sustainability

Article

Forecasting the Potential Number of Influenza-like Illness Cases by Fusing Internet Public Opinion

Yu-Chih Wei [1], Yan-Ling Ou [1], Jianqiang Li [2,*] and Wei-Chen Wu [3]

[1] Department of Information and Finance Management, National Taipei University of Technology, Taipei City 10608, Taiwan; vickrey@mail.ntut.edu.tw (Y.-C.W.); t109ab8016@ntut.org.tw (Y.-L.O.)

[2] Faculty of Information, Beijing University of Technology, Beijing 100124, China

[3] Department of Finance, National Taipei University of Business, Taipei City 10051, Taiwan; weichen@ntub.edu.tw

* Correspondence: lijianqiang@bjut.edu.cn

Abstract: As influenza viruses mutate rapidly, a prediction model for potential outbreaks of influenza-like illnesses helps detect the spread of the illnesses in real time. In order to create a better prediction model, in this study, in addition to using the traditional hydrological and atmospheric data, features, such as popular search keywords on Google Trends, public holiday information, population density, air quality indices, and the numbers of COVID-19 confirmed cases, were also used to train the model in this research. Furthermore, Random Forest and XGBoost were combined and used in the proposed prediction model to increase the prediction accuracy. The training data used in this research were the historical data taken from 2016 to 2021. In our experiments, different combinations of features were tested. The results show that features, such as popular search keywords on Google Trends, the numbers of COVID-19 confirmed cases, and air quality indices can improve the outcome of the prediction model. The evaluation results showed that the error rate between the predicted results and the actual number of influenza-like cases form Week 15 to Week 18 fell to less than 5%. The outbreak of COVID-19 in Taiwan began in Week 19 and resulted in a sharp rise in the number of clinic or hospital visits by patients of influenza-like illnesses. After that, from Week 21 to Week 26, the error rate between the predicted and actual numbers of influenza-like cases in the later period dropped down to 13%. It can be confirmed from the actual experimental results in this research that the use of the ensemble learning prediction model proposed in this research can accurately predict the trend of influenza-like cases.

Keywords: forecasting of influenza-like illnesses; public opinion analysis; COVID-19; monitoring and early warning

check for updates

Citation: Wei, Y.-C.; Ou, Y.-L.; Li, J.; Wu, W.-C. Forecasting the Potential Number of Influenza-like Illness Cases by Fusing Internet Public Opinion. *Sustainability* **2022**, *14*, 2803. https://doi.org/10.3390/su14052803

Academic Editors: Jason C. Hung, Hwa-Young Jeong and Neil Yuwen Yen

Received: 21 December 2021
Accepted: 8 February 2022
Published: 28 February 2022

1. Introduction

The COVID-19 pandemic broke out at the end of 2019. It has spread all over the world at lightning speed. Large-scale pandemics, such as SARS, H1N1, Influenza A, and MERS, etc. have drawn much global attention to the damage that pandemics can bring to the world [1,2]. Due to the convenience in public transport, viruses these days can be easily spread to every corner of the world these days [3,4]. An influenza-like illness means any illness caused by a virus with symptoms similar to those that are caused by influenza viruses ("flu"), including symptoms such as fever, respiratory symptoms, muscle pain, and fatigue, etc. If they are not diagnosed as influenza, they are called influenza-like illnesses. In this research, an "influenza-like illness" is defined as a sudden onset of illness with a fever over of 38° or more, accompanied by respiratory symptoms, muscular soreness, headache, or extreme fatigue, excluding mild rhinitis, tonsillitis, and bronchitis [5]. The results of many studies have shown the correlation between survival rates and outbreak periods of most viruses and seasonal climate changes. Prel et al. [6] explored the effects of different climates on acute respiratory tract infections (ARI) and found that different viruses had different survival rates and tolerance of different temperatures, humidity, and other climatic conditions and seasons. Chan et al. [7] observed Hong Kong from 1997 for 10 years and found that there were two seasonal flu peaks every year and that although there were many seasonal factors of influenza, changes in weather were likely to play a key role. Taking Taiwan as an example [8], influenza outbreaks in Taiwan mainly occur in autumn and winter. The number of influenza cases gradually increases from November and peaks between December and March. Influenza may cause acute respiratory infections in patients. Common symptoms include fever, runny nose, and muscle pain, etc.

In addition to meteorological factors that may affect the timing of virus outbreaks, keyword searches from social network platforms have also been used in many studies to monitor epidemic virus outbreaks [9–14], including the number of times a keyword appears in posts and discussions on social media or search platforms, such as Facebook, Twitter, and Wikipedia. Therefore, Internet search volumes may immediately reflect how a virus is affecting people. For example, the popularity of keyword discussions can be calculated on Google Trends in real time. The data obtained from Google Trends are a better factor than that from other social media to effectively predict outbreaks of illnesses [15]. Furthermore, Google Trends has long been used by the Taiwan Centers for Disease Control (CDC) for tracking and predicting the spread of influenza [16].

According to Ginsberg et al. [16], after analyzing a large number of Google search queries, the frequency of people searching for related keywords was very close to the number of clinic or hospital visits by patients with influenza-like symptoms, and as a result, it was possible to accurately estimate the weekly numbers of influenza-like cases in various regions of the United States by putting forward 45 search keywords that were most relevant to influenza outbreaks. Kang et al. [17] studied the time correlation between influenza keywords used in Google Trends and the routine monitoring data in the Guangdong province in China to verify whether increases or decreases in Internet search volumes might match the actual number of influenza cases in the province.

An ensemble learning approach was proposed in this research using multiple features, such as the hydrometeorological data, the emergency infectious disease monitoring statistics—influenza-like illnesses, Google Trends keyword search volumes, the Taiwan public holiday information, the population data, air pollution indices, and the number of COVID-19 confirmed cases as features. Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and ensemble learning were used to predict the number of influenza-like cases in Taiwan. Root Mean Squared Log Error (RMSLE) was used in this research for model error evaluation. RMSLE is widely used to evaluate regression models and is an evaluation indicator used in many data sciences. It is similar to Root Mean Square Error (RMSE), but logarithms are used in RMSLE for calculation. The advantage of using RMSLE as an indicator is that it is robust against outliers [18]. In our experiments, different models were used for comparison and evaluation, and an ensemble learning model was used to predict the number of influenza-like illnesses in Taiwan. The results of our experiments are useful and can be applied widely in practice, as they can provide an early warning of an influenza outbreak. Therefore, the proposed ensemble learning approach can be used to predict sudden large-scale outbreaks of influenza-like illnesses. Furthermore, the proposed model can be used to prevent possible threats from these illnesses in a timely manner, to allocate medical resources reasonably to reduce morbidity and mortality, and to reduce the risk of transmissions of these illnesses.

Section 1 of this research is the introduction. Related works contributed by scholars in similar fields in the past are discussed in Section 2. The methodology used in this research is described in Section 3. The empirical results of this research are discussed in Section 4. Section 5 contains the conclusion and future work of this research.

2. Related Works

In this section, the past related works containing discussions and reviews on factors affecting influenza-like illnesses prediction models for outbreaks of influenza-like illnesses in machine learning will be reviewed.

2.1. The Definition of Influenza-Like Illnesses

Influenza is an acute viral respiratory illness, often accompanied by fever, cough, headaches, muscle pain, and other symptoms. It is mainly transmitted from person to person by droplets produced while coughing or sneezing or by touching a contaminated object or surface. It is impossible to accurately diagnose whether patients with influenza-like symptoms, severe community-infected pneumonia, or other similar illnesses are caused by influenza viruses or other pathogens from their clinical symptoms, routine examinations, and chest X-rays, etc. [19–21]. There are four types of influenza viruses: influenza A, B, C, and D. However, influenza A (H1N1 and H3N2) and influenza B are the main influenza viruses that cause current seasonal influenza [22–24]. Although in clinical diagnosis, influenza cannot be easily distinguished from other acute respiratory illnesses, such as common cold, bronchitis, or viral pneumonia, etc., influenza is usually more serious than the common cold, and the duration of treatment is longer than the common cold. Table 1 shows a comparison between influenza and the common cold [5].

Table 1. A comparison of influenza and the common cold [5].

	Influenza	**Common Cold**
Pathogens	Influenza viruses	More than 200 viruses, such as commonly seen respiratory syncytial viruses and adenoviruses, etc.
Affected parts of body	Whole body	Respiratory tracts mainly
Main clinical symptoms	Fever, cough, muscle aches, fatigue, runny nose, sore throat	Sore throat, sneezing, stuffy nose, runny nose
Complications	Pneumonia, myocarditis, encephalitis, neurological symptoms (Reye's syndrome), and other complications	Less common (otitis media, pneumonia)
Modes of transmission	Droplet and contact transmission	Droplet and contact transmission

Seasonal viruses cause respiratory illnesses when human bodies are infected by influenza viruses. In most countries, there are repeated periodic epidemics every year. The timing for seasonal influenza outbreaks is different between the southern and northern hemispheres. In the southern hemisphere, seasonal outbreaks occur between June and September every year, whereas in the northern hemisphere, they occur between November and March [5]. Seasonal outbreaks in Taiwan occur between November and March (winters) every year, as it is in the northern hemisphere.

2.2. The Selection of Training Features

There are four possible modes of transmission of influenza viruses [25]. They are: (1) transmission through direct physical contact with an infected person; (2) transmission through mediums, usually inanimate objects (such as droplets on objects or surface); (3) transmission through droplets of an infected person produced through sneezing, coughing, etc., which are transmitted to the nasal cavity or oral mucosa of a recipient; and (4) transmission through particles of a radius of 2.5 µm propelled by coughing or sneezing into the air. Viruses can survive in particles that float in the air for a long time and be transmitted through the particles.

