ILI rates, along with the validation methods associated with these models.

2.3.1 Nowcasting Models

Ginsberg et al.[9] first proposed the Google Flu Trends (GFT) model, which uses Google search query data to predict influenza trends in real-time. The core idea of GFT is to utilize the frequency of searches related to flu symptoms on the internet as an indicator of flu activity. The research team selected flu-related keywords from Google search data covering 50 states in the United States as potential indicators of flu activity and constructed a linear regression model based on this data, comparing the selected search query frequencies with the CDC-reported influenza-like illness (ILI) data.

The model was constructed using a four-fold cross-validation method, fitting four 96-point subsets of 128 data points for each region to ensure the robustness and predictive capability of the model. To validate the accuracy of the model, the research team compared the ILI rates predicted by the GFT model with historical CDC data. Across nine different regions, the GFT model achieved an average correlation of 0.90 with the CDC-reported ILI percentages. Furthermore, in a separate validation conducted in Utah, the model also demonstrated a correlation of 0.90 across 42 validation points.

Nevertheless, GFT faces some challenges. For instance, when the media extensively reports on flu or related health issues, public search behavior may be influenced, leading to biases in the model's predictions. To address this issue, Lampos et al.[11] enhanced the model's resilience to noise by optimizing feature selection and regularization techniques, thereby improving its stability across datasets from different seasons and regions.

In their study, the team first used an Elastic Net regression model to filter out flu-related keywords from a large number of search queries, identifying these keywords as potential indicators of flu activity. After selecting the relevant search keywords, these were fed into a nonlinear regression framework based on Gaussian Processes (GP) to capture the complex nonlinear relationships in the data. Furthermore, to enhance the model's predictive capability, the study integrated autoregressive elements, particularly the ARMAX model from the Autoregressive Moving Average (ARMA) series. This model not only included the regression of past Influenza-Like Illness values but also incorporated seasonal components targeting annual patterns, trained using maximum likelihood estimation.

Overall, the Elastic Net model achieved a Mean Absolute Percentage Error (MAPE) of 11.9% compared to 20.4% for the GFT model, while the GP model further reduced the MAPE to 10.8%. Regarding the performance of autoregressive models, the AR+GFT model achieved a MAPE of 10.2% with a 2-week lag; the AR+Elastic Net model reduced the MAPE to 7.7% under the same lag; and the AR+GP model achieved the best MAPE of 7.3% with a 2-week lag.

Lampos et al. [18] improved feature selection by using word2vec for word embeddings and filtering noise. They used similarity scores S to exclude unrelated terms. The study also introduced a hybrid approach, combining correlation-based feature selection with word embeddings for optimization. Results showed that the correlation-based model (with $r > 0.4$) outperformed the word embedding model, achieving a mean absolute error (MAE) of 2.137 compared to 3.006. The hybrid method, with thresholds of $r > 0.30$ and $S > \mu_S + \sigma_S$, yielded the best performance, with a correlation of 0.913, an MAE of 1.880, and a MAPE of 36.23%.

2.3.2 Forecasting Models

In Hickmann et al.'s study [19], Wikipedia page views related to influenza were used to predict the ILI rate one week ahead. They primarily utilized CDC's ILI report data and collected access data from relevant Wikipedia articles like "Human Flu," "Influenza," and "Oseltamivir." The model was based on the SEIR (Susceptible-Exposed-Infected-Recovered) differential equation model, simulating influenza transmission in the U.S. Enhancements included incorporating seasonal variations and heterogeneity in contact structures. The Kalman smoother (enKS) method was used to dynamically adjust model parameters with new ILI and Wikipedia data. The researchers compared their SEIR model to a baseline "straw man model," which generated predictions based on average and standard deviation of previous flu seasons. The SEIR model was more accurate in the first half of the flu season, reducing Mahalanobis distance (M-distance) by up to 20%. However, the study also noted that the SEIR model had limitations during the later stages, where the assumption of immunity caused rapid declines in predicted ILI values, making it challenging to predict the season's tail end. This highlights the need for further improvements to enhance the model's performance throughout the entire flu season, particularly in predicting subsequent peaks.

In their study, Morris et al. [20] utilized Google Health Trends web search activity data and historical influenza-like illness rates from the CDC to forecast flu trends using neural network models. To enhance prediction accuracy, they incorporated Bayesian methods to estimate the uncertainty of the model's predictions. During data processing, they analyzed the correlation and semantic similarity of search queries, selecting the most relevant ones for model training, and standardized these queries to create a composite score guiding the model inputs. Additionally, they interpolated the weekly ILI rate data into daily data to align with the search frequency time series. The study developed three neural network architectures: a Feedforward Neural Network (FF), a Simple Recurrent Neural Network (SRNN) using Gated Recurrent Units (GRU), and an Iterative Recurrent Neural Network (IRNN). A Bayesian layer was introduced in the final layer of these neural networks to better capture the uncertainty in the model's predictions. Cross-validation and Bayesian optimization were used to find the optimal model parameters. The results showed that the IRNN consistently outperformed the other models in predicting ILI rates across different time horizons (e.g., 14 days, 21 days, and 28 days). Compared to the traditional Dante model, the IRNN demonstrated an 11.93% higher skill score, 4.97% lower MAE, and 5.96% higher correlation, highlighting its superior ability to capture ILI rates.

In their study, Wei et al. [21] utilized weekly ILI% data from the Chinese National Influenza Center (CNIC) and daily search data for 30 influenza-related keywords from the Baidu search platform, covering 31 provinces, to predict influenza-like illness rates in China.During the data processing phase, the researchers first calculated the Pearson correlation coefficient to assess the relationship between Baidu search queries and ILI rates. Keywords with a correlation coefficient above 0.4 were selected for use in the predictive model. Additionally, the researchers normalized the Baidu search index and ILI% data to a range between 0 and 1 to facilitate more in-depth analysis and model training.

The study employed a three-layer Long Short-Term Memory (LSTM) neural network model. To prevent overfitting, the researchers applied L2 regularization and other techniques. The model's performance was validated by fine-tuning the model parameters using the last year of the training set data to ensure it did not overfit. Moreover, they expanded the dataset threefold through data augmentation, thereby enhancing the robustness of the training process. The results showed that the LSTM model, combined with the Baidu search index, significantly improved the accuracy of ILI rate predictions compared to models using only ILI rate data. In northern mainland China, the combination of ILI% data with the "mask" keyword index (lagged by one week) provided the best predictive performance, with an R^2 of 0.9055 and a 16.75% reduction in RMSE. In southern mainland China, the best results were achieved using a combination of the "influenza name" keyword index with ILI% data, with an R^2 of 0.75579 and a 4.20% reduction in RMSE.

2.3.3 Summary

Overall, web search data and neural networks have been widely used to predict influenza-like illness rates, demonstrating excellent performance and predictive capabilities. However, in reviewing existing studies, the selection of model parameters typically follows one of several approaches: manual tuning, cross-validation, or using the last year's data as a validation set. Given that different stages of a flu season (such as the early phase, peak phase, and end phase of ILI activity) contain varying amounts and types of information, choosing different periods of data as the validation set may significantly impact parameter selection and, consequently, model performance. Therefore, in this project, I will explore the effects of using different periods as validation sets to assess how various validation strategies influence model performance, aiming to optimize the model's predictive ability.

3 | Methodology

In this section, I will provide a detailed presentation of the datasets used in this study, the steps and methods of data preprocessing, the architecture of the neural networks, and the validation strategies employed.

3.1 Dataset

In this study, we utilized a time series dataset spanning from August 2008 to September 2019. This dataset includes both weekly ILI rates and daily web search query frequencies.

The ILI rates were sourced from the Royal College of General Practitioners (RCGP) and the UK Health Security Agency (UKHSA), representing the incidence of doctor consultations for ILI symptoms per 100,000 individuals in England. Figure 3.1 displays the interpolated daily ILI rates based on the weekly ILI data.

Additionally, we obtained 22,571 web search query frequencies through an academic API provided by Google Health Trends for academic research purposes with a health-oriented focus. These are a non-standardized version of the publicly available Google Trends outputs. Examples of these queries are provided in Table 3.1. The search query frequencies were normalized using min-max scaling, with detailed information mentioned in Section 3.3.2. This normalization adjusts the search frequencies to a consistent range, enhancing the reliability and comparability of the analysis.

3.2 Data Preprocessing

In this section, we introduce how the queries embedding job is done, how the query frequency processed and ILI rate interpolation.

