International Journal of
*Environmental Research and Public Health*

MDPI

*Article*

# The Prediction of Influenza-like Illness and Respiratory Disease Using LSTM and ARIMA

**Yu-Tse Tsan [1,2,3], Der-Yuan Chen [4,5], Po-Yu Liu [6], Endah Kristiani [7,8], Kieu Lan Phuong Nguyen [9] and Chao-Tung Yang [7,10,*]**

[1] Department of Emergency Medicine, Taichung Veterans General Hospital, Taichung 407204, Taiwan; janyuhjer@gmail.com
[2] School of Medicine, Chung Shan Medical University, Taichung 40201, Taiwan
[3] Division of Occupational Medicine, Department of Emergency Medicine, Taichung Veterans General Hospital, Taichung 407204, Taiwan
[4] College of Medicine, China Medical University, Taichung 406040, Taiwan; dychen1957@gmail.com
[5] Rheumatology and Immunology Center, China Medical University Hospital, Taichung 404332, Taiwan
[6] Division of Infection, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung 40704, Taiwan; pyliu@vghtc.gov.tw
[7] Department of Computer Science, Tunghai University, Taichung 407224, Taiwan; endahkristi@thu.edu.tw
[8] Department of Informatics, Krida Wacana Christian University, Jakarta 11470, Indonesia
[9] Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, Ho Chi Minh City 70000, Vietnam; nklphuong@ntt.edu.vn
[10] Research Center for Smart Sustainable Circular Economy, Tunghai University, No. 1727, Sec. 4, Taiwan Boulevard, Taichung City 407224, Taiwan
[*] Correspondence: ctyang@thu.edu.tw

**Abstract:** This paper proposed the forecasting model of Influenza-like Illness (ILI) and respiratory disease. The dataset was extracted from the Taiwan Environmental Protection Administration (EPA) for air pollutants data and the Centers for Disease Control (CDC) for disease cases from 2009 to 2018. First, this paper applied the ARIMA method, which trained based on the weekly number of disease cases in time series. Second, we implemented the Long short-term memory (LSTM) method, which trained based on the correlation between the weekly number of diseases and air pollutants. The models were also trained and evaluated based on five and ten years of historical data. Autoregressive integrated moving average (ARIMA) has an excellent model in the five-year dataset of ILI at 2564.9 compared to ten years at 8173.6 of RMSE value. This accuracy is similar to the Respiratory dataset, which gets 15,656.7 in the five-year dataset and 22,680.4 of RMSE value in the ten-year dataset. On the contrary, LSTM has better accuracy in the ten-year dataset than the five-year dataset. For example, on average of RMSE in the ILI dataset, LSTM has 720.2 RMSE value in five years and 517.0 in ten years dataset. Also, in the Respiratory disease dataset, LSTM gets 4768.6 of five years of data and 3254.3 of the ten-year dataset. These experiments revealed that the LSTM model generally outperforms ARIMA by three to seven times higher model performance.

**Keywords:** air pollution; PM2.5; ILI; influenza-like illness; respiratory disease; LSTM; ARIMA

check for updates

**Citation:** Tsan, Y.-T.; Chen, D.-Y.; Liu, P.-Y.; Kristiani, E.; Nguyen, K.L.P.; Yang, C.-T. The Prediction of Influenza-like Illness and Respiratory Disease Using LSTM and ARIMA. *Int. J. Environ. Res. Public Health* **2022**, *19*, 1858. https://doi.org/10.3390/ijerph19031858

Academic Editor: Caroline Beardsmore

Received: 25 December 2021
Accepted: 4 February 2022
Published: 7 February 2022

## 1. Introduction

The primary factor of global death and disease leads to outdoor air pollutants. The increasing number of hospital and emergency room visitations indicates when environmental exposures happened indicates that there is a correlation between them [1]. World Health Organization (WHO) addressed that around 4.2 million premature deaths cases are correlated with air pollution [2,3]. The cases account for 29% from lung cancer, 17% from an acute lower respiratory infection, 24% from stroke, 25% from ischaemic heart disease, and 43% from chronic obstructive pulmonary disease. Therefore, the study of modeling the correlation of air pollutants and diseases such as Influenza-like Illness (ILI) and respiratory illness is notable [4].

This paper proposed the prediction model of air pollutants, ILI, and respiratory disease using Long short-term memory (LSTM) and Autoregressive integrated moving average (ARIMA). This paper aims to predict the amount case of ILI and respiratory disease based on air pollutants parameters. In this case, we compare the different datasets of five years and ten years terms. Also, we applied feature selection methods for the training model of LSTM [5]. The specific objectives are listed as follows:

- We empirically study the training model of ARIMA and LSTM with the input of air pollutants, and the output is the number of diseases predicted.
- We Evaluate and compare the training model of ARIMA and LSTM for air pollutant and diseases cases.

This paper's contribution is to empirically study the comparison models of ILI and Respiratory disease in two different time-spans of 5 and 10 years datasets. Moreover, we compare the trained models using three feature selections from matrix correlation, extra trees classifier, and chemical-based methods. In addition, we also trained all parameters of the air quality dataset. Therefore, we have 20 models of ILI disease and 20 models of Respiratory disease for evaluating the model's performance. Furthermore, other research might refer to our results for further development. However, this paper has the limitation as follows.