The relative importance among the four transmission modes is a controversial issue. Lowen et al. [26] used guinea pigs as mammalian test objects to test the hypothesis that temperature and relative humidity would affect the transmission rate of influenza viruses. They found that guinea pigs are very sensitive to influenza viruses that infect humans and that the pups of guinea pigs exposed to the viruses are more likely to be infected. They used a variety of relative humidity and temperature conditions and various combinations of them to evaluate the transmission rates of influenza viruses and found that guinea pigs were very sensitive to influenza viruses that infected humans and that the pups of guinea pigs exposed to the viruses were more likely to be infected. They used a variety of relative humidity and temperature conditions and various combinations of them to evaluate the transmission rates of influenza viruses. They found that transmission speeds of influenza viruses depended on the temperature and relative humidity of the environment. Their findings support the hypothesis that meteorological conditions affect the spread of influenza viruses and help establish the link between meteorological factors and the spread and evolution of viruses, which was troublesomely uncertain in the past.

In influenza-related prediction studies, people tend to associate them with the climate and hydrological information. Prel et al. [6] explored the impacts of the climate on acute respiratory tract infection (ARI) hospitalization. Globally, ARI-related pneumonia is the leading cause of childhood deaths. It is worth noting that not all known ARI viruses cause epidemics in cold seasons, and many countries regard ARI as a common cold. The survival rates of ARI viruses may be influenced by the cold air, but the cold air is by no means the main reason that determines the survival of the viruses. Low temperature and other climatic factors may cause the viruses to increase their activity levels, adaptability, infection rates, and degrees of infection in virus hosts, pathogens, and the environment. For example, activity levels of influenza A, respiratory syncytial viruses, and adenoviruses are related to temperature, and rhinoviruses are related to relative humidity. In a study conducted by Cox and Subbarao [27], they pointed out that influenza had an obvious and consistent seasonal distribution in temperate regions and that peak outbreak seasons in winter were from November to March in the northern hemisphere and from May to September in the southern hemisphere for 5–10 weeks. Yap et al. [28] proposed that in tropical and subtropical regions, influenza-prone periods varied greatly, and there might be several peak periods within a year. Chan et al. [7] investigated the relationship between influenza activity and two key meteorological factors, namely, temperature and relative humidity, in Hong Kong from 1997 to 2006.

There are many controversies about the impacts of wind speed on the spread of viruses. Xiao et al. [29] used multiple sets of climatic conditions to conduct their research and found that slow wind speeds helped the spread of influenza A virus pandemics. Sundell et al. [30] conducted a study on the impacts of four seasons on the transmission rates of influenza A virus pandemics in temperate climates. They speculated that when an infected person coughed, certain wind speed conditions helped spread particles of droplets that contained the virus for a longer time. In addition, wind speed can help lower outdoor temperature and reduce outdoor humidity. These two effects of wind speed increase the speed of the spread of influenza A virus pandemics. However, there are other scholars who do not consider that wind speed affects the spread of viruses. Peci et al. [31] used a variety of climatic factors to conduct their research. They found that there was no correlation between wind speed and any influenza virus test results, so wind speed did not affect influenza transmission.

Air quality is currently a public health issue. Air pollution is a by-product of a civilized society. Many studies have shown that air pollution causes a variety of diseases that are harmful to the human bodies [32–34]. There are significant interactions between different types of air pollutants and respiratory diseases. Influenza-like illnesses are respiratory illnesses. Influenza-like viruses spread through air transmission or droplets, so suspended particles in the air are also one of the factors that affect influenza. Huang et al. [35] used the wavelet coherence analysis method to explore the possible correlation between suspended particles and influenza-like illnesses. Their results showed that that there was a significant correlation between suspended particles PM2.5, PM10, and NO_2 and influenza-like illnesses but that there was no correlation between suspended particles and a crowd of people over 25 years old in Nanjing, China during a peak season of influenza. Contrary to Huang's finding, Feng et al. [36] found that PM2.5 particles had a positive correlation with influenza-like illnesses in all age groups, which was most evident for the age group of 25–29 years old, followed by the age group of 15–24 years old and then the 5–14 years old and the over 60 years old groups. It had the least impact on children under 5 years old. Su et al. [37] explored the potential relationship between air pollutants and influenza-like illnesses in Jinan, China. They found a potential correlation between PM2.5, PM10, and SO_2 particles and peak periods of influenza-like illnesses. However, they found no correlation between NO_2 and O_3 particles and influenza-like illnesses. Xu et al. [38] discussed the impacts of air pollution and temperature on the occurrences of influenza cases for people aged between 0 and 14 years old in Brisbane, Australia. They used a regression model to analyze the correlation between occurrence rates of influenza cases in winter and air pollution and temperature. Studies have shown that temperature is negatively correlated with occurrence rates of influenza cases, and highly concentrated O_3 and PM10 have a significant correlation with occurrence rates of influenza cases. Therefore, O_3 and PM10 are also important indicators when assessing occurrence rates of influenza cases.

2.3. Machine Learning Models for Predicting Outbreaks of Influenza-Like Illnesses

Cheng et al. [39] used four machine learning algorithms, namely, ARIMA, Random Forest, SVM, and XGBoost, to establish a real-time national system to monitor influenza outbreaks and predict influenza-like cases for a four-week period for the Taiwan Centers for Disease Control (CDC). To combine the prediction results of the four different machine learning models, a stacking ensemble learning method was used to form the final prediction model. Its most accurate prediction result for a week scored a MAPE of less than 0.75 and a hit rate 0.75. Darwish et al. [40] used machine learning and deep learning multiple algorithms to establish a model to predict the number of influenza-like cases in Syria. The lowest MAPE of its prediction results was 3.52% and the lowest RMSE 0.01662. Chen et al. [41] used the Seasonal Autoregressive Integrated Moving Average (SARIMA) to predict outpatient rates of the influenza-like illnesses in Shenyang, China. The authors mentioned that the predicted values of influenza-like illnesses could be used as a reference for outbreaks of influenza-like cases in the short term, but other factors should be taken into consideration when forming strategies for influenza prevention and control. Hu et al. [42] proposed an IAT-BPNN model to predict the number of influenza-like illnesses in different regions of the United States. They used the artificial tree (AT) algorithm to train the model, which optimized the initial parameters of the BP neural network. They used BPNN, AT-BPNN, and IAT-BPNN in their experimental tests and comparisons. Their results showed that IAT-BPNN reduced the error rates and produced the most accurate predictions. Tapak et al. [43] used support vector machine (SVM), artificial neural-network, and Random Forest time series models to predict weekly influenza-like illnesses in Iran. The results showed that the Random Forest time series models outperformed the other three methods in simulating the weekly ILI frequencies. The comparison of related works on using machine learning for predicting outbreaks of influenza-like illnesses in Table 2.

Table 2. The comparison of related works on using machine learning for predicting outbreaks of influenza-like illnesses.

Author	Dataset	Model	Evaluation Metrics
Cheng et al. [39]	ILI dataset, records of patients with severe influenza with complications	ARIMA, Random Forest, SVR, XGBoost	Pearson correlation, MAPE, Hit rate of trend prediction
Darwish et al. [40]	EWARS data	GLM, SVR, Gradient boosting, Random Forest, LSTM	MAPE, RMSE
Chen et al. [41]	Influenza surveillance data	SARIMA	MAPE, RMSE, R2 (coefficient of determination)
Hu et al. [42]	Twitter dataset, ILI dataset	IAT-BPNN	MSE, RMSE, MAPE
Tapak et al. [43]	ILI dataset from FluNet	SVM, ANN, random forest	RMSE, MAE, ICC (intra-class correlation coefficient)

The definition of influenza, how to choose eigenvalues, and the proposed influenza prediction system have been discussed in this Section. As stated above, some scholars have proposed different methods to build an influenza-like illness prediction system, but in terms of feature selections, few studies have included weather, air pollution factors, public holidays, and other data into their model training. Therefore, different features have been incorporated in the experiments and discussions of this research.

3. Methodology

3.1. Prediction Framework

In this section, the methodology and techniques used in this research on the prediction of outbreaks of influenza-like illnesses will be described. Hydrometeorological data taken from meteorological observation data, statistics on emergency infectious diseases—influenza-like illnesses, data on keyword search volumes on Google Trends, air quality indices, data on total population, population density, and Taiwan public holiday information were used in this research. XGBoost, Random Forest, SVR, and ensemble learning were selected for experiments and verifications in this research. Descriptions of the experimental environment and related package versions are shown in Table 3:

Table 3. The experimental environment on hardware and software library.

Item	Specifications and Version Description
CPU	Intel i7-8700 3.2 GHz–4.6 GHz
GPU	Intel$^{®}$ UHD Graphics 630
RAM	8 GB DDR4 2400 MHz $*$ 4
SSD	Micron Crucial MX500 500GB SATAIII
System OS	Windows Pro 10 1909
Anaconda	1.9.7
Jupyter Notebook	6.0.1
Python	3.7.4
scikit-learn	0.23.1
XGBoost	1.1.1
Pandas	1.1.0
Numpy	1.19.1

The observation stations, where the hydrometeorological data were taken for this research, were the ground weather observation stations and the automatic weather/rainfall observation stations of the Central Weather Bureau, Taiwan. The observation data consisted of several parts. The first part comprised the data taken from the “Data Bank for

Atmospheric and Hydrologic Research" in Taiwan up to April 2020 and the data taken from the "Open Weather Data" in Taiwan from that date up to the date of this research. This research focused on predicting the number of influenza-like cases in counties and cities in Taiwan. However, as there were no ground weather observation stations of the Central Weather Bureau in some counties and cities, such as Miaoli County and Chiayi County, etc., the data for those regions were taken from the automatic weather/rainfall observation stations to fill in the missing data of these counties and cities.

Data on statistics on emergency infectious diseases—influenza-like illnesses—were taken from the "Taiwan National Infectious Disease Statistics System" of the Taiwan Centers for Disease Control (CDC), which contained statistical data on the number of visits to emergency departments at hospitals by patients with influenza-like illnesses of every age in every county/city in every week of the year.