3.2.1 ILI Rates Interpolation

We receive weekly ILI rate data, assuming that the weekly ILI rate represents the ILI rate on Thursday of each week. We then use linear interpolation to generate the daily ILI rate between each pair of Thursdays. The interpolation formula is as follows:

$$I_t = I_w + \frac{I_w - I_{w-1}}{7} \times d \tag{3.1}$$

where I_t represents the ILI rate on day t between the Thursday of week $w-1$ and the Thursday of week w, I_w represents the weekly ILI rate (considered as the ILI rate on the Thursday of week w), and d represents the number of days between day t and the Thursday of week $w-1$. The interpolation result is shown in Figure 3.1.

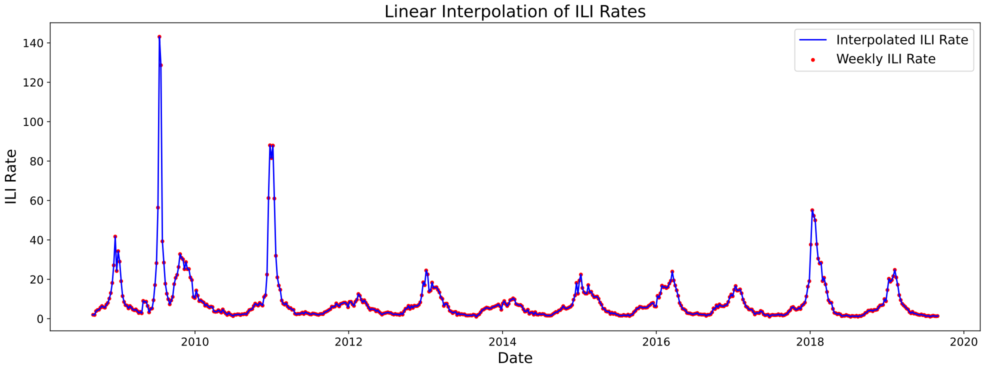

Figure 3.1: Daily influenza-like illness (ILI) rates in England from September 1, 2008, to August 31, 2019, as reported by RCGP and UKHSA.

3.2.2 Query Embedding

We obtained a total of 22,570 search queries; however, not all of them are related to influenza-like illness (ILI). Our approach involves first embedding these queries and then calculating their cosine similarity with flu-related phrases.

To obtain the embeddings of the queries, we used the SBERT model, specifically the *all-MiniLM-L6-v2*, a pre-trained sentence transformer model from the SentenceTransformers Python library. This model, based on BERT's MiniLM, employs a 6-layer Transformer encoder and produces 384-dimensional embedding vectors as output. It has been extensively trained on various datasets, including those with healthcare-related terminology, ensuring high-quality sentence embeddings suitable for computing cosine similarity[4].

Next, we identified the most flu-related queries by calculating their cosine similarity with seven flu-related phrases: 'flu', 'fever', 'flu medicine', 'gp', 'hospital', 'flu symptoms', and 'influenza vaccine'. For each query, we computed a score by averaging the cosine similarities with these seven phrases. A higher score indicates a higher probability that the query is related to the flu. We then ranked these scores and selected the top 2,000 queries for further processing.

3.2.3 Query Frequency Smoothing

The daily search frequencies of queries are influenced by various factors, resulting in a significant amount of short-term fluctuations and noise. By applying a moving average smoothing process to the data, we can obtain a smoother query frequency trend, allowing for a clearer observation of the long-term trend in query frequency. In this study, we use the harmonic mean with a 14-day window to smooth the query frequencies:

$$f_t = \frac{f_t + \frac{1}{2}f_{t-1} + \frac{1}{3}f_{t-2} + \cdots + \frac{1}{14}f_{t-13}}{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{14}} \tag{3.2}$$

where f_t represents the frequency on day t for a given query.

Figure 3.2 shows the original search frequency of the query "flu medicine" and the data after smoothing.

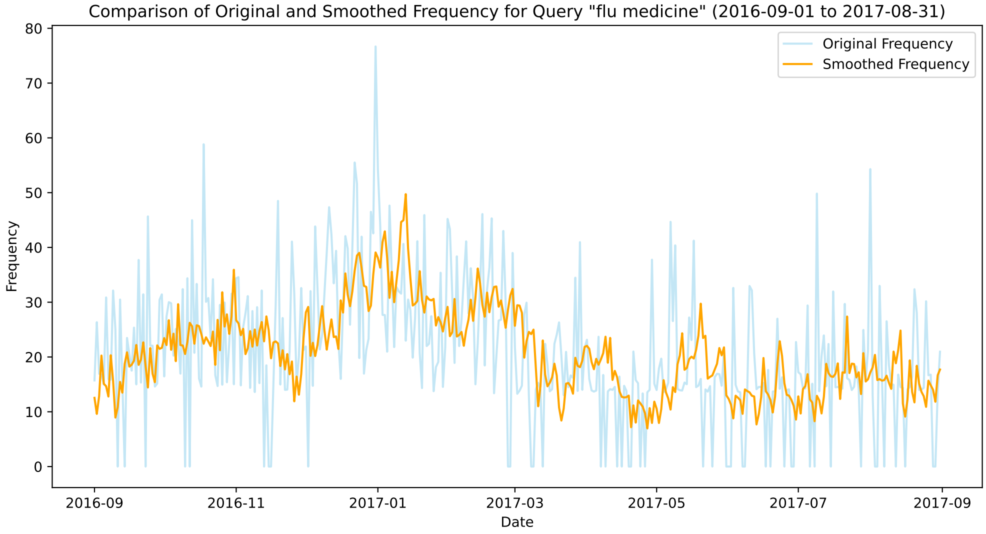

Figure 3.2: The search frequencies of "flu medicine" from September 1, 2016, to August 31, 2017.

3.3 Feedforward Neural Network

In this section, we will provide a detailed illustration of the neural network structure, the training and testing data split, and the validation strategy settings.

3.3.1 Definition of Feedforward Neural Network

A feedforward neural network (FFNN) is one of the most fundamental architectures in artificial neural networks. In this network, data flows from the input layer through one or more hidden layers, and finally reaches the output layer. The data flow in an FFNN is unidirectional, meaning there are no feedback connections, and information only propagates forward, forming a feedforward pattern. Figure 3.3 illustrates a three-layer FFNN, which is the model used in this project.

Structure

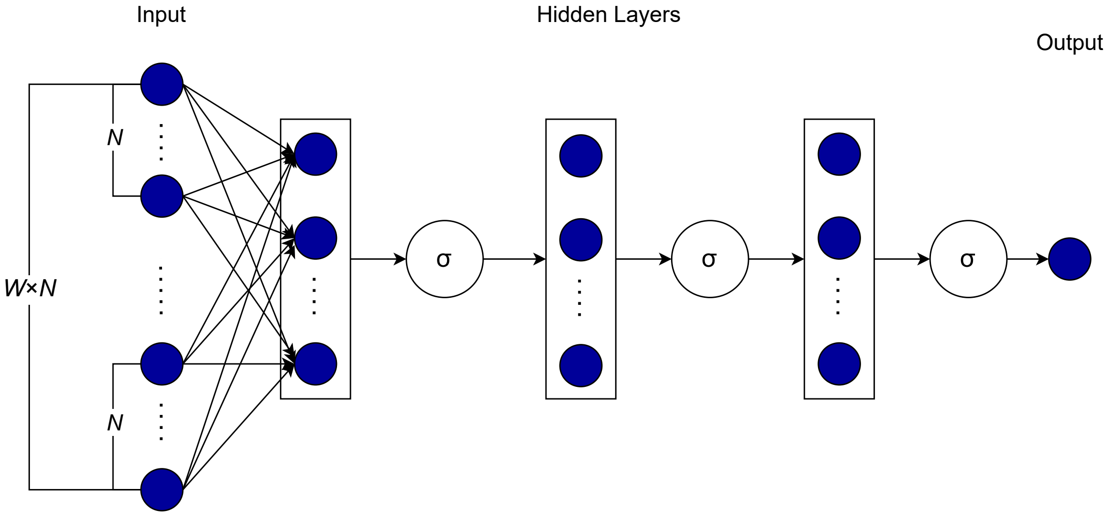

Figure 3.3: Computation graph of the FFNN model. In this figure, W represents the window size ($W \in \{7, 14\}$), and N denotes the number of selected top queries ($N \in \{100, 200, 300, 400\}$). The activation function σ is implemented using ReLU.