- We trained ARIMA, which is the statistical-based and univariate input.
- We trained LSTM, the neural networks-based, and multivariate input.
- We implemented three feature selections on LSTM models.
- We just compared the result of ARIMA and LSTM models.

## 2. Background Reviews and Related Works

### 2.1. Root Mean Squared Error (RMSE)

The RMSE is the standard deviation of residuals (prediction errors). The residuals are a metric about how far these data points are from the regression line, and the RMSE reflects how wide the spread out of these residuals are. In other words, it indicates how narrowly the information is concentrated on the best-fit line. Root mean square error is commonly used to assess experimental findings in climatology, forecasting, and regression analysis. The following is a description of the equation:

$$RMSE = (1/X) \sum_{X}^{(i=1)} (z_i - \cap z_i)^2 \tag{1}$$

where $X$ is sample size, $z_i$ is the actual expected output, and $\cap z_i$ is the model's prediction.

### 2.2. Feature Selections

Before conducting to analyze time-series data, the input data should be prepared to achieve the best performance. Data pre-processing might include data transformation, handling with missing data, and feature selection. In addition, feature selection is a critical step to identify significant features for noise filtration purpose leading to reduce the dimensionality of the data. This procedure can also improve the accuracy of the model Panda et al. [6]. Several studies propose some approaches to select essential factors to improve time-series forecasting. Zhai and Chen (2018) [7] apply four different feature extraction involving simplification, polynomial, transformation, and combination. As a result of the four methods, $NO_2$ and CO contribute the most in forecasting $PM_{2.5}$ concentrations. Sethi and Mittal (2019) [8] denote measuring the relationship between air pollutants and $PM_{2.5}$ play a vital role in predict accurately the concentration of $PM_{2.5}$. The dependencies among features can also decide the relevance of the feature in the dataset, determined using linear regression. The association between $PM_{2.5}$ and $PM_{10}$, and $NO_2$, $SO_2$ are also measured by Pearson correlation coefficients in a study at 31 Chinese cities (Xie et al., 2015) [9].

### *2.3. Air Pollutants and Public Health Correlations*

There is a strong association between deteriorating air quality and public health problems [10,11]. According to the WHO, the exposure to indoor air pollution causes approximately 4.2 million deaths worldwide every year, while the number of deaths relating to ambient air pollution is 3.8 million (World Health Organization, 2016) [12]. The effect of air pollutants has been demonstrated in previous studies on respiratory, nervous, and cardiovascular systems [13]. For example, nitrogen oxides (NOx), sulfur dioxide (SO2), ozone (O3), fine-particular matter (PM2.5) are influencing factors that lead to air pollution-related diseases such as asthma, ischemic heart disease, and neurodevelopmental disorders (Genc et al., 2012 [14]; Kampa and Castanas, 2008 [15]; Özkaynak et al., 2013) [16]. A combination of SO2, NO2, and some heavy metals will result in nose and throat irritation and could become more serious for asthmatic people because of bronchoconstriction and dyspnea symptoms. Gianfrancesco et al. [17] proposed the big data analysis such as the evaluation of the influence of air pollution on ILI requires an adaptation of machine leaning approach. However, the machine learning approach used in their study may also be subject to bias.

### *2.4. ARIMA and LSTM*

ARIMA and LSTM are the conventional statistical and deep learning models, respectively, widely applied for analyzing and forecasting time series data in various sectors. Eymen and Köylü (2019) [18] use ARIMA tool to predict relative humidity, which was slipped into pre-dam and post-dam. Pre-dam relative humidity fitted the ARIMA model, while post-dam relative humidity is compatible with data predicted by ARIMA. In addition, Chakraborty et al. (2019) [19] propose hybrid models that integrate ARIMA and some models of machine learning such as artificial neural networks (ANN), support vector machines (SVM), LSTM, and neural network autoregressive (NNAR) applying to dengue time-series data. In the business and finance field, Siami-Namini and Namin (2018) [20] investigate how performance of ARIMA and LSTM in stock data of some companies between Jan 1985 and Aug 2018. Consequently, by comparing the RMSE metric, LSTM is confirmed to outperform the statistical-based algorithm, ARIMA. Due to the difficulty of obtaining parameters hydrogeological characteristics, Zhang et al. (2018) [21] proposes an approach that combines a fully connected layer associated with the dropout method on the top and LSTM layer to predict water table depth in the agricultural area. The result shows a higher R2 score comparing to a model with double LSTM layers [22]. Similarly, Kristiani et al. (2020) [23] claim a high correlation between PM2.5 and NO2, SO2 and prove an improved accuracy of forecasting PM2.5 by sequence to sequence LSTM model.