Data on keyword search volumes were based on Google keyword searches. Various flu symptoms were selected as keywords, and their search volume values on Google were collected. The search volume values are relative values and refer to the popularity of a search term in a specific area within a specific period. The value range was set at [0, 100].

Monitoring data of the Environmental Protection Administration of the Executive Yuan in Taiwan were used as air quality indices in this research. Data on total population and population density were based on the statistical data of all counties, cities, towns, and villages in Taiwan as provided by the Department of Statistics of the Ministry of the Interior, Taiwan. The total population was the statistical data of the statistical population, and the population density was the population indicator data. The Taiwan public holiday information was taken from the open government data platform at "data.gov.tw" (accessed on 1 August 2021).

Datasets required for this research were first imported from their sources. They were then pre-processed using its applicable data processing method, and then all the processed data were grouped into its applicable training and testing datasets. Assuming prediction took place in week 0 (lag0) to predict the number of influenza-like cases in the following week, as the features used in this research to predict outbreaks of influenza-like cases did not predict outbreaks for the same week but they lagged behind for a week or longer, data of the week before (lag1) were used for week 0 for prediction. Machine learning was then used to predict the number of influenza-like cases for the following week. Figure 1 shows the framework used in this research for predicting influenza-like cases. The following subsection in this section will discuss the techniques used and the reasons why they were chosen for this research.

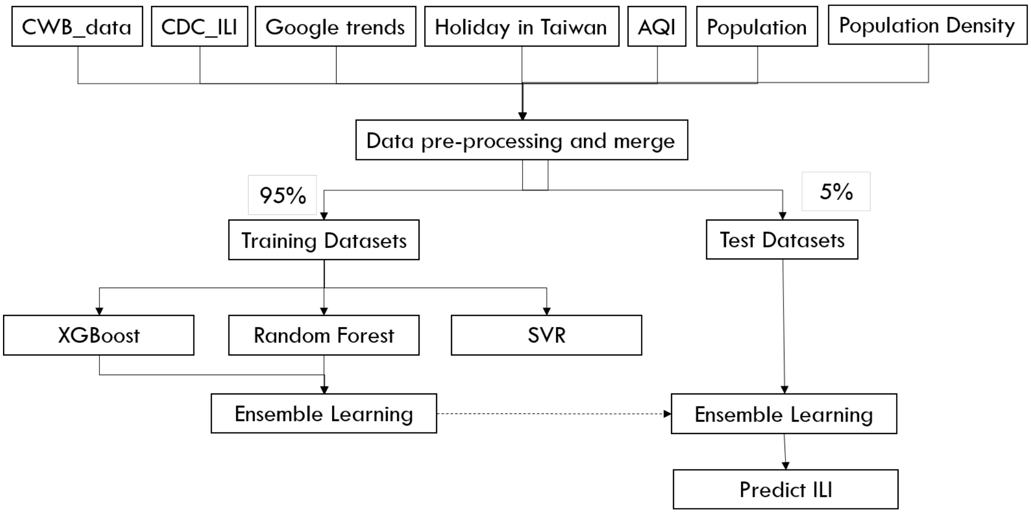


Figure 1. The framework for prediction of influenza-like cases in this research.

3.2. Data Pre-Processing

The pre-processed datasets discussed in this section are datasets that are publicly available as mentioned in the section above. The features required for this research were selected from among the datasets. The features contained in Table 4 are the original features used in this research. The features contained in Table 5 are the features related to influenza-like symptoms subsequently added to this research, as mentioned in Section 2.2, and the additional features will be compared at the end of this section. PP and RH may show negative values in the observatory instruments due to various reasons, and TX, WD may show abnormal negative values in the observation instruments due to various reasons. Please refer to Table 6. However, anomalous values for PM10, PM2.5, SO_2, O_3, and NO_2 will be removed, as mentioned in Table 7.

Table 4. Original features.

Features	Measurement Units	Descriptions	Notes
PP	Millimeter (mm)	Precipitation	The minimum value of precipitation in this research is 0. All negative values that may be caused by instrumental and human factors are replaced by 0. After excluding outliers of a distance greater than three standard deviations, the average value for a week is calculated (using one week as a unit), and a log value is taken.
RH	Percentage (%)	Average relative humidity	After excluding outliers of a distance greater than three standard deviations, the average value for a week is calculated (using one week as a unit), and a log value is taken.
TX	Celsius(°C)	Average temperature	After excluding any average temperature of a negative value and obvious outliers of a distance greater than three standard deviations, the average value for a week is calculated (using one week as a unit), and a log value is taken.
TD	Celsius(°C)	Daily temperature differences	A week is taken as a unit. After the data for a week are tallied up, a log value is taken.
WD	Meter per second (m/s)	Average wind speeds	After excluding outliers of a distance greater than three standard deviations, the average value for a week is calculated (using one week as a unit), and a log value is taken.
ILI	Number of people	The number of influenza-like cases	The number of emergency visits by patients of influenza-like illnesses is obtained and tallied up for all age groups in each county/city.
ILI_D	Number of people	Differences in the numbers of influenza-like cases	Weekly changes in the number of emergency visits by patients of influenza-like illnesses are obtained by deducting the number of emergency visits of this week with the number of visits of the week before.
GT_I	[0, 100]	Keyword search volumes on Google Trends—influenza (flu)	Values of search volumes of a keyword, "influenza (flu)," on Google Trends by people in Taiwan on a weekly basis within five years.
HoC	Number of days	The number of public holidays in Taiwan per week	The weekly number of public holidays in Taiwan in five years

Table 5. Additional features.

Features	Measurement Units	Descriptions	Notes
ILI_LW	Number of people	The number of influenza-like cases from the week before	The number of emergency visits by patients with influenza-like illnesses from the week before is obtained and tallied up in accordance with all age groups in each county/city.
GT_IS	[0, 100]	Keyword search volumes on Google Trends—influenza (flu) symptoms	Values of search volumes of keywords, influenza (flu) symptoms, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_C	[0, 100]	Keyword search volumes on Google Trends—common cold	Values of search volumes of keywords, common cold, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_CS	[0, 100]	Keyword search volumes on Google Trends—common cold symptoms	Values of search volumes of keywords, common cold symptoms, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_D	[0, 100]	Keyword search volumes on Google Trends—diarrhea	Values of search volumes of keyword, diarrhea, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_N	[0, 100]	Keyword search volumes on Google Trends—nausea	Values of search volumes of a keyword, nausea, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_FR	[0, 100]]	Keyword search volumes on Google Trends—fever	Values of search volumes of a keyword, fever, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_T	[0, 100]	Keyword search volumes on Google Trends—tiredness	Values of search volumes of a keyword, tiredness, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_FE	A value range of [0, 100]	Keyword search volumes on Google Trends—fatigue	Values of search volumes of a keyword, fatigue, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_SM	A value range of [0, 100]	Keyword search volumes on Google Trends—muscle pain	Values of search volumes of a keywords, muscle pain, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_H	A value range of [0, 100]	Keyword search volumes on Google Trends—headaches	Values of search volumes of a keyword, headaches, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_ST	A value range of [0, 100]	Keyword search volumes on Google Trends—sore throat	Values of search volumes of keywords, sore throat, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_SN	A value range of [0, 100]	Keyword search volumes on Google Trends—stuffy nose	Values of search volumes of keywords, stuffy nose, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_RN	A value range of [0, 100]	Keyword search volumes on Google Trends—runny nose	Values of search volumes of keywords, runny nose, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_CH	A value range of [0, 100]	Keyword search volumes on Google Trends—coughing	Values of search volumes of a keyword, coughing, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_S	A value range of [0, 100]	Keyword search volumes on Google Trends—sneezing	Values of search volumes of a keyword, sneezing, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
GT_DA	A value range of [0, 100]	Keyword search volumes on Google Trends—breathing difficulty	Values of search volumes of keywords, breathing difficulty, on Google Trends by people in Taiwan on a weekly basis within five years are obtained.
PIR	Number of people	Total population in each county/city/town/village	The statistics on total population in each county/city/town/village for every four quarters are obtained and converted into the weekly population data.

Table 5. *Cont.*

Features	Measurement Units	Descriptions	Notes
PD	Population density	Population Density	The population indicators in each county/city/town/village for every four quarters are obtained and converted into the weekly population density data.
PDoI	Population density	Population density of influenza	The number of patients with influenza this week (ILI) is divided by the total population of a county/city (PIR) and then multiplied by the population density of the county/city (PD) to obtain the weekly population density of influenza cases.
PM10	μg/m^3	Air quality index —PM10	The PM10 data are obtained from the air quality index and computed to obtain weekly averages.
PM2.5	μg/m^3	Air quality index —PM2.5	The PM2.5 data are obtained from the air quality index and computed to obtain weekly averages.
SO_2	ppb	Air quality index —SO_2	The PMSO2 data are obtained from the air quality index and computed to obtain weekly averages.
O_3	ppb	Air quality index —O_3	The O_3 data are obtained from the air quality index and computed to obtain weekly averages.
NO_2	ppb	Air quality index —NO_2	The NO_2 data are obtained from the air quality index and computed to obtain weekly averages.
Cov19	Number of people	The number of confirmed cases of COVID-19	The number of confirmed cases of COVID-19 is obtained and tallied up for all age groups in each county/city.

Table 6. Descriptions for the negative values at the Atmospheric Hydrological Observation Stations mentioned in this research [44].

Negative Values	Descriptions
−9991	Instrument failures, to be repaired
−9996	Data accumulated later
−9997	No information available due to unknown reasons or malfunctions
−9998	Traces of rain
−9999	No data due to no observation

Table 7. Descriptions for the anomalies at the Air Quality Index Observation Stations mentioned in this research.