In an FFNN, the input signals start from the input layer and pass through the linear transformations and nonlinear activation functions of the hidden layers, until the output layer generates the final prediction. This process can be described using the following equations:

For the neurons in the l-th layer, the relationship between the input and output is given by:

$$\begin{aligned} \mathbf{z}^{(l)} &= \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \\ \mathbf{a}^{(l)} &= \sigma(\mathbf{z}^{(l)}) \end{aligned} \tag{3.3}$$

where $\mathbf{z}^{(l)}$ is the weighted input at the l-th layer, $\mathbf{W}^{(l)}$ is the weight matrix of the l-th layer, $\mathbf{b}^{(l)}$ is the bias, $\mathbf{a}^{(l)}$ is the activation value, and $\sigma(\cdot)$ is the activation function.

In an FFNN, the input layer receives external data, where each neuron represents a feature variable. Thus, the number of neurons in the input layer corresponds to the number of input features. The details of the input will be elaborated in Section 3.3.2.

The hidden layers are situated between the input and output layers, and an FFNN can have one or more hidden layers. Each hidden layer consists of several neurons that are connected to the neurons in the previous layer through a trainable weight matrix. The neurons in each hidden layer perform a weighted summation of the input signals, followed by a nonlinear transformation through an activation function, which enables the network to capture complex patterns in the data. In this project, I used three hidden layers and employed the ReLU function as the activation function, i.e., $\sigma(x) = \mathrm{ReLU}(x)$. The ReLU function is defined as follows:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{3.4}$$

The output layer is used to generate the final predictions. The number of neurons in the output layer typically corresponds to the number of classes for classification tasks or the dimensionality of the output for regression tasks. In my project, since the task is to nowcast or forecast the ILI rate 14 or 21 days ahead, the output consists of a single neuron. This neuron is connected to the last hidden layer through a ReLU activation function.

Training Process

In the training process of an FFNN, the core task is to minimize the loss function (such as mean squared error or cross-entropy; in this project, MAE Loss is used) to adjust the weights and biases of the network. The training steps are as follows:

First, the output of the network is computed by propagating the input data through each layer according to equations 3.3 and 3.4. The loss value $L(\mathbf{y}, \hat{\mathbf{y}})$, which quantifies the difference between the predicted and true values, is then calculated, where $\hat{\mathbf{y}}$ represents the network's predicted output and $\mathbf{y}$ represents the true labels. The function L can represent various types of loss functions, such as Mean Absolute Error (MAE) or Mean

Squared Error (MSE). In my project, I used the MAE loss, which is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.5}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of samples. MAE measures the average magnitude of errors in a set of predictions, without considering their direction.

Next, backpropagation is used to compute the gradients of the loss function with respect to each parameter, i.e., $\frac{\partial L}{\partial \mathbf{W}^{(l)}}$ and $\frac{\partial L}{\partial \mathbf{b}^{(l)}}$. These gradients are used to update the weights and biases according to the following rules:

$$\begin{aligned} \mathbf{W}^{(l)} &\leftarrow \mathbf{W}^{(l)} - \eta\frac{\partial L}{\partial \mathbf{W}^{(l)}} \\ \mathbf{b}^{(l)} &\leftarrow \mathbf{b}^{(l)} - \eta\frac{\partial L}{\partial \mathbf{b}^{(l)}} \end{aligned} \tag{3.6}$$

where η is the learning rate.

In my project, I employed the Adam optimizer [22], which is a variant of gradient descent with adaptive learning rates. The Adam optimizer combines the ideas of momentum and RMSProp, updating the weights and biases according to the following equations:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\nabla L(\mathbf{W}^{(l)}) \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)(\nabla L(\mathbf{W}^{(l)}))^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\ \mathbf{W}^{(l)} &\leftarrow \mathbf{W}^{(l)} - \frac{\eta \hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \end{aligned} \tag{3.7}$$

where β_1 and β_2 are the decay rates for the moving averages of the gradient and its square, respectively, and ϵ is a small constant added for numerical stability.

3.3.2 Implementation of Feedforward Neural Network

Input

We utilized the top 2000 query frequencies based on the methodology described in Section 3.2.2. To identify the most correlated queries, we calculated the correlations between query frequencies and ILI rates using the last five years of training data. From this, we selected the top N related queries, where $N \in \{100, 200, 300, 400\}$.

The 2009 H1N1 flu pandemic significantly influenced certain queries, such as "swine flu," resulting in high correlations with ILI rates during that period. However, including data from 2009 would disproportionately rank such queries high in correlation, despite their decreased relevance to recent flu trends. To address this, we excluded 2009 data and used only the most recent five years of training data to ensure the selected queries remain relevant to current flu trends. As shown in Table 3.1, this approach helps maintain the relevance of the identified queries.

The input vector $\mathbf{i}$ of the neural network consists of the top N query frequencies over the latest W days ($W \in \{7, 14\}$): $\mathbf{i} = [\mathbf{q}_t, \mathbf{q}_{t-1}, \ldots, \mathbf{q}_{t-W+1}]$, where each $\mathbf{q}_t$ represents the selected query frequencies for day t.

From Figure 3.2, the query searching frequency spans a wide range, which can lead to gradient vanishing or exploding issues in the neural network. Therefore, I scaled the training data using the min-max method. The scaling is performed using the following formula:

$$\mathbf{q}' = \frac{\mathbf{q} - \min(\mathbf{q})}{\max(\mathbf{q}) - \min(\mathbf{q})} \tag{3.8}$$

where $\mathbf{q}$ represents the original query frequencies and $\mathbf{q}'$ represents the scaled query frequencies. To prevent data leakage and improve the model's predictive accuracy and generalization ability, I used the minimum and maximum values obtained from the training data to scale the validation and test data. Specifically, the scaled data is obtained

Top 50 using All Training Data	**Top 50 using Last 5 Years**
'flu treatment', 'flu symptoms nhs', 'nhs flu symptoms', 'pregnancy and flu', 'flu pregnancy', 'pregnancy flu', 'flu and pregnancy', 'flu contagious', 'symptoms flu', 'flu symptoms', 'symptoms of flu', 'flu infection', 'flu symptoms uk', 'flu how long', 'how long flu', 'flu symptom', 'baby flu', 'treatment for flu', 'pregnant flu', 'incubation period for flu', 'flu pregnant', 'flu signs', 'flu in babies', 'the flu symptoms', 'flu in pregnancy', **'symptoms of swine flu'**, 'signs of flu', **'symptoms swine flu'**, **'swine flu symptoms'**, 'flu recovery', 'winter flu', 'flu in children', 'flu incubation period', 'i have flu', 'nhs direct flu', 'flu last', 'flu cough', **'signs of swine flu'**, 'flu pain', 'cough flu', 'flu incubation', 'flu contagious period', **'swine flu treatment'**, 'flu symptom checker', 'flu virus symptoms', 'flu symptoms in children', 'flu complications', 'flu symptoms last', 'cure flu', 'flu cure'	'flu symptoms', 'symptoms flu', 'flu how long', 'flu virus', 'virus flu', 'how long flu', 'symptoms of flu', 'flu last', 'pneumonia', 'cold cough', 'colds', 'how long does flu last', 'flu and cold', 'flu pain', 'flu contagious', 'cold and flu', 'cold virus', 'flu cold', 'cold flu', 'the flu symptoms', 'bad cold', 'cold and cough', 'cold', 'nhs cold', 'cold nhs', 'flu temperature', 'cough and cold', 'flu infection', 'i have cold', 'a cold', 'difference between cold and flu', 'cough flu', 'is flu contagious', 'flu cough', 'how long does the flu last', 'fever flu', 'flu fever', 'coughing', 'a cough', 'flu treatment', 'flu remedies', 'flu virus symptoms', 'cough medicine', 'flu incubation', 'i have a cold', 'get rid of flu', 'flu headache', 'got flu', 'influenza'

Table 3.1: Top 50 queries using 2016-2017 flu season training data vs. recent 5 years

using:

$$\mathbf{q}' = \frac{\mathbf{q} - \min(\mathbf{q}_{\text{train}})}{\max(\mathbf{q}_{\text{train}}) - \min(\mathbf{q}_{\text{train}})} \tag{3.9}$$

where $\mathbf{q}$ represents the original query frequencies, $\mathbf{q}'$ represents the scaled query frequencies, and $\min(\mathbf{q}_{\text{train}})$ and $\max(\mathbf{q}_{\text{train}})$ are the minimum and maximum values from the training data, respectively.

Model Training and Testing

Due to the influence of the COVID-19 pandemic starting in December 2019, the ILI rates during the pandemic are not accurate. Therefore, in this experiment, we use the flu seasons from 2016-2017, 2017-2018, and 2018-2019 (spanning from September 1st to August 31st of the following year) as the three test datasets. The training dataset includes data from September 1st, 2008, up to the day before the first test set begins.

We use an early stopping method to find the best validation loss during the training process. First, we train the model for 20 epochs, then we start the early stopping method with a patience of 5. To get the best validation loss, when the model stops at epoch n, we choose to use the model trained by epoch $n - 5$ with the best validation loss. The advantage of using early stopping is that it helps prevent overfitting by halting training when the model's performance on the validation set stops improving. This ensures that the model generalizes better to new, unseen data by selecting the point where it performs best on the validation set rather than continuing to train and potentially overfitting to the training data [23].

3.3.3 Validation Strategy

This section will outline the definitions of the onset, outset, and peak of a flu season, as well as the validation strategies for our models.

Onset, Outset, and Peak

To simplify the explanation and illustration, we will first define the validation period. For model validation, we select data from the last three years of the training dataset, considering these three years as the validation period.

Onset and outset refer to the start and end of each flu season, respectively. We use the mean and standard deviation of the five years preceding the validation period to estimate the onset and outset thresholds. At the beginning of the flu season, there may be fluctuations around the threshold. Therefore, we define the onset day as the first day when the ILI rates remain above the threshold for 14 consecutive days. The outset day is defined as the last day when the ILI rates cross the threshold for that flu season. The threshold is calculated using the equation:

$$threshold = mean - 0.25 \times std \tag{3.10}$$

where *std* stands for standard deviation. Both *mean* and *std* are calculated from the data of the five years preceding the validation period.

The peak is defined as the highest ILI rate during the flu season.

Using the methods described above, Figure 3.4 illustrates the onset and outset thresholds for each validation period, and Table 3.2 provides detailed information on the onset, peak, and outset dates for each test season.

Validation Strategies

This section details the validation strategies used in the experiments.

The goal of these validation strategies is to identify the best approach for validating hyperparameter choices, thereby improving the model's performance compared to random data selection. The days used for validation consists of 180 days, selected according to different strategies.

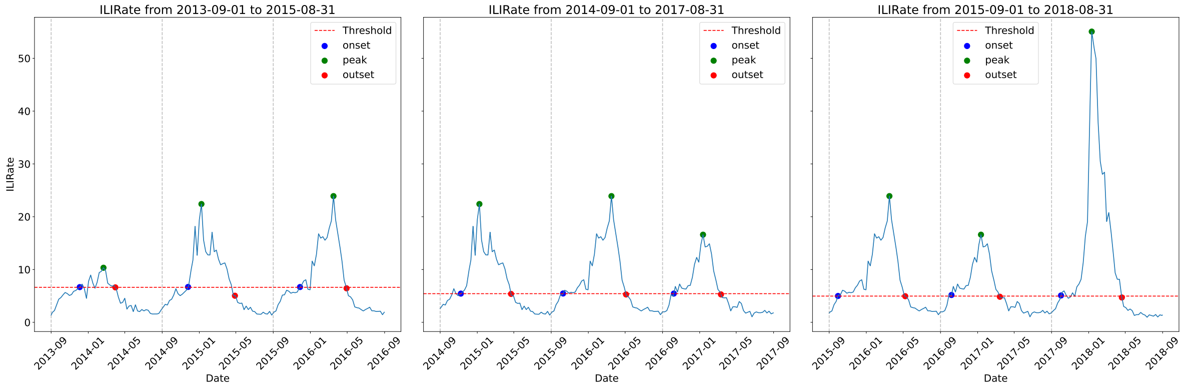

Figure 3.4: Thresholds for the validation period for each test flu season (September 1 to August 31 of 2016-2017, 2017-2018, and 2018-2019).

To clearly describe the validation methods, we define a 60-day window as the period encompassing the 29 days before, the day of, and the 30 days after a specified day.

Here are the six strategies used in this experiment:

1. Utilize the last 180 days of the training dataset.

2. For each of the three flu seasons in the validation period, use a 60-day window centered around the onset day.

3. Use a 60-day window centered around the peak day for each of the three flu seasons in the validation period.

4. Use a 60-day window centered around the outset day for each of the three flu seasons in the validation period.

5. For the validation period, consider the following specific windows for each flu season:

   - For the first flu season: use a 60-day window centered around the onset day.
   - For the second flu season: use a 60-day window centered around the peak day.
   - For the third flu season: use a 60-day window centered around the outset day.

Test Season	Onset Date	Peak Date	Outset Date
2016-2017	2016-09-01 2017-11-01 2018-10-01	2016-12-15 2018-01-15 2018-12-15	2017-02-28 2018-03-31 2019-02-28
2017-2018	2017-09-01 2018-11-01 2019-10-01	2017-12-15 2019-01-15 2019-12-15	2018-02-28 2019-03-31 2020-02-28
2018-2019	2018-09-01 2019-11-01 2020-10-01	2018-12-15 2020-01-15 2020-12-15	2019-02-28 2020-03-31 2021-02-28

Table 3.2: Detailed onset, peak, and outset dates of validation period for each test flu season.

6. For the validation period, consider the following alternative windows for each flu season:

    - For the first flu season: use a 60-day window centered around the outset day.
    - For the second flu season: use a 60-day window centered around the peak day.
    - For the third flu season: use a 60-day window centered around the onset day.

The reason most strategies use data around the onset day, outset day, or peak day is that these three days are the most critical in a flu season. The onset day marks the beginning of the increase in ILI rates and the start of the flu pandemic. The peak day is when the ILI rate reaches its highest value, indicating the peak of the flu activity. The outset day signifies the end of the flu pandemic as ILI rates start to level off. Using days around these key points helps capture the significant trends and variations in ILI rates, providing a robust basis for validation. This approach ensures that the model is effectively validated against the most relevant and impactful periods of flu activity.

Validated Hyperparameters

We provided various choices of hyperparameters for the model to validate and select the best one for the test flu season. Hyperparameters such as the number of units in the neural network layers, learning rate, and batch size were optimized using Grid Search. Grid Search is a systematic method for hyperparameter optimization, aimed at finding the best combination of hyperparameters within a defined search space. Specifically, Grid Search exhaustively searches through all possible values of each hyperparameter to form a "grid" of combinations. It then trains the model for each combination and evaluates its performance. Ultimately, Grid Search selects the combination that yields the best performance on the validation set or through cross-validation as the final model configuration.

Additionally, we evaluated the model's performance across different top N queries for input and various window sizes W for input days. The detailed choices are listed in Table 3.3.

Hyperparameter	**Choices**
$units\_1$	$[50, 100]$
$units\_2$	$[25, 50]$
$units\_3$	$[25, 50]$
$learning\_rate$	$[1e-4, 1e-5]$
$batch\_size$	$[14, 28, 56]$
N	$[100, 200, 300, 400]$
W	$[7, 14]$

Table 3.3: Choices of Hyperparameters. $units\_n$ denotes the number of units in layer n of the feedforward neural network. N represents the top N related queries based on the correlations mentioned in Section 3.2.2. W denotes the use of the latest W days' top N queries as input.

4 | Results and Analysis

This chapter presents and analyzes the experimental results for nowcasting and forecasting the ILI rate 14 and 21 days ahead. The experiments were conducted using ten different random seeds {1024, 2048, 4096, 8192, 16384, 15510, 1854, 18296, 16652, 11085}. The analysis begins with a comparison of the performance of different validation strategies against a baseline model. The baseline model uses default parameters, including the search frequencies of the top 200 queries, a window size of 14 days, 100 units in the first layer, 50 units in the second layer, 50 units in the third layer, a learning rate of 1e-4, and a batch size of 28. Given that previous studies [21, 20] have used the last several days of the training set as the validation set, I adopt a similar approach. Specifically, I use the first validation strategy (using the last 180 days of the training data) as a benchmark to compare with other validation strategies. Ultimately, my goal is to identify the optimal validation strategy among all the strategies considered.

4.1 Nowcasting

The detailed results of the nowcasting experiment are presented in Table 4.1. In this table, MAE and ρ represent the averages of the best validation loss models across 10 different random seeds. The "Baseline" in Table 4.1 refers to the model with default parameters. Analyzing the results, it is evident that the Baseline model performs the worst across all three flu seasons. In the 2016-17 flu season, the best MAE achieved with a validation strategy is 1.475, while the Baseline model's best performance is 2.196, which is 48.9% higher than the best validation result. In the 2017-18 flu season, the best MAE using a validation strategy is 2.862, compared to 3.642 for the Baseline model, indicating an increase of 27.3%. Similarly, for the 2018-19 flu season, the best MAE with a validation strategy is 1.883, while the Baseline model achieves 3.366, which is 78.8% higher than the best validation result.