Anomalies	Descriptions
#	Indicates an invalid value after instrument checks
*	Indicates an invalid value after program checks
x	Indicates an invalid value after manual checks
NR	Indicates no rain fall
blank	Indicates no value
888	Indicates no wind
999	Indicates instrument failures

Lastly, to consolidate the knowledge about influenza-like illnesses, we selected the features introduced at Tables 4 and 5 to perform different processing on different data according to the categories they belonged to. The features of hydrometeorological data were combined and calculated according to the observation stations of the county and city, from which the data was collected. The minimum value of the PP data was 0, and there should be no negative value for RH data. As to the TX data, as only the mountainous areas at a high altitude might be subjected to a temperature below 0 degree Celsius in winter while the rest of the areas should be above 0 degree Celsius, negative values of this set of data were also excluded. There should be no negative values for the WD set of data either. All anomalous negative values of these four sets of data were caused by instrumental or human factors (please refer to Table 6, and therefore, the negative values of the PP data were replaced by 0; those of the RH, TX, and WD data were excluded due to their characteristics. Additionally, outliers of a distance greater than three standard deviations were excluded. Weekly average values were calculated, and log values (using one week as a unit) were taken for the PP and RH data. TD was the most special among the hydrometeorological features. TD represented the differences between the highest and the lowest temperature of TX, which were then calculated on a weekly basis to take logs. ILI represented the emergency infectious disease monitoring statistics—the influenza-like illness, the aggregate data of the total number of emergency visits by patients of influenza-like illnesses for all age groups in each county/city. ILI_D represented the weekly changes in the number of emergency visits by patients of influenza-like illnesses by deducting the number of emergency visits in this week with the number of visits in the week before. ILI_D was the total number of emergency visits by patients of influenza-like illnesses from the week before. Each feature value of GT was the numerical data of a Google Trends search volume within 5 years. The numerical values were floating values, not absolute values. HoC was the number of public holidays per week, from Sunday to Saturday, in Taiwan within 5 years. PIR was the statistical data on the total population in each county/city in Taiwan, converted from the data for four quarters into weekly units. Data on PD were the statistical data on the population density of each county/city in Taiwan and were converted, like PIR, from the data for four quarters into weekly units. PDoIs represented the values of an IL, divided by a PIR and multiplied by a PD. As mentioned earlier, there are six severity levels of air quality indices, i.e., PM10, PM2.5, SO_2, O_3, and NO_2. Anomalies (see Table 7 for anomalies) were excluded to calculate weekly averages. Cov19 represented the number of confirmed cases of COVID-19. As symptoms of COVID-19 are similar to those of influenza, the public often finds it difficult to tell whether they catch an influenza or COVID-19 virus. It is considered that the number of confirmed cases of COVID-19 has an impact on the number of influenza-like illness cases. Therefore the number of confirmed cases of COVID-19 was taken into account in this research.

3.3. Keyword Volumes from Google Trends

Google Trends displays the search volumes of users on the Google search engine within a specific geographic region in time series indices. A keyword search index is based on its search volume proportion, meaning a search volume of a keyword is divided by the total search volume in the geographic region within a specific period to compare the relative popularity of its discussions. The percentage of a total volume of a keyword search in a designated region within a designated period is normalized to the range of [0, 100]. The maximum search volume percentage is 100, and the contrary to that is 0 [45,46].

Figure 2 is the visual presentation of keyword search volumes of "common cold" on Google Trends, adjusted to show the search popularity of the keywords in Taiwan, as the designated region, in the last 5 years. Google Trends can also use keywords to view the search popularity in each sub-region in a specific geographic area within a specific period. The search popularity in this research was calculated in the range of [0, 100]. If a search volume of a keyword in a sub-region showed the highest popularity in the total search volume in the relevant geographic region, that sub-region was marked 100. If a search volume of a keyword in a sub-region occupied only half of the total search volume in the relevant geographic region, that sub-region was marked 50. If a search volume of a keyword in a sub-region was insufficient, that sub-region was marked 0. Figure 3 presents the search popularity of the keyword of "common cold" in each sub-region in 5 years in Taiwan. The sub-regions are the counties and cities, such as Taipei City, New Taipei City, and Taichung City, etc., in Taiwan. It shows in Figure 3 that the keyword, "common cold," is the most searched in New Taipei City, whose search popularity is marked 100.

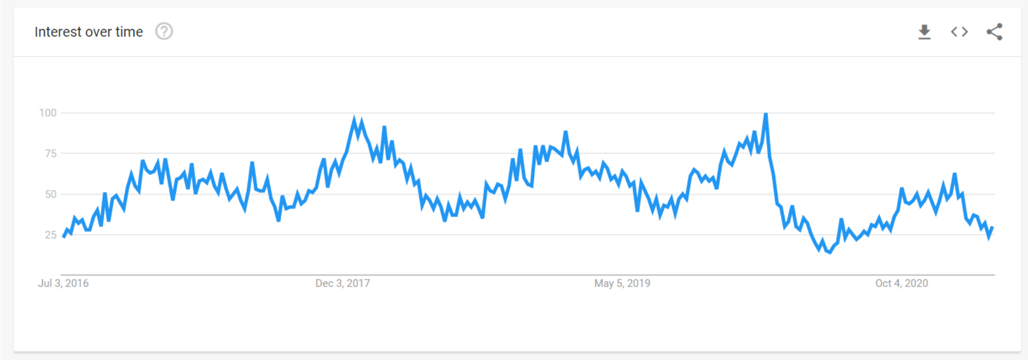
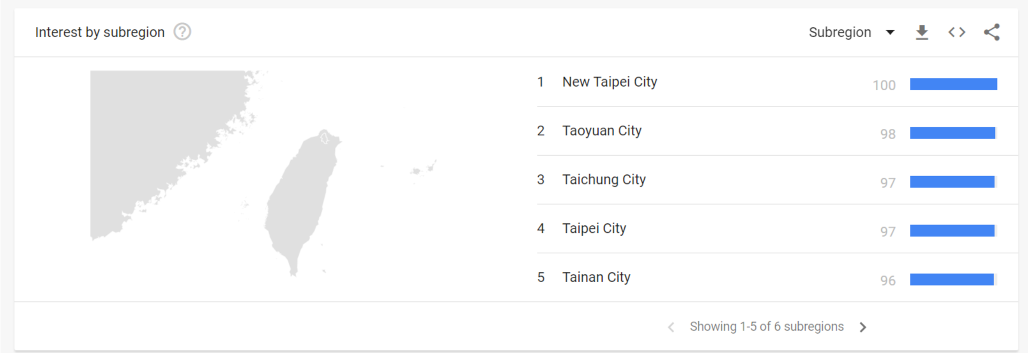


Figure 2. Trends of the search popularity of "common cold".



1	New Taipei City	100
2	Taoyuan City	98
3	Taichung City	97
4	Taipei City	97
5	Tainan City	96

Showing 1-5 of 6 subregions

Figure 3. The search popularity of "common cold" in sub-regions of Taiwan.

3.4. Lag Features

Individual data in a data series are arranged according to their time sequence (e.g., a second, minute, day, week, or month apart) in a chronological order in the time series. Time series can be divided into two types: systematic and non-systematic time series. A non-systematic time series contains random data changes, called noise. A systematic time series is divided into two types: a trend and a periodic time series. A trend time series refers to a trend of changes according to time periods, e.g., linear or exponential increases or decreases. A periodic time series refers to periodic changes, e.g., seasonal increases or changes in peak and off-peak seasons.

The datasets used in this research to predict the weekly numbers of influenza-like cases were all time series data. The data were organized into weekly units, as it did not have any impact on the daily numbers of influenza-like cases but usually had a delay impact on the number of influenza-like cases in a week later or longer. As shown in Table 8, assuming that the target prediction week of this research is Week 13 of 2021, the data of Week 11 (Lag1) of 2021 are used to fill in the data in Week 12 (Lag0), so that the data of Week 11 are used to predict the number of influenza-like cases for Week 13.

Table 8. The illustration of processing lag features.

Week	City	Flue _Amt	Lag	New	Lag	New	Lag	New	Lag
			Flue_Amt	PP	PP	GT_I	GT_I	PM10	PM10
2021-11	Nantou County	158	148	0	0	5	3	50.302	54.827
2021-11	Taoyuan City	468	459	0	0.068	5	3	47.811	36.383
2021-12	Nantou County	133	158	1.292	0	3	5	45.599	50.302
2021-12	Taoyuan City	570	468	2.047	0	3	5	40.372	47.811
2021-13	Nantou County	Predict	133	0.080	1.292	3	3	42.813	45.599
2021-13	Taoyuan City	Predict	570	0	2.047	3	3	43.098	40.372

3.5. Periodicity

In this section, the steps of feature selections are discussed. There were 34 features in total selected for this research, as mentioned in the previous section. Considering the periodic issues associated with influenza-like illnesses, one-hot encoding was used to process the time data of the numbers of years and weeks to deal with periodicity. There are 6 years from 2016 to 2021, and there are 52 or 53 weeks in each of those years, as shown in Table 9.

Table 9. The illustration of using one-hot encoding to process the time data of the numbers of years and weeks to deal with periodicity.

Year	Week	Months	Year	Week	Week	Week	Week	Week	Week	Week
			2021	01	02	03	04	05	06	07
2021	1	1	1	1	0	0	0	0	0	0
2021	2	1	1	0	1	0	0	0	0	0
2021	3	1	1	0	0	1	0	0	0	0
2021	4	1	1	0	0	0	1	0	0	0
2021	5	2	1	0	0	0	0	1	0	0
2021	6	2	1	0	0	0	0	0	1	0
2021	7	2	1	0	0	0	0	0	0	1

3.6. Machine Learning Models

Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and ensemble learning were selected as training models in this research. Random Forest affects proportions of features and facilitates the verification of hypotheses and ideas. Similar principles are used in XGBoost and Random Forest to predict results more accurately. SVR is different from the previous two. SVR was used as a control group. The final prediction results of models were compared with the past data to obtain the accuracy rates. In this research, the values of the XGBoost and Random Forest models were combined by ensemble learning as the final prediction results of this research.