Strategy	**16-17 Flu Season**		**17-18 Flu Season**		**18-19 Flu Season**	
	MAE	ρ	MAE	ρ	MAE	ρ
Baseline	2.196	0.951	3.642	0.937	3.366	0.890
1	1.674	0.952	**2.862**	0.960	2.701	0.903
2	**1.475**	**0.968**	3.182	0.945	2.334	0.924
3	1.567	0.954	3.458	0.968	2.252	0.918
4	1.759	0.954	3.139	0.955	2.919	0.884
5	1.928	0.952	3.365	0.946	2.303	0.916
6	1.743	0.954	3.021	**0.972**	**1.883**	**0.937**

Table 4.1: Detailed Performance of Different Validation Strategies and Window Sizes Across Three Flu Seasons. The table shows the mean absolute error (MAE) and correlation (ρ) for each validation strategy in the 16-17, 17-18, and 18-19 flu seasons for the nowcasting task. The best performance in each season is highlighted in bold and with a green background.

Then, I calculated the average performance over three flu seasons for each strategy and the baseline model. The results are presented in Table 4.2. In this table, MAE and ρ represent the averages of the best validation loss models across 10 different random seeds, further averaged over three test flu seasons. It is evident from the table that Strategy 6 performed the best, achieving the lowest MAE and the highest correlation. Additionally, the baseline model shows the worst performance overall, highlighting the critical role of selecting an effective validation strategy for identifying the best model for the nowcasting task. In the following section, I will conduct a detailed comparison and ranking of each validation strategy across the three flu seasons.

4.1.1 Ranking of Validation Strategies

In this section, I will present a ranking of the validation strategies used for each test flu season. The ranking is primarily based on MAE loss. In cases where strategies have

Strategy	**Average MAE**	**Average ρ**
Baseline	3.068	0.926
1	2.399	0.940
2	2.330	0.946
3	2.426	0.947
4	2.426	0.931
5	2.532	0.938
6	**2.216**	**0.954**

Table 4.2: Average Performance of 3 Flu Seasons Across Different Validation Strategies for Nowcasting Task

similar MAE loss, the average correlation is used as a secondary criterion. If the null hypothesis is that the difference in correlation is not significant, and the p-value supports this, the strategies are considered tied. If the difference in correlation is significant, the strategies are ranked accordingly.

2016-2017 Flu Season

Based on the observations from Table 4.1, Strategy 2 achieved the best MAE (Mean Absolute Error) and the highest correlation during the 2016-17 flu season. Compared to Strategy 2, Strategy 1's MAE was 13.5% worse than the best MAE, and its correlation was 0.017 lower than the best correlation.

To further evaluate the performance of each validation strategy, a comparison and ranking were conducted across different flu seasons. The ranking, based on MAE loss, is as follows: from best to worst, S2, S3, S1, S6, S4, and S5. Since the experiments were conducted using ten random seeds, a t-test was applied to the test MAE loss results to statistically compare the strategies and assess their degree of similarity. The results of the t-tests, summarized in Table 4.3, show the statistical significance of the pairwise differences in

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
S1	-	0.244	0.507	0.641	0.089	0.622
S2	0.244	-	0.559	0.122	0.004	0.060
S3	0.507	0.559	-	0.269	0.012	0.173
S4	0.641	0.122	0.269	-	0.281	0.918
S5	0.089	0.004	0.012	0.281	-	0.086
S6	0.622	0.060	0.173	0.918	0.086	-

Table 4.3: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2016-17 flu season for nowcasting task. Red indicates significant differences ($p < 0.05$).

MAE loss between the strategies. Overall, only Strategy 2 and Strategy 5 exhibited significant differences in MAE loss, while the other strategies showed high similarity. Therefore, based on the MAE loss, Strategy 2 should be ranked first, and Strategy 5 should be ranked last. The remaining strategies are ranked based on their correlation values.

Thus, the final ranking of the validation strategies for the nowcasting task in the 2016-17 flu season is as follows:

1. S2
2. S3, S4 and S6 (tied)
3. S1
4. S5

2017-2018 Flu Season

In Table 4.1, we observe that Strategy 1 achieved the lowest MAE loss, with a value of 2.862, while Validation Strategy 6 achieved the best correlation, reaching 0.972. The results of the t-tests, presented in Table 4.4, indicate that there are significant differences

Strategy	S1	S2	S3	S4	S5	S6
S1	-	0.053	0.035	0.298	0.024	0.398
S2	0.053	-	0.306	0.869	0.385	0.391
S3	0.035	0.306	-	0.352	0.755	0.135
S4	0.298	0.869	0.352	-	0.448	0.674
S5	0.024	0.385	0.755	0.448	-	0.147
S6	0.398	0.391	0.135	0.674	0.147	-

Table 4.4: Pairwise p-values from t-tests between six different strategies for the 2017-18 flu season based on MAE loss for nowcasting task. Red indicates significant differences ($p < 0.05$).

between Strategy 1 and Strategies 3 and 5. Therefore, Strategy 1 is ranked first. Since there is no significant difference between Strategies 3 and 5, they are ranked based on their correlation values. Strategy 3 has a correlation of 0.968, while Strategy 5 has a correlation of 0.946, placing Strategy 3 second to last and Strategy 5 in the last position.

Next, the remaining strategies are ranked based on their correlation values, resulting in the following ranking for the performance of each validation strategy in the 2017-18 flu season:

1. Strategy 1
2. Strategy 6
3. Strategy 4
4. Strategy 2
5. Strategy 3
6. Strategy 5

2018-1019 Flu Season

Firstly, by analyzing Table 4.1, we observe that Strategy 6 performed the best, with the lowest MAE loss of 1.883 and a correlation of 0.937. Its MAE was 30.2% better than Strategy 1, and its correlation was 0.034 higher than that of Strategy 1.

The t-test results, shown in Table 4.5, indicate that there are significant differences between Strategy 6 and Strategies 1, 2, and 4. Furthermore, Strategy 4 shows significant differences with Strategies 2, 3, 5, and 6. Based on their average MAE loss, we conclude that Strategy 6 had the best performance, while Strategy 4 performed the worst, and Strategy 1 ranked second to last, with Strategy 2 ranking third to last.

For the remaining strategies, we rank them based on their correlation values, placing Strategy 3 ahead of Strategy 5. Therefore, the final ranking is as follows:

1. Strategy 6
2. Strategy 3
3. Strategy 5
4. Strategy 2
5. Strategy 1
6. Strategy 4

Summary

Based on the above rankings, we can conclude that Strategy 6 performed the best, consistently ranking in the top two across all three flu seasons. In contrast, Strategy 5 showed the weakest performance, ranking last in two of the seasons. The performance of the other strategies was inconsistent and varied across the different flu seasons.

Strategy	S1	S2	S3	S4	S5	S6
S1	-	0.068	0.060	0.327	0.087	6.85e-05
S2	0.068	-	0.749	0.025	0.904	0.033
S3	0.060	0.749	-	0.022	0.855	0.123
S4	0.327	0.025	0.022	-	0.031	1.32e-4
S5	0.087	0.904	0.855	0.031	-	0.078
S6	6.85e-05	0.033	0.123	1.32e-4	0.078	-

Table 4.5: Pairwise p-values from t-tests between six different strategies for the 2018-19 flu season based on MAE loss for nowcasting task. Red indicates significant differences ($p < 0.05$).

4.1.2 Parameter Selection Across Different Flu Seasons

In Table 4.6, we present the average hyperparameters selected across different flu seasons for the nowcasting task. The table shows that, except for the query number (QN), other parameters such as batch size (BS), learning rate (LR), and the number of units in the neural network layers remain relatively stable across the flu seasons.

For the 17-18 flu season, there is a noticeable increase in the number of queries compared to the 16-17 flu season and 18-19 flu season. Upon examining Figure 3.1, we observe that the flu peak during the 17-18 season was significantly higher than in the other seasons. This unusually high peak suggests increased variability during that period. The elevated peak and greater fluctuations likely required the model to capture more complex patterns, which is why a higher number of queries was necessary to accurately reflect the dynamic trends in flu transmission. Additionally, the increase in query numbers could be attributed to the availability of more comprehensive and higher-quality data during the 17-18 flu season.

Season	W	**BS**	**LR**	**Units1**	**Units2**	**Units3**	N
16-17	11.67	38.50	7.8e-5	70.83	38.75	40.00	265.00
17-18	10.61	37.57	8.4e-5	78.33	39.17	34.17	320.00
18-19	9.45	36.17	9.2e-5	74.17	40.42	40.42	246.67

Table 4.6: Averaged Hyperparameters by Season for nowcasting task. WS stands for window size, BS for batch size, LR for learning rate, Units1/2/3 represent the number of units in the first, second, and third layers, and QN for the query number.