3.6.1. eXtreme Gradient Boosting (XGBoost)

XGBoost [47] is a scalable machine learning system, used for tree boosting and based on the extension and improvement in the Gradient Boosted Tree (GBDT), while retaining the original model. New functions can be added to XGBoost to adjust mistakes of the last tree, i.e., to add a new tree to the last tree to rectify the insufficiency of the last tree to boost the overall efficiency, known as additive training, the formula of which is shown in Equation (1). Features are segmented each time a tree is constructed. Method (1): A greedy algorithm is used to obtain a best segmentation point. After all features are listed, a feature is used as a segmentation point, and parameters, to which features are corresponded, are calculated in accordance with Equation (2). The larger the value is, the more the loss value decreases. The best segmentation point can then be found. Method (2): A proximity algorithm is used to select features to assemble quantiles of feature distributions into sets of split points. Features of continuous information are assembled into corresponding buckets according to their segmentation points, and then samples in the buckets are accumulated, and the best split point is found by its accumulated value. Method (3): A weighted quantile algorithm is used to solve the problems that the data cannot be accessed in one go or the low efficiency of the greedy algorithm. Method (4): A sparse perception algorithm is used when the content of a dataset is sparse as there are missing values in most datasets. The datasets without missing values are used for node branching. When a feature with missing data is to be placed on a node, it will directly determine which branch node the missing value should be assigned to [48]. XGBoost is used to solve classification and regression problems.

It can generate a set of classification and regression trees (CART). Each leaf of the CART corresponds to a set of scores, which is used as the basis for classification.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)\right) + \Omega(f_t) \tag{1}$$

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\epsilon I_L} g_i)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{(\sum_{i\epsilon I_R} g_i)^2}{\sum_{i\epsilon I_R} h_i + \lambda} + \frac{(\sum_{i\epsilon I} g_i)^2}{\sum_{i\epsilon I} h_i + \lambda}\right] - \gamma \tag{2}$$

3.6.2. Random Forest (RF)

Random Forest [49] consists of decision tree classifiers. Each of the classifiers is generated independently from random vectors in input vectors. A bagging algorithm is used for each feature or each feature combination, i.e., samples are randomly taken from the training data to train multiple classifiers. Gini coefficients are used to select features, to measure the impurity of the features to their categories, and to segment each feature. The smallest Gini is selected for segmentation. In the end, if it is the classified data, weights are used to vote. The averaging method is used in the regression model to obtain results [50].

3.6.3. Support Vector Regression (SVR)

SVR [51] is an extension of the support vector machine (SVM). SVR can handle continuous prediction problems. Consistent with the classification method, SVR is characterized by the use of kernel functions, sparse solutions, VC marginal controls, and the number of support vectors. One of the main advantages of SVR is that its complex calculation does not depend on the dimensionality of the input space. It has an excellent generalization ability and a high prediction accuracy [52].

3.7. Model Evaluation

RMSLE was used in this research to measure the effects of the models. RMSLE is RMSE in the log form. It considers relative errors in the same ways as MSPE and MAPE, but RMSLE error curves are asymmetrical. The closer its value is to 0, the less often errors occur. The RMSLE formula is shown in Equation (3). The error rates of this research for influenza-like prediction results were calculated by the differences in percentage between the predicted number of cases and the actual number of confirmed cases of influenza-like illnesses. The formula for calculating the prediction error rates is shown in Equation (4).

$$\begin{aligned} RMSLE &= \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2} \\ &= RMSE(\log(y_i+1), \log(\hat{y}_i+1)) \end{aligned} \tag{3}$$

$$Predict_{err} = \frac{1}{n}\sum_{i=1}^{n}\left|1 - \frac{Predict_i}{Actual_i}\right| * 100\% \tag{4}$$

3.8. Model Testing and Adjustments

When conducting model testing and adjustments in this research, much time was spent in the data pre-processing stage to carry out the numerical processing of the data and sorting out the data from different time periods, such as excluding outliers of distances greater than three standard deviations, replacing negative PP values with 0, and excluding other anomalous data of negative values. After multiple adjustments, the processing method that produced better results was selected. During the adjustment process, it was discovered that the number of predicted cases was suddenly reduced. During the anomaly exclusion process, it was discovered during the data pre-processing that the data from the “Data Bank for Atmospheric and Hydrologic Research” in Taiwan were missing from the datasets received from the Hsin-Wu Observation Station, which was subsequently supplemented by the data taken from the "Open Weather Data" in Taiwan. After the use of the supplementary data, the problem of the sudden reduction in the predicted result was resolved. As it was found that, during the consolidation of the data, the data from Miaoli County and Chiayi County were missing, the data were then taken from the automatic weather/rainfall observation stations of the Central Weather Bureau to fill in the missing data.

The data taken from 10 remote observation stations, such as Wu-Fen-Shan Radar Station and An-Bu-Peng-Jia-Yu, etc., were first excluded, and the data taken from stations located in more populated areas were retained. Then, features, such as temperature differences (TD), differences in the number of influenza-like cases (ILI-D), and the number of public holidays in Taiwan per week (HoC), were then added. The process of additions and adjustments of various features was recorded. The first step was to add features, such as daily temperature differences (TD) and the number of influenza-like cases from the previous week. The results showed that some data reflected more accurately the actual number of confirmed cases, but some deviated more from it. However, the average differences between the predicted results and the actual confirmed cases of influenza-like illnesses of those after the addition of these features were slightly smaller than those before the addition. In the second step, the data taken from remote observation stations were excluded, and the number of Taiwan public holidays (HoC) was added. After the additions and adjustments of the features, the best parameters for the models of this research were selected. The predicted results were compared with the results in step two, and it showed better results. The average differences are relatively reduced [39,53].

Previous studies have shown that relative humidity affects the spread and survival rates of influenza viruses. Shaman and Kohn [25] mentioned in their study that absolute humidity had more obvious impacts on the spread rates and activity of viruses than relative humidity. In temperate regions, there are strong seasonal cycles of both absolute humidity indoor and outdoor. These seasonal cycles are consistent with the increases in virus activity and transmissions in winter and can be used to explain the seasonality of influenza. Therefore, differences in absolute temperature provide single, coherent, and more physical explanations for observed changes in activity, transmission, and the seasonality of influenza viruses in temperate regions. However, absolute humidity was not included as a feature in this research as there was insufficient hydrometeorological data to calculate absolute humidity.

As to the feature of the average temperature differences, Suntronwong et al. [54] explored the relationship among each influenza virus, influenza activity, and meteorological variables. After analyzing average temperature differences, relative humidity, and accumulated rainfall, it was found that all flu activity is positively correlated with average temperature, relative humidity, and rainfall. Kamigaki et al. [55] found that in the Philippines, average temperature differences were positively correlated with respiratory infections. Therefore, average temperature differences were added as a feature in this research.

As to the choices for the data for Week 0 (Lag0), assuming the number of influenza-like cases was to be predicted for Week 1, as there were no complete data for Week 0, there were a few options. Option 1 was to use the known data from the week before Week 0 to fill in the data for Week 0. Option 2 was to use the average data of the data from the month before to fill in the data for Week 0. After testing and the adjustment of parameters and the comparison with historical data, it was found that Option 1 brought the predicted results closer to the actual confirmed cases than Option 2. Option 1 was therefore adopted in this research.

Features mentioned in Section 2.2 were also selected. Features, such as the popularity or search volumes of keywords selected from influenza-like symptoms on Google Trends, various air quality indices, the data on total population and population density in each county/city, and the number of confirmed cases of COVID-19 were also adopted as features in this research. All these factors have effects on the outbreaks of influenza-like cases. These features were used in our experiments as discussed in the following section.

3.9. Experimental Design

In terms of predicting the number of clinic or hospital visits by patients of influenza-like illnesses, it is known from previous studies that the number of the clinic or hospital visits is affected by atmospheric and hydrological factors, such as precipitation (PP), relative humidity (RH), temperature (TX), temperature differences (TD), and wind speed differences (WD), etc. A new factor, keyword search volumes on Google Trends, that has not been included as a factor affecting the number of outbreaks of influenza-like cases in previous predictions of influenza-like cases was added as a feature in this research. The use of search volumes on Google Trends is the highlight of this research in the prediction of influenza-like cases. The popularity of keyword searches of influenza-like illnesses and symptoms in specific regions within specific periods was calculated through search volumes on Google Trends. The values of the popularity of keyword search volumes were set to range from 0 to 100. The larger the number was, the more popular was a keyword being discussed on Google. The smaller the number was, the less popular was a keyword being discussed on Google. The number of Taiwan public holidays was added to this research as a feature, excluding the data taken from remote observation stations and adding features of the total population and population density. Air quality indices were added as a feature as well, as they too would affect the speed of the spread of influenza-like illnesses accordingly to previous studies.

In Section 2.2, factors that affect transmission rates of influenza-like illnesses were discussed. Some scholars consider that wind speed is not significantly related to the spread of influenza-like illnesses. Some study air quality indices and discover that air pollution is significantly related to outbreaks of influenza-like illnesses. The symptoms of COVID-19 in 2019 were very similar to those of influenza-like illnesses. Most people could not tell them apart. Additionally, people had different symptoms, which made it even harder for them to diagnose themselves with the virus that caused their illness. Therefore, COVID-19 was included as a feature in this research. In order to evaluate and select features, various combinations of features were experimented in this research to conduct training, testing, and subsequent evaluation.

Random Forest, XGBoost, SVR, and ensemble learning, the most seen models used in the prediction of data, were used to construct models in this research to predict the number of influenza-like cases. Random Forest and XGBoost were used in this research to carry out the prediction and then to carry out comparisons through SVR. Random Forest and XGBoost can be used to output weights that affect features and are convenient in verifying hypotheses and ideas. Similar principles in carrying out the prediction of data are used in Random Forest and XGBoost, but, generally speaking, XGBoost is more accurate in its prediction outputs than Random Forest. To this end, in this research, an ensemble learning model combining Random Forest and XGBoost was used to obtain more accurate results. In addition to using these three models to carry out predictions, SVR was used to predict results, and its results were compared with the other three in this research. As the SVR prediction results were worse than the other three, only the ensemble learning model of Random Forest and XGBoost were used as the prediction models. The SVR prediction results were not adopted.

4. Experiments and Evaluation Results

4.1. Datasets

The data used in this research were all taken from the open data provided by various government agencies or the data publicly available on the Internet, including the hydrometeorological data, the emergency infectious disease monitoring statistics—the influenza-like illness, the Google Trends search volume data, the Taiwan public holiday information, the data on air quality indices, and the data on the total population and population density.

There were a total of 35 selected features in this research, as listed in Tables 4 and 5. Adding periodic features of 52 or 53 weeks for 6 years that were converted by one-hot encoding, and four columns of "City," "Year," "Month," and "Week," it tallied up to a total of 97 features in this research. As discussed in Section 2.2, past scholars have proposed that various hydrometeorological factors and air quality indices have significant or insignificant effects on the number of influenza-like cases. Therefore, models were trained with different combinations of features using the above-mentioned research methods and processes in our experiments. The effects of different feature combinations will be discussed at the end of this section.

4.2. Comparisons of Different Combinations of Features

As mentioned before, features used in this research were adjusted many times before being finalized. Some previous studies have suggested that many features directly affect the number and the spread of influenza-like cases, whereas some are not significantly related to outbreaks of influenza-like illnesses. Various combinations of features were therefore used in this research to conduct training, testing, and subsequent evaluation, as shown in Table 10.

Table 10. Combinations of features.

Combinations	Features Used
Org_df	Atmospheric hydrological data, the number of influenza-like cases, Google Trends search volumes: influenza, public holidays, as listed in Table 4
GT_df	Similar to Org_df, but adding the features of 17 keywords relating to influenza-like symptoms on Google Trends, such as runny nose, common cold, sore throat, etc., also adding the population data
GT_noWD_df	Similar to GT_df but excluding the feature of wind speed differences (WD)
AQI_df	Similar to GT_df but adding air quality indices, e.g., PM10, PM2.5, NO_2, SO_2 and O_3
AQI_noWD_df	Similar to AQI_df but excluding the feature of wind speed differences (WD)
Covid_noWD_df	Similar to AQI_noWD_df but adding the feature of the number of COVID-19 confirmed cases

Two RMSLE evaluation indicators were used for the performance comparison of this model. As mentioned above, various feature combinations were used for training. The numbers of influenza-like cases were predicted from Week 15 to Week 28 of 2021. Next, performance evaluations were performed on six different feature combinations. The results are shown in Tables 11–16. Predictions on the number of influenza-like cases were carried out, using three RMSLE models each week. The average method was used to calculate the error rates for these 14 weeks. Table 17 shows a comparison of the performance evaluation of the six feature combinations.

According to the error rates of all feature combinations, it was found that most of the error rates of SVR were higher than those of XGBoost and Random Forest. As a result, SVR-predicted results were not used in this research. Only the predicted results of XGBoost and Random Forest were used in this research. Table 17 shows the overall comparison of average evaluation results of XGBoost and Random Forest for each feature combinations. When the RMSLE values were used for comparison, the top three feature combinations with the lowest error rates were Covid_noWD_df, GT_noWD_df, and AQI_noWD_df. Therefore, in Section 4.3, only these three prediction results are discussed.

Table 11. Org_df performance evaluation.

Week	XGR	RF	SVR
2021-15	0.234	0.226	0.247
2021-16	0.244	0.200	0.232
2021-17	0.164	0.182	0.198
2021-18	0.215	0.194	0.220
2021-19	0.222	0.198	0.235
2021-20	0.203	0.187	0.205
2021-21	0.247	0.223	0.230
2021-22	0.187	0.198	0.224
2021-23	0.192	0.212	0.229
2021-24	0.159	0.193	0.218
2021-25	0.198	0.208	0.237
2021-26	0.203	0.187	0.218
2021-27	0.248	0.224	0.242
2021-28	0.259	0.234	0.244

Table 12. GT_noWD_df performance evaluation.

week	XGR	RF	SVR
2021-15	0.220	0.222	0.250
2021-16	0.223	0.196	0.226
2021-17	0.154	0.176	0.190
2021-18	0.166	0.193	0.219
2021-19	0.204	0.204	0.246
2021-20	0.161	0.178	0.198
2021-21	0.226	0.224	0.226
2021-22	0.162	0.190	0.220
2021-23	0.187	0.205	0.223
2021-24	0.159	0.186	0.219
2021-25	0.186	0.211	0.235
2021-26	0.192	0.186	0.229
2021-27	0.230	0.217	0.226
2021-28	0.233	0.227	0.232

Table 13. GT_noWD_df performance evaluation.

Week	XGR	RF	SVR
2021-15	0.214	0.223	0.251
2021-16	0.203	0.197	0.226
2021-17	0.155	0.178	0.190
2021-18	0.162	0.194	0.218
2021-19	0.207	0.202	0.246
2021-20	0.166	0.177	0.198
2021-21	0.203	0.227	0.226
2021-22	0.168	0.191	0.220
2021-23	0.169	0.203	0.223
2021-24	0.160	0.186	0.220
2021-25	0.192	0.212	0.235
2021-26	0.194	0.188	0.229
2021-27	0.230	0.219	0.226
2021-28	0.224	0.226	0.232

Table 14. AQI_df performance evaluation.

Week	XGR	RF	SVR
2021-15	0.216	0.224	0.247
2021-16	0.206	0.199	0.225
2021-17	0.161	0.179	0.190
2021-18	0.167	0.194	0.218
2021-19	0.198	0.207	0.244
2021-20	0.145	0.177	0.200
2021-21	0.209	0.225	0.227
2021-22	0.173	0.194	0.220
2021-23	0.204	0.208	0.223
2021-24	0.176	0.187	0.220
2021-25	0.187	0.212	0.237
2021-26	0.195	0.189	0.229
2021-27	0.229	0.221	0.227
2021-28	0.221	0.228	0.231

Table 15. AQI_noWD_df performance evaluation.

Week	XGR	RF	SVR
2021-15	0.217	0.224	0.247
2021-16	0.204	0.199	0.225
2021-17	0.164	0.180	0.190
2021-18	0.161	0.196	0.217
2021-19	0.186	0.208	0.244
2021-20	0.170	0.178	0.200
2021-21	0.215	0.224	0.227
2021-22	0.161	0.193	0.220
2021-23	0.167	0.208	0.223
2021-24	0.163	0.188	0.220
2021-25	0.183	0.213	0.236
2021-26	0.196	0.192	0.229
2021-27	0.223	0.221	0.227
2021-28	0.238	0.227	0.231

Table 16. Covid_noWD_df performance evaluation.

Week	XGR	RF	SVR
2021-15	0.204	0.188	0.224
2021-16	0.152	0.171	0.223
2021-17	0.172	0.170	0.200
2021-18	0.206	0.194	0.237
2021-19	0.178	0.175	0.232
2021-20	0.176	0.168	0.204
2021-21	0.162	0.177	0.199
2021-22	0.233	0.167	0.213
2021-23	0.270	0.169	0.212
2021-24	0.235	0.178	0.212
2021-25	0.208	0.206	0.238
2021-26	0.249	0.211	0.245
2021-27	0.219	0.191	0.226
2021-28	0.140	0.168	0.219

Table 17. A comparison of the performance evaluation of the six feature combinations.

Combinations	RMSLE
Org_df	0.2085
GT_df	0.197
GT_noWD_df	0.1955
AQI_df	0.1975
AQI_noWD_df	0.1965
Covid_noWD_df	0.1905

4.3. Evaluation Results and Discussion

In terms of results, the ensemble learning model combining RandomForest and XGBoost was used as a predictive model, and the data of the week before Week 0 were used as the data of Week 0; the number of influenza-like cases for Week 1 was predicted. It was mentioned in the previous section that the three feature combinations with the smallest error rates were: Covid_noWD_df, GT_noWD_df, and AQI_noWD_df. So, next, the results of these three feature combinations will be discussed.

The weekly numbers of influenza-like cases for 14 weeks of 2021 were examined and predicted in this research. As seen from Figures 4–6, the prediction results from Week 15 to Week 18 were close to the actual number of influenza-like cases, and the prediction error rate for that period was only about 5%. From Week 19, the difference between the predicted and the actual number of influenza-like cases was huge, and the prediction error rate increased to 50%. The reason for the increase in the error rate was that in Week 19, there was an outbreak of COVID-19 in Taiwan, and the number of confirmed cases of COVID-19 surged. However, the symptoms of COVID-19 are similar to those of influenza, such as fever, coughs, fatigue, etc.. The incubation period of influenza is 1–4 days, and it takes approximately 2–14 days for COVID-19 symptoms to appear [56]. Therefore, it is difficult for people to quickly tell whether they catch a COVID-19 or an influenza-like virus. According to the guidelines issued by the Taiwan Centers for Disease Control, all patients who exhibit symptoms of COVID-19 or influenza-like illnesses must report to the relevant authorities, take appropriate protective measures, and seek medical treatment. When people find that they have similar symptoms, they choose to seek medical treatment directly, which leads to an increase in the number of patients of influenza-like illnesses. At present, the number of confirmed cases of COVID-19 in Taiwan has gradually slowed down, and the number of clinic or hospital visits by patients of influenza-like illnesses has also decreased. From Week 21 to Week 26, the predicted and the actual numbers of influenza-like cases gradually became close to each other, and the prediction error rate for that period was reduced to 13%.