4.2 Forecasting 14 Days Ahead

In this section, I will conduct a comparative analysis between the models using validation strategies and the baseline model. Additionally, I will rank the performance of these models on the test flu season based on different validation strategies and analyze the parameter selection.

Table 4.7 presents the average of the best performances for each validation strategy across different random seeds. From this table, we can observe that Strategy 3 performed the best during the 16-17 flu season, with a minimum MAE loss of 2.204, which is 19.0% better than the Baseline model. Additionally, it achieved a correlation of 0.890, which is 0.012 higher than the Baseline model.

However, during the 17-18 flu season, we see that the Baseline model significantly outperforms all validation strategies. Its MAE loss is 21.8% lower than that of Strategy 5, which had the best MAE loss, and its correlation is 0.085 higher than that of Strategy 6, which had the best correlation. This discrepancy could be due to the higher peak in the 17-18 season, where the parameter set that performed best on the validation set did not generalize well to the 17-18 test flu season, resulting in the model missing the optimal performance for that season.

For the 18-19 flu season, Strategy 3 had the best MAE loss of 2.952. In contrast, the Baseline model performed poorly in this season, with an MAE loss 2.595 times higher than that of Strategy 3. Its correlation was also lower than that of Strategy 2 by 0.052.

Based on these results, we can conclude that although the Baseline model performed better than all models selected through validation strategies during the 17-18 flu season, its performance fluctuated significantly. In forecasting tasks, while using a validation strategy might not always yield the best-performing model on the test set, it helps ensure prediction stability across most seasons, especially those with smoother trends. Only during volatile seasons, like 17-18, might the performance deviate.

Strategy	**16-17 Flu Season**		**17-18 Flu Season**		**18-19 Flu Season**	
	MAE	ρ	MAE	ρ	MAE	ρ
Baseline	2.721	0.878	**3.665**	**0.938**	7.659	0.841
1	2.349	0.864	5.025	0.817	3.595	0.727
2	2.360	0.886	5.184	0.813	3.780	**0.893**
3	**2.204**	**0.890**	4.896	0.850	**2.952**	0.840
4	3.376	0.845	4.833	0.842	3.513	0.830
5	2.468	0.878	4.688	0.842	3.563	0.853
6	2.339	0.875	4.923	0.853	3.061	0.866

Table 4.7: Detailed Performance of Different Validation Strategies and Window Sizes Across Three Flu Seasons. The table shows the mean absolute error (MAE) and correlation (ρ) for each validation strategy in the 16-17, 17-18, and 18-19 flu seasons for forecasting 14 days ahead task. The best performance in each season is highlighted in bold.

Table 4.8 presents the average performance of each strategy across different flu seasons, as summarized from Table 4.7. From this, we can observe that Strategy 3 achieved the best average MAE, while the Baseline model exhibited the highest correlation, primarily due to its outstanding performance during the 17-18 flu season. In the following sections, I will analyze the performance of various validation strategies across different flu seasons and rank them accordingly.

Strategy	**Average MAE**	**Average ρ**
Baseline	4.682	**0.886**
1	3.656	0.803
2	3.775	0.864
3	**3.351**	0.860
4	3.907	0.847
5	3.573	0.858
6	3.441	0.865

Table 4.8: Average Performance Across Different Validation Strategies for Forecasting 14 Days ahead Task

4.2.1 Ranking of Validation Strategies

2016-2017 Flu Season

By examining Table 4.9, we can see that Strategy 4 has a significant gap compared to the other strategies. Combined with the performance of Strategy 4 in the 16-17 flu season shown in Table 4.7, it is clear that Strategy 4 is the worst validation strategy. The remaining validation strategies show minimal differences in MAE loss, so we rank them based on their correlation. The ranking of strategies for predicting the ILI rate 14 days ahead during the 16-17 flu season, from best to worst, is as follows:

1. Strategy 3
2. Strategy 2
3. Strategy 5
4. Strategy 6
5. Strategy 1

Strategy	S1	S2	S3	S4	S5	S6
S1	-	0.944	0.310	0.001	0.544	0.950
S2	0.944	-	0.363	0.002	0.623	0.908
S3	0.310	0.363	-	2.5e-4	0.196	0.423
S4	0.001	0.002	2.5e-4	-	0.006	0.001
S5	0.544	0.623	0.196	0.006	-	0.550
S6	0.950	0.908	0.423	0.001	0.550	-

Table 4.9: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2016-17 flu season for forecasting 14 days ahead task. Red indicates significant differences ($p < 0.05$).

6. Strategy 4

Although the MAE loss difference between Strategy 1 and Strategy 3 (which has the best MAE) is small, Strategy 1's correlation is 0.026 lower than the best MAE loss achieved by Strategy 3.

2017-2018 Flu Season

By examining Table 4.10, we observe a significant gap between Strategy 2 and Strategy 5. Further analysis of Table 4.7 reveals that Strategy 2 has the worst MAE loss, while Strategy 5 achieves the best MAE loss. Thus, Strategy 5 is identified as the best validation strategy, and Strategy 2 as the worst. The remaining validation strategies show little difference in MAE loss, so they are ranked based on their correlation. The resulting ranking is as follows:

1. Strategy 5

2. Strategy 6

3. Strategy 3

4. Strategy 4

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
S1	-	0.502	0.655	0.461	0.158	0.698
S2	0.502	-	0.269	0.127	0.017	0.261
S3	0.655	0.269	-	0.819	0.415	0.926
S4	0.461	0.127	0.819	-	0.513	0.724
S5	0.158	0.017	0.415	0.513	-	0.307
S6	0.698	0.261	0.926	0.724	0.307	-

Table 4.10: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2017-18 flu season for forecasting 14 days ahead task. Red indicates significant differences ($p < 0.05$).

5. Strategy 1

6. Strategy 2

Although the MAE loss difference between Strategy 1 and Strategy 5 is not large in Table 4.10, the correlation of Strategy 5 is 0.025 higher than that of Strategy 1.

2018-2019 Flu Season

By examining Table 4.11, we observe a significant gap in MAE loss between Strategy 1 and both Strategy 3 and Strategy 6, as well as between Strategy 2 and Strategy 3. Based on Table 4.7 and the information above, we can conclude that Strategy 3 has the best MAE loss, while Strategy 2 has the worst MAE loss. Therefore, Strategy 2 is identified as the worst validation strategy, and Strategy 3 as the best. Since Strategy 1 shows a considerable gap compared to Strategy 3 and Strategy 6, we can infer that Strategy 1 is the second-worst, while Strategy 6 is the second-best validation strategy. As the remaining strategies show no significant differences in MAE loss, they are ranked according to their correlation. The resulting order is as follows:

1. Strategy 3

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
S1	-	0.559	0.008	0.829	0.918	0.036
S2	0.559	-	0.023	0.540	0.562	0.051
S3	0.008	0.023	-	0.150	0.054	0.641
S4	0.829	0.540	0.150	-	0.907	0.255
S5	0.918	0.562	0.054	0.907	-	0.124
S6	0.036	0.051	0.641	0.255	0.124	-

Table 4.11: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2018-19 flu season for forecasting 14 days ahead task. Red indicates significant differences ($p < 0.05$).

2. Strategy 6
3. Strategy 5
4. Strategy 4
5. Strategy 1
6. Strategy 2

Comparing Strategy 1 with the best-performing Strategy 3, we find that Strategy 3's MAE loss is 17.9% better, and its correlation is 0.113 higher than that of Strategy 1.

Summary

Overall, strategy 3 performs only slightly worse during the 17-18 flu season. However, based on Table 4.10, we can see that the difference in MAE loss between strategy 3 and the best model, strategy 5, is not statistically significant. Both strategy 5 and strategy 6 show relatively stable performance, frequently ranking in the top three. Therefore, it can be concluded that strategy 3 is the optimal model. Additionally, we observe that only during the 16-17 flu season does strategy 2 rank relatively high, whereas the performances of strategy 1, strategy 4, and strategy 2 are generally poor, often appearing in the bottom three.

4.2.2 Parameter Selection Across Different Flu Seasons

Table 4.12 presents the average parameters selected by different strategies across various flu seasons under different random seeds. By comparing this table with the one from the nowcasting task, Table 4.6, we can observe that both tend to select larger window sizes. A larger window size can capture more historical information, which is crucial for prediction tasks because it allows the model to detect long-term trends and patterns. Specifically, a larger window size includes the search frequencies of top N selected queries over more days, enabling the model to more accurately identify potential changes in disease transmission and thus improve prediction accuracy.