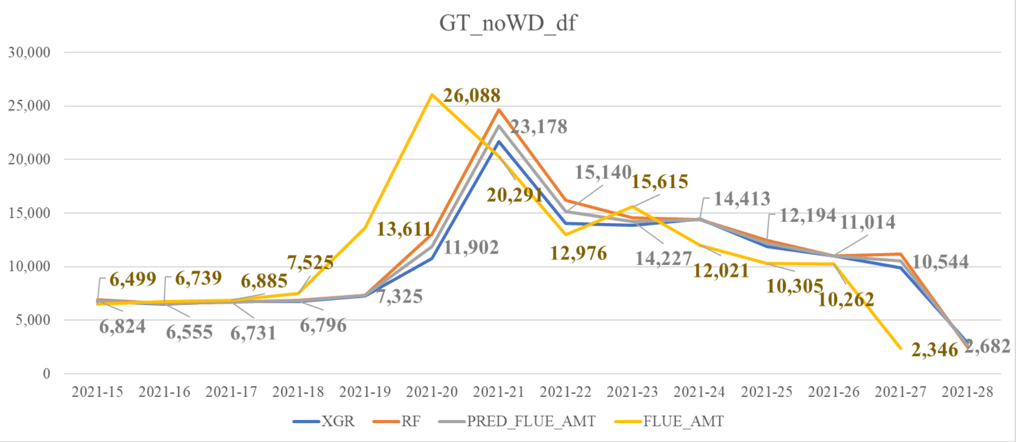


Figure 4. Predicted results of GT_noWD_df.

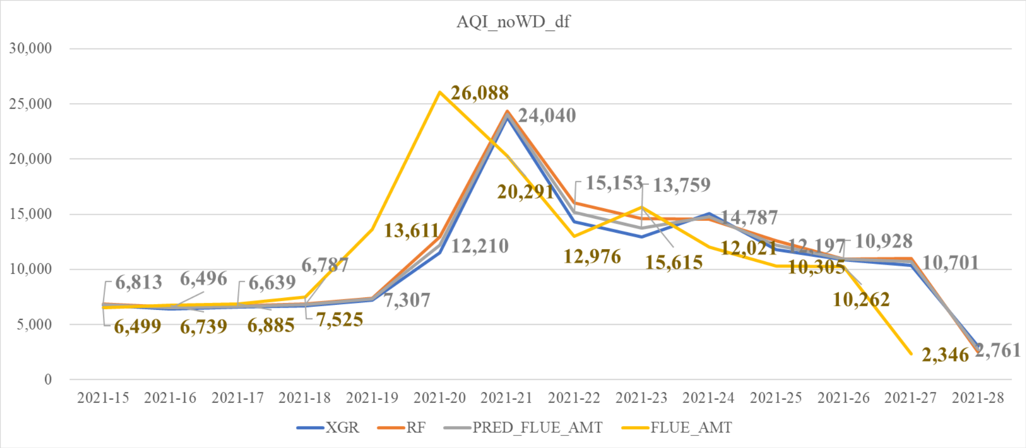


Figure 5. Predicted results of AQI_noWD_df.

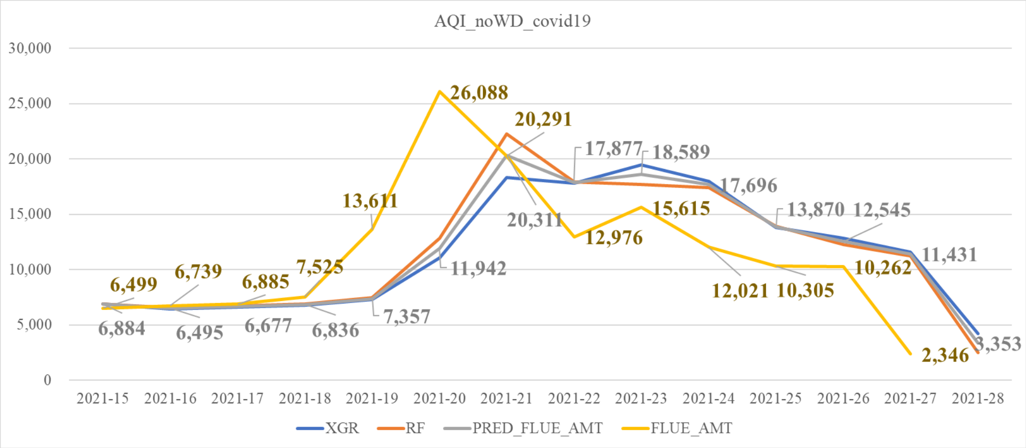


Figure 6. Predicted results of Covid_noWD_df.

5. Conclusions and Future Work

Due to the convenience in public transport, viruses can be easily spread to every corner of the world these days, especially as it can be seen in the COVID-19 pandemic. However, the symptoms of COVID-19, SARS, H1N1, and Influenza A are very similar. Before pandemics are said to be caused by influenza viruses, they are called influenza-like illnesses. If the potential number of influenza-like illness cases can be predicted earlier and accurately, the predicted results can help the government, hospitals, pharmacies, and companies quickly prepare for the spread of influenza-like cases, as they can help form informed decisions and take preventive measures. In this research, an ensemble learning approach, fusing Random Forest and XGBoost learning models, is proposed. Multiple features, such as the hydrometeorological data, the emergency infectious disease monitoring statistics on influenza-like illnesses, Google Trends keyword search volumes, the Taiwan public holiday information, the population density, average temperature differences, air pollution indices, and the number of COVID-19 confirmed cases, were used in the proposed model.

In our experiments, the weekly numbers of influenza-like cases were predicted for 14 weeks in 2021. The experimental results were compared with the actual numbers of influenza-like cases. The error rate for the period from Week 15 to Week 18 was within 5%. In Week 19, there was a sudden surge in the number of influenza-like cases. According to seasonal flu periods in Taiwan, outbreaks do not occur in summers. Nevertheless, the number of COVID-19 confirmed cases suddenly increased at that time. It is speculated that it is because people cannot tell for sure whether they have a common cold, a flu-like illness, or COVID-19. Moreover, the Taiwan Centers for Disease Control requires all patients who exhibit symptoms of COVID-19 or influenza-like illnesses to seek medical treatment. This led to an increase in the number of patients of influenza-like illnesses being reported at that time. Three weeks later, the number of confirmed cases of COVID-19 in Taiwan gradually slowed down, and the number of clinic or hospital visits by patients of influenza-like illnesses also decreased. The prediction error rate between the predicted and the actual number of influenza-like cases for that period was reduced to 13%, getting closer to the actual number of influenza-like cases. The experimental results showed that our proposed ensemble learning approach could accurately predict the number of influenza-like cases. The outcomes of our experiments can be practically useful and applied widely, as they can provide an early warning of an influenza outbreak. The proposed model can be used to prevent possible threats from these illnesses in a timely manner, to allocate medical resources reasonably to reduce morbidity and mortality, and to reduce the risk of transmission of these illnesses.

In the future, the model will be built into a prediction system, which will be provided to the government, hospitals, pharmacies, and companies to predict the number of influenza-like illnesses at any time. This will enable them to quickly understand the spread of influenza-like cases in the future, so that they can form informed decisions and take preventive measures. This can also help the public understand, through the government and hospitals, potential large-scale outbreaks of influenza-like illnesses in the near future, so that they can take measures to protect their own health and safety.

Author Contributions: Conceptualization, Y.-C.W., J.L. and W.-C.W.; Funding acquisition, Y.-C.W.; Investigation, Y.-L.O.; Methodology, Y.-C.W. and Y.-L.O.; Project administration, Y.-C.W.; Supervision, Y.-C.W.; Writing–original draft, Y.-C.W. and Y.-L.O.; Writing–review & editing, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Onward Security (No. 209A136), Ministry of Science and Technology (No. 110-2637-H-027-004-), and National Taipei University of Technology-Beijing University of Technology Joint Research Program (No. NTUT-BJUT-110-01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Related data discussed in this article can be found at the following project websites: https://e-service.cwb.gov.tw/HistoryDataQuery/, https://scidm.nchc.org.tw/dataset/best_wish14584, https://data.gov.tw/dataset/14718, https://data.epa.gov.tw/dataset/detail/AQX_P_488 (accessed on 7 February 2022).

Acknowledgments: We would like to thank the anonymous reviewers for their valuable feedback. We are also grateful to Lien Wang and another anonymous individual for proofreading the manuscript.

Conflicts of Interest: The authors declare no conflict interest.

References

1. Petersen, E.; Koopmans, M.; Go, U.; Hamer, D.H.; Petrosillo, N.; Castelli, F.; Storgaard, M.; Al Khalili, S.; Simonsen, L. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect. Dis.* **2020**, *20*, e238–e244. [CrossRef]
2. Abdelrahman, Z.; Li, M.; Wang, X. Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses. *Front. Immunol.* **2020**, *11*. [CrossRef] [PubMed]
3. Wei, J.T.; Liu, Y.X.; Zhu, Y.C.; Qian, J.; Ye, R.Z.; Li, C.Y.; Ji, X.K.; Li, H.K.; Qi, C.; Wang, Y.; et al. Impacts of transportation and meteorological factors on the transmission of COVID-19. *Int. J. Hyg. Environ. Health* **2020**, *230*, 113610. [CrossRef] [PubMed]
4. Du, Z.; Wang, L.; Cauchemez, S.; Xu, X.; Wang, X.; Cowling, B.J.; Meyers, L.A. Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerg. Infect. Dis.* **2020**, *26*, 1049. [CrossRef]
5. Taiwan Centers for Disease Control. *Practical Guidelines for Prevention and Control of Seasonal Influenza*; Report; Taiwan Centers for Disease Control: Taipei, Taiwan, 2020.
6. du Prel, J.B.; Puppe, W.; Gröndahl, B.; Knuf, M.; Weigl, F.; Schaaff, F.; Schaaff, F.; Schmitt, H.J. Are meteorological parameters associated with acute respiratory tract infections? *Clin. Infect. Dis.* **2009**, *49*, 861–868. [CrossRef]
7. Chan, P.K.; Mok, H.; Lee, T.; Chu, I.M.; Lam, W.; Sung, J.J. Seasonal influenza activity in Hong Kong and its association with meteorological variations. *J. Med Virol.* **2009**, *81*, 1797–1806. [CrossRef]
8. Taiwan Centers for Disease Control. *Severe Complicated Influenza*; Taiwan Centers for Disease Control: Taipei, Taiwan, 2020.

9. Wang, Y.; Xu, K.; Kang, Y.; Wang, H.; Wang, F.; Avram, A. Regional influenza prediction with sampling Twitter data and PDE model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 678. [CrossRef]
10. Seo, D.W.; Shin, S.Y. Methods using social media and search queries to predict infectious disease outbreaks. *Healthc. Inform. Res.* **2017**, *23*, 343. [CrossRef]
11. Daughton, A.R.; Chunara, R.; Paul, M.J. Comparison of social media, syndromic surveillance, and microbiologic acute respiratory infection data: Observational study. *Jmir Public Health Surveill.* **2020**, *6*, e14986. [CrossRef]
12. Lampos, V.; Zou, B.; Cox, I.J. Enhancing feature selection using word embeddings: The case of flu surveillance. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 695–704.
13. Volkova, S.; Ayton, E.; Porterfield, K.; Corley, C.D. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLoS ONE* **2017**, *12*, e0188941. [CrossRef]
14. Lee, K.; Agrawal, A.; Choudhary, A. Forecasting influenza levels using real-time social media streams. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017; pp. 409–414.
15. Huang, L.H. A Deep Learning Based Approach to Forecasting Influenza-Like Illness Rate. Master's Thesis, Tzu Chi University, Hualien, Taiwan, 2020.
16. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012–1014. [CrossRef]
17. Kang, M.; Zhong, H.; He, J.; Rutherford, S.; Yang, F. Using google trends for influenza surveillance in South China. *PLoS ONE* **2013**, *8*, e55205. [CrossRef]
18. Zeroual, A.; Harrou, F.; Dairi, A.; Sun, Y. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos Solitons Fractals* **2020**, *140*, 110121. [CrossRef] [PubMed]
19. Paules, C.I.; Subbarao, K. Influenza vaccination and prevention of cardiovascular disease mortality–Authors' reply. *Lancet* **2018**, *391*, 427–428. [CrossRef]
20. Masters, B.R. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*; Bennett, J.E., Dolin, R., Blaser, M.J., Eds.; Elsevier Saunders: Philadelphia, PA, USA, 2016; ISBN 13-978-1-4557-4801-3.
21. Centers for Disease Control and Prevention. *Epidemiology and Prevention of Vaccine-Preventable Diseases*; Department of Health & Human Services, Public Health Service, Centers for Disease Control and Prevention: Atlanta, GA, USA, 2005.
22. Hause, B.M.; Collin, E.A.; Liu, R.; Huang, B.; Sheng, Z.; Lu, W.; Wang, D.; Nelson, E.A.; Li, F. Characterization of a novel influenza virus in cattle and swine: Proposal for a new genus in the Orthomyxoviridae family. *MBio* **2014**, *5*, e00031-14. [CrossRef] [PubMed]
23. Liu, R.; Sheng, Z.; Huang, C.; Wang, D.; Li, F. Influenza D virus. *Curr. Opin. Virol.* **2020**, *44*, 154–161. [CrossRef]
24. Ferguson, L.; Olivier, A.K.; Genova, S.; Epperson, W.B.; Smith, D.R.; Schneider, L.; Barton, K.; McCuan, K.; Webby, R.J.; Wan, X.F. Pathogenesis of influenza D virus in cattle. *J. Virol.* **2016**, *90*, 5636–5642. [CrossRef]
25. Shaman, J.; Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3243–3248. [CrossRef]
26. Lowen, A.C.; Mubareka, S.; Steel, J.; Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog.* **2007**, *3*, e151. [CrossRef]
27. Cox, N.J.; Subbarao, K. Global epidemiology of influenza: Past and present. *Annu. Rev. Med.* **2000**, *51*, 407–421. [CrossRef]
28. Yap, F.H.; Ho, P.; Lam, K.; Chan, P.K.; Cheng, Y.; Peiris, J.S. Excess hospital admissions for pneumonia, chronic obstructive pulmonary disease, and heart failure during influenza seasons in Hong Kong. *J. Med Virol.* **2004**, *73*, 617–623. [CrossRef] [PubMed]
29. Xiao, H.; Tian, H.; Lin, X.; Gao, L.; Dai, X.; Zhang, X.; Chen, B.; Zhao, J.; Xu, J. Influence of extreme weather and meteorological anomalies on outbreaks of influenza A (H1N1). *Chin. Sci. Bull.* **2013**, *58*, 741–749. [CrossRef] [PubMed]
30. Sundell, N.; Andersson, L.M.; Brittain-Long, R.; Lindh, M.; Westin, J. A four year seasonal survey of the relationship between outdoor climate and epidemiology of viral respiratory tract infections in a temperate climate. *J. Clin. Virol.* **2016**, *84*, 59–63. [CrossRef]
31. Peci, A.; Winter, A.L.; Li, Y.; Gnaneshan, S.; Liu, J.; Mubareka, S.; Gubbay, J.B. Effects of absolute humidity, relative humidity, temperature, and wind speed on influenza activity in Toronto, Ontario, Canada. *Appl. Environ. Microbiol.* **2019**, *85*, e02426-18. [CrossRef]
32. Brunekreef, B.; Holgate, S.T. Air pollution and health. *Lancet* **2002**, *360*, 1233–1242. [CrossRef]
33. Lelieveld, J.; Klingmüller, K.; Pozzer, A.; Pöschl, U.; Fnais, M.; Daiber, A.; Münzel, T. Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *Eur. Heart J.* **2019**, *40*, 1590–1596. [CrossRef] [PubMed]
34. Mannucci, P.M.; Franchini, M. Health effects of ambient air pollution in developing countries. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1048. [CrossRef]
35. Huang, L.; Zhou, L.; Chen, J.; Chen, K.; Liu, Y.; Chen, X.; Tang, F. Acute effects of air pollution on influenza-like illness in Nanjing, China: A population-based study. *Chemosphere* **2016**, *147*, 180–187. [CrossRef]
36. Feng, C.; Li, J.; Sun, W.; Zhang, Y.; Wang, Q. Impact of ambient fine particulate matter (PM 2.5) exposure on the risk of influenza-like-illness: A time-series analysis in Beijing, China. *Environ. Health* **2016**, *15*, 1–12. [CrossRef]
37. Su, W.; Wu, X.; Geng, X.; Zhao, X.; Liu, Q.; Liu, T. The short-term effects of air pollutants on influenza-like illness in Jinan, China. *BMC Public Health* **2019**, *19*, 1–12. [CrossRef]

38. Xu, Z.; Hu, W.; Williams, G.; Clements, A.C.; Kan, H.; Tong, S. Air pollution, temperature and pediatric influenza in Brisbane, Australia. *Environ. Int.* **2013**, *59*, 384–388. [CrossRef] [PubMed]
39. Cheng, H.Y.; Wu, Y.C.; Lin, M.H.; Liu, Y.L.; Tsai, Y.Y.; Wu, J.H.; Pan, K.H.; Ke, C.J.; Chen, C.M.; Liu, D.P.J. Applying machine learning models with an ensemble approach for accurate real-time influenza forecasting in Taiwan: Development and validation study. *J. Med Internet Res.* **2020**, *22*, e15394. [CrossRef] [PubMed]
40. Darwish, A.; Rahhal, Y.; Jafar, A. A comparative study on predicting influenza outbreaks using different feature spaces: Application of influenza-like illness data from Early Warning Alert and Response System in Syria. *BMC Res. Notes* **2020**, *13*, 1–8. [CrossRef]
41. Chen, Y.; Leng, K.; Lu, Y.; Wen, L.; Qi, Y.; Gao, W.; Chen, H.; Bai, L.; An, X.; Sun, B.J.E.; et al. Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018. *Epidemiol. Infect.* **2020**, *148*, e29. [CrossRef] [PubMed]
42. Hu, H.; Wang, H.; Wang, F.; Langley, D.; Avram, A.; Liu, M. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Sci. Rep.* **2018**, *8*, 1–8. [CrossRef] [PubMed]
43. Tapak, L.; Hamidi, O.; Fathian, M.; Karami, M. Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Res. Notes* **2019**, *12*, 1–6. [CrossRef]
44. Central Weather Bureau. *Central Meteorological Administration Station Data Description*; Report; Central Weather Bureau: Taipei, Taiwan, 2018.
45. Google. FAQ about Google Trends Data. Available online: https://support.google.com/trends/answer/4365533?hl=en (accessed on 7 February 2022).
46. Choi, H.; Varian, H. Predicting the present with Google Trends. *Econ. Rec.* **2012**, *88*, 2–9. [CrossRef]
47. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.
48. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
49. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote. Sens.* **2005**, *26*, 217–222. [CrossRef]
51. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 155–161.
52. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines*; Springer: Berkeley, CA, USA, 2015; pp. 67–80.
53. Chakraborty, P.; Lewis, B.; Eubank, S.; Brownstein, J.S.; Marathe, M.; Ramakrishnan, N. What to know before forecasting the flu. *PLoS Comput. Biol.* **2018**, *14*, e1005964. [CrossRef] [PubMed]
54. Suntronwong, N.; Vichaiwattana, P.; Klinfueng, S.; Korkong, S.; Thongmee, T.; Vongpunsawad, S.; Poovorawan, Y. Climate factors influence seasonal influenza activity in Bangkok, Thailand. *PLoS ONE* **2020**, *15*, e0239729. [CrossRef] [PubMed]
55. Kamigaki, T.; Chaw, L.; Tan, A.G.; Tamaki, R.; Alday, P.P.; Javier, J.B.; Olveda, R.M.; Oshitani, H.; Tallo, V.L. Seasonality of influenza and respiratory syncytial viruses and the effect of climate factors in subtropical–tropical asia using influenza-like illness surveillance data, 2010–2012. *PLoS ONE* **2016**, *11*, e0167712.
56. The NYC Health Department. *Is It the FLU OR COVID-19?* Report; The NYC Health Department: New York, NY, USA, 2020.