Season	W	**BS**	**LR**	**Units1**	**Units2**	**Units3**	N
16-17	12.71	36.40	7.45e-5	72.03	38.14	40.68	235
17-18	13.07	46.20	8.05e-5	79.17	35.83	33.75	305
18-19	11.20	26.13	8.80e-5	78.33	41.67	40.00	280

Table 4.12: Averaged Hyperparameters by Season for forecasting 14 days ahead task. W stands for window size, BS for batch size, LR for learning rate, Units1/2/3 represent the number of units in the first, second, and third layers, and N for the query number.

We can also observe that during the 2017-2018 flu season, the number of top N selected queries is significantly higher than in the 2016-2017 season, while only slightly higher than in the 2018-2019 season. This phenomenon is consistent with the findings from the nowcasting task, further indicating that the 2017-2018 flu season exhibited greater variability in flu transmission, which required more queries to capture the complex dynamic changes.

4.3 Forecasting 21 Days Ahead

Table 4.13 presents the average performance of the baseline model and various strategies across 10 random seeds for predicting the ILI rate 21 days ahead. We can observe once again that during the 17-18 flu season, the baseline model performs the best, with an MAE loss that is 31.5% higher than strategy 5, which had the lowest MAE loss, and a correlation of 0.942. However, the baseline model performs poorly in the 18-19 flu season, with an MAE loss 4.91 times higher than that of strategy 3, which had the lowest MAE loss. In the 16-17 flu season, strategy 3, which had the best MAE loss, outperformed the baseline model by 25.7%.

Strategy	**16-17 Flu Season**		**17-18 Flu Season**		**18-19 Flu Season**	
	MAE	ρ	MAE	ρ	MAE	ρ
Baseline	2.965	0.856	**3.667**	**0.942**	14.651	0.705
1	2.410	0.815	5.643	0.765	3.448	0.788
2	3.305	**0.876**	5.885	0.723	7.224	0.911
3	**2.202**	0.839	5.800	0.760	**2.981**	0.851
4	2.559	0.807	5.561	0.749	3.571	0.811
5	2.428	0.845	5.350	0.763	3.810	0.843
6	2.243	0.814	5.665	0.805	4.565	**0.913**

Table 4.13: Detailed Performance of Different Validation Strategies and Window Sizes Across Three Flu Seasons. The table shows the mean absolute error (MAE) and correlation (ρ) for each validation strategy in the 16-17, 17-18, and 18-19 flu seasons for forecasting 21 days ahead task. The best performance in each season is highlighted in bold.

Table 4.14 presents the Average Performance Across Different Validation Strategies. We can observe that strategy 3 has the best average MAE loss of 3.661, which is 48.4% better than the baseline model. Strategy 6 shows the highest correlation of 0.844, which is 0.010

higher than the baseline model.

Strategy	**Average MAE**	**Average ρ**
Baseline	7.094	0.834
1	3.834	0.789
2	4.213	0.837
3	**3.661**	0.817
4	3.897	0.789
5	3.863	0.817
6	4.158	**0.844**

Table 4.14: Average Performance Across Different Validation Strategies for Forecasting 21 Days ahead Task

4.3.1 Ranking of Validation Strategies

2016-2017 Flu Season

By examining Table 4.15, we can determine that the performance of strategy 2 is significantly different from that of all other strategies. Additionally, from Table 4.13, we observe that strategy 2 has the worst MAE loss, which leads to the conclusion that strategy 2 should be ranked last. Since there is no significant difference in MAE loss among the remaining validation strategies, they are ranked based on their correlation values. The ranking from best to worst is as follows:

1. Strategy 5
2. Strategy 3
3. Strategy 1
4. Strategy 6

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
S1	-	9.59e-5	0.130	0.437	0.922	0.320
S2	9.59e-5	-	1.83e-6	0.002	4.16e-4	2.14e-5
S3	0.130	1.83e-6	-	0.048	0.179	0.771
S4	0.437	0.002	0.048	-	0.541	0.121
S5	0.922	4.16e-4	0.179	0.541	-	0.341
S6	0.320	2.14e-5	0.771	0.121	0.341	-

Table 4.15: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2016-17 flu season for forecasting 21 days ahead task. Red indicates significant differences ($p < 0.05$).

5. Strategy 4
6. Strategy 2

We can also observe that while strategy 1 has a similar MAE loss to the other strategies (except strategy 2), it is ranked lower due to its relatively poor correlation, which is 0.030 lower than strategy 5.

2017-2018 Flu Season

In Table 4.16, we can observe that the MAE loss of strategy 5 differs significantly from that of strategy 2 and strategy 3. From Table 4.13, we see that strategy 5 has the best MAE loss, while the performance of strategy 2 and strategy 3 is comparatively poor. Therefore, strategy 5 is ranked first. Since the difference in MAE loss between strategy 2 and strategy 3 is not significant, we rank them based on their correlation. Strategy 3 has a higher correlation than strategy 2, making strategy 2 the last and strategy 3 the second to last. The remaining strategies are then ranked based on their correlation, resulting in the following ranking from best to worst:

1. Strategy 5

Strategy	S1	S2	S3	S4	S5	S6
S1	-	0.288	0.541	0.714	0.154	0.924
S2	0.288	-	0.725	0.123	0.007	0.309
S3	0.541	0.725	-	0.321	0.049	0.584
S4	0.714	0.123	0.321	-	0.239	0.624
S5	0.154	0.007	0.049	0.239	-	0.104
S6	0.924	0.309	0.584	0.624	0.104	-

Table 4.16: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2017-18 flu season for forecasting 21 days ahead task. Red indicates significant differences ($p < 0.05$).

2. Strategy 6
3. Strategy 1
4. Strategy 4
5. Strategy 3
6. Strategy 2

Here, the MAE loss of strategy 1 is not significantly different from that of strategy 5, which has the best MAE loss, and the correlation difference between them is also small. However, strategy 1's correlation is 0.040 lower than that of strategy 6, which has the highest correlation.

2018-2019 Flu Season

Here, we can observe that strategy 2 has significant differences compared to all other strategies, and strategy 6 shows significant differences with strategy 1 and strategy 3. By examining Table 4.13, we can see that strategy 2 has the worst MAE loss, so it should be ranked last. Additionally, strategy 6 shows significant differences with strategy 1 and strategy 3, and based on their performance, strategy 6 performs worse, while strategy 1

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
S1	-	1.33e-9	0.099	0.766	0.365	0.003
S2	1.33e-9	-	8.44e-9	1.54e-6	2.69e-6	1.96e-5
S3	0.099	8.44e-9	-	0.237	0.093	0.001
S4	0.766	1.54e-6	0.237	-	0.674	0.071
S5	0.365	2.69e-6	0.093	0.674	-	0.153
S6	0.003	1.96e-5	0.001	0.071	0.153	-

Table 4.17: Pairwise p-values from t-tests between six different strategies based on test MAE loss for the 2018-19 flu season for forecasting 21 days ahead task. Red indicates significant differences ($p < 0.05$).

and strategy 3 perform better. Therefore, strategy 6 should be ranked second to last. Since there is no significant difference in MAE loss between strategy 1 and strategy 3, we rank them based on their average correlation, with strategy 3 ranked first and strategy 1 second. As the remaining strategies do not show significant differences in MAE loss, they are ranked according to their correlation. The final ranking from best to worst is as follows:

1. Strategy 3
2. Strategy 1
3. Strategy 5
4. Strategy 4
5. Strategy 6
6. Strategy 2

Although there is no significant difference in MAE loss between strategy 1 and the best-performing strategy 3, strategy 1's correlation is 0.063 lower than that of strategy 3.

Summary

In the task of predicting the ILI rate 21 days ahead, we can observe that strategy 5 consistently ranks in the top three, securing first place in both the 16-17 and 17-18 flu seasons. Strategy 3 only ranks fifth during the 17-18 flu season, but it ranks second in the 16-17 flu season and first in the 18-19 flu season.

4.3.2 Parameter Selection Across Different Flu Seasons

In Table 4.18, we observe that both the 16-17 and 18-19 flu seasons tend to favor larger window sizes (with a window size of 14). The 18-19 season shows a more balanced selection, slightly favoring a window size of 14. However, the average query numbers for the 16-17 flu season are slightly higher than for the 17-18 flu season, with both seasons showing values around 300.

Season	**WS**	**BS**	**LR**	**Units1**	**Units2**	**Units3**	**QN**
16-17	13.53	46.20	9.25e-05	74.17	39.58	40.00	315.00
17-18	13.18	45.50	8.95e-05	73.33	35.42	36.25	281.67
18-19	10.85	33.13	8.35e-05	76.67	38.33	40.00	258.33

Table 4.18: Averaged Hyperparameters by Season for forecasting 21 days ahead task. WS stands for window size, BS for batch size, LR for learning rate, Units1/2/3 represent the number of units in the first, second, and third layers, and QN for the query number.

5 | Discussion

5.1 Main Findings

1. In the nowcasting task, all validation strategies outperform the baseline model. However, in the forecasting tasks for 14 days ahead and 21 days ahead, the baseline model performs better than the models using validation strategies during the 17-18 flu season. This may be due to the unusually high peak in the 17-18 flu season, where models that performed best on the validation set underperformed on the test set compared to the discarded combinations.

2. In the nowcasting task, validation strategy 3 is the best-performing strategy. For the forecasting tasks of 14 days ahead and 21 days ahead, strategies 3, 5, and 6 exhibit strong performance, while other strategies are slightly less effective.

3. In the forecasting tasks, each tested flu season tends to favor a larger window size.

4. The 17-18 flu season selects a higher number of average queries, likely due to the higher peak during that season. Using more frequent queries provides additional information for the model to learn from.

5.2 Limitations and Future Work

Due to time and computational resource constraints, the parameter selection was relatively limited. Each model should be given more parameter options to further optimize its performance. In terms of validation strategies, it might be beneficial to move beyond the conventional approach of selecting 180 days from the last three years of the training set. For instance, selecting a total of 365 days from the last five years could provide a more robust validation. Additionally, the composition of the validation strategy could be more varied. More combinations of onset, peak, and outset periods could be considered to validate the models.

Additionally, we observed that the best-performing validation strategy in the nowcasting task, strategy 6, and the well-performing strategies in the forecasting task, strategies 3 and 5, all included one or more peak periods from the flu seasons. In contrast, the less effective strategies did not use these peak periods. This suggests that using peak periods as part of the validation set may have a positive impact on model performance. Therefore, exploring more combinations of periods, particularly those involving the peak period, could provide valuable insights into how different phases of the flu season affect the results.

This project utilized a Feed Forward Neural Network (FFNN), but the model scope could be expanded to include other models, such as Long Short-Term Memory (LSTM) networks. This would allow for an examination of whether the choice of validation strategy is consistent across different models and tasks, or if different models require different validation strategies to achieve optimal performance.

When forecasting the 17-18 flu season, the baseline model performed better. This raises the question of how to handle special cases like the 17-18 flu season, where the model with the lowest validation loss does not necessarily represent the best-performing model. Moving forward, alternative strategies should be considered to ensure robustness in such scenarios.

6 | Conclusion

The primary focus of this project was to examine how different validation sets, composed of various time periods, influence parameter selection in a feedforward neural network (FFNN) for predicting ILI rates using web search data. This study demonstrated that the choice of validation set can lead to variations in the optimal parameter combinations, ultimately affecting the model's performance.

In this study, we developed a three-layer feedforward neural network (FFNN) and used validation sets composed of 60-day windows of the onset, outset, or peak periods from three years of data to perform grid search for parameter selection. These parameters were then used to predict and nowcast ILI rates for the next 14 and 21 days. In the nowcasting task, the model performed best when the validation set was composed of 60 days around the outset in the first year, 60 days around the peak in the second year, and 60 days around the onset in the third year of the last three years of training data. The next best performance was achieved using a validation set of 60 days around the onset in each of the last three years.

We found that, compared to the baseline model, the models selected by validation strategies outperformed the baseline in all tasks except for the forecasting task during the 17-18 flu season. In this case, the larger peak in the 17-18 flu season caused the parameters selected based on the validation set to perform poorly on the test set. In the nowcasting task, the baseline model's MAE loss was 38.4% higher than that of the best-performing validation strategy, strategy 6. Strategy 6 demonstrated the best performance in the nowcasting task, with an MAE loss 8.2% lower than that of strategy 1. In the forecasting tasks for 14 and 21 days ahead, strategy 3 achieved the best average MAE across the three flu seasons, with MAE losses 9.1% and 4.7% lower than strategy 1 for 14 and 21 days ahead, respectively.

Future research could involve expanding the parameter options available in grid search to allow for greater flexibility and selection, leading to potentially better-performing models. Other parameter optimization methods, such as Bayesian optimization, could also be explored to find the optimal parameter combinations. Additionally, future studies could investigate how different types of neural networks perform with various validation strategies.

References

[1] World Health Organization, "Influenza (seasonal)." https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal), 2024. Accessed: 2024-08-08.

[2] G. Verikios, M. Sullivan, P. Stojanovski, J. A. Giesecke, and G. Woo, *The global economic effects of pandemic influenza*. Centre of Policy Studies (CoPS), 2011.

[3] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman, "Influenza forecasting with google flu trends," *PloS one*, vol. 8, no. 2, p. e56176, 2013.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[5] Q. Xu, Y. R. Gel, L. L. Ramirez Ramirez, K. Nezafati, Q. Zhang, and K.-L. Tsui, "Forecasting influenza in hong kong with google search queries and statistical model fusion," *PloS one*, vol. 12, no. 5, p. e0176690, 2017.

[6] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

[7] A. D. Iuliano, K. M. Roguski, H. H. Chang, D. J. Muscatello, R. Palekar, S. Tempia, C. Cohen, J. M. Gran, D. Schanzer, B. J. Cowling, *et al.*, "Estimates of global seasonal influenza-associated respiratory mortality: a modelling study," *The Lancet*, vol. 391, no. 10127, pp. 1285–1300, 2018.

[8] B. K. White, E. Wilhelm, A. Ishizumi, S. Abeyesekera, A. Pereira, B. Yau, A. Kuzmanovic, T. Nguyen, S. Briand, and T. D. Purnat, "Informing social media analysis

for public health: a cross-sectional survey of professionals," *Archives of Public Health*, vol. 82, no. 1, p. 1, 2024.

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

[10] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[11] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen, "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific reports*, vol. 5, no. 1, p. 12760, 2015.

[12] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLoS computational biology*, vol. 11, no. 10, p. e1004513, 2015.

[13] J. J. Szczygielski, A. Charteris, P. R. Bwanya, and J. Brzeszczyński, "Google search trends and stock markets: sentiment, attention or uncertainty?," *International Review of Financial Analysis*, vol. 91, p. 102549, 2024.

[14] G. Aoki, K. Ataka, T. Doi, and K. Tsubouchi, "Data-driven estimation of economic indicators with search big data in discontinuous situation," *The Journal of Finance and Data Science*, vol. 9, p. 100106, 2023.

[15] C. Prado-Román, R. Gómez-Martínez, and C. Orden-Cruz, "Google trends as a predictor of presidential elections: the united states versus canada," *American Behavioral Scientist*, vol. 65, no. 4, pp. 666–680, 2021.

[16] P. Kassraie, A. Modirshanechi, and H. K. Aghajan, "Election vote share prediction using a sentiment-based fusion of twitter data with google trends and online polls.," in *DATA*, pp. 363–370, 2017.

[17] V. Agarwal, L. Zhang, J. Zhu, S. Fang, T. Cheng, C. Hong, and N. H. Shah, "Impact of predicting health care utilization via web search behavior: a data-driven analysis," *Journal of medical Internet research*, vol. 18, no. 9, p. e251, 2016.

[18] V. Lampos, B. Zou, and I. J. Cox, "Enhancing feature selection using word embeddings: The case of flu surveillance," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 695–704, 2017.

[19] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. Del Valle, "Forecasting the 2013–2014 influenza season using wikipedia," *PLoS computational biology*, vol. 11, no. 5, p. e1004239, 2015.

[20] M. Morris, P. Hayes, I. J. Cox, and V. Lampos, "Neural network models for influenza forecasting with associated uncertainty using web search activity trends," *PLoS Computational Biology*, vol. 19, no. 8, p. e1011392, 2023.

[21] S. Wei, S. Lin, Z. Wenjing, S. Shaoxia, Y. Yuejie, H. Yujie, Z. Shu, L. Zhong, and L. Ti, "The prediction of influenza-like illness using national influenza surveillance data and baidu query data," *BMC Public Health*, vol. 24, no. 1, p. 513, 2024.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.

A | Source Code

Source code for all of the methods implemented in Chap. 3 for the project can be found in the GitHub repository:
https://github.com/yueling-16/Master_Final_Project.