### *2.5. Related Works*

The air quality and disease cases dataset are belonging to spatiotemporal time-series data, which is a series of data points in time order and specific geographical (or spatial) points [24,25]. Spatial variation in disease cases has also demonstrated dependence patterns and noise levels in the data. Spatiotemporal analyzes have additional advantages as they simultaneously learn pattern persistence over time and investigate anomalies behavior [26]. The ability to interact with space and time might also detect data clustering, which leads to discover the emerging of environmental hazards [9,27]. The analytic results by using LSTM and ARIMA may aid in making early detection of IL1 and optimal preventive strategies for air pollution [28,29]. A comparative study between a model applies an ensemble approach with and without feature selection to predict the different classes of skin disease by Verma et al. (2019) [30]. An ensemble method combines six classification algorithms in order including passive-aggressive classifier, linear discriminant analysis, radius neighbor classifier, Bernoulli naive Bayesian, Gaussian naive Bayesian, and Extra Trees Classifier. The findings show efficient use in the detection of erythemato-squamous diseases with improvement in speed and accuracy.

## 3. Material and Methods

In this section, the system design of the system is presented in three subsections of the system architecture, the experimental workflows, and the parameter used in this paper.

### 3.1. System Environment

The specific environment is required to provide machine learning experiments. The dataset of air pollutants, ILI, and respiratory disease are trained using ARIMA and LSTM. TensorFlow's deep learning library was used to train air quality and disease history integration data. The library used Numpy and Pandas to process data manipulation and preprocessing. Matplotlib was used to visualize real and predictive information. Scikit-learn applied to evaluate the model's accuracy. Figure 1 describes the architecture of the environmental scheme used in the project.

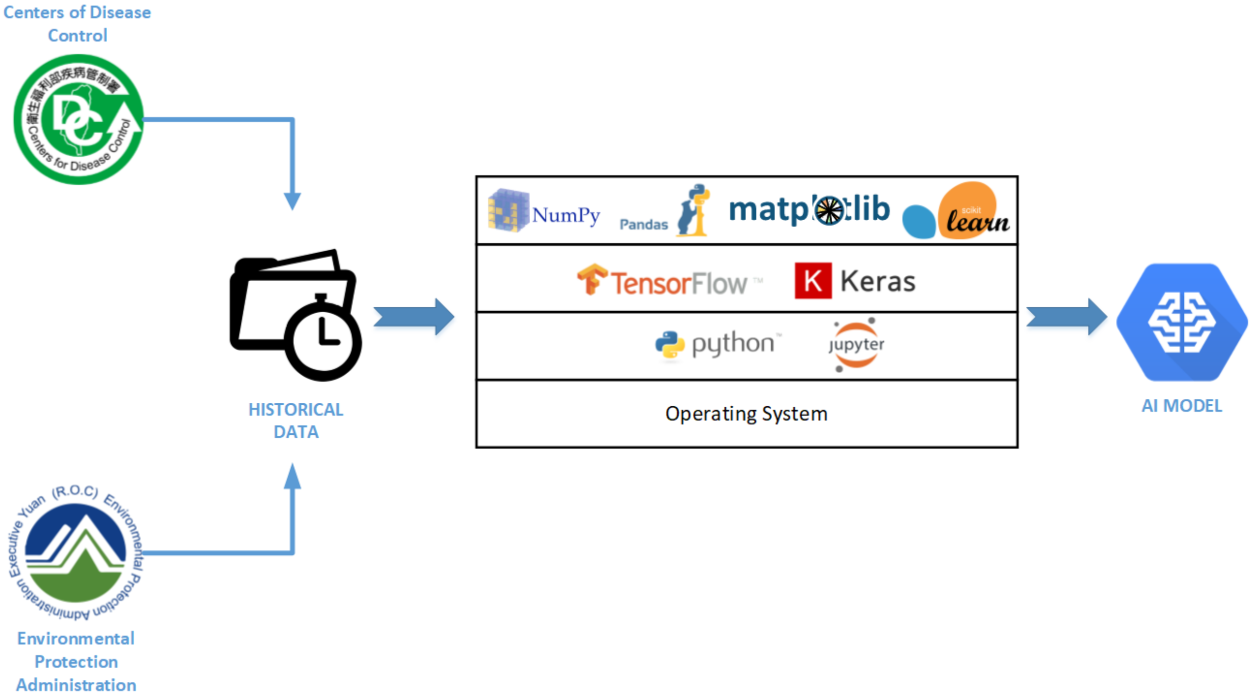


**Figure 1.** System Architecture.

### 3.2. System Workflow

Figure 2 describes the experimental procedure. First, the data were collected from the Taiwan Environmental Protection Administration (EPA) and The Taiwan Centers for Disease Control (CDC). Then, we did the data preprocessing by filling the missing data with 0 value and removing the outliers. The dataset was then divided into two categories of five years and ten years. For ARIMA model training, the data used weekly disease cases. For LSTM model training, we correlated the dataset with air pollutants. Therefore, we applied feature selection. There are three methods of feature selection that is Matrix Correlation (MC), Extra Trees Classifier (ETC), and Chemical (CHE) base. After the feature selection process, we train the dataset in five models of ARIMA, LSTM MC, LSTM ETC, LSTM CHE, and LSTM all Parameters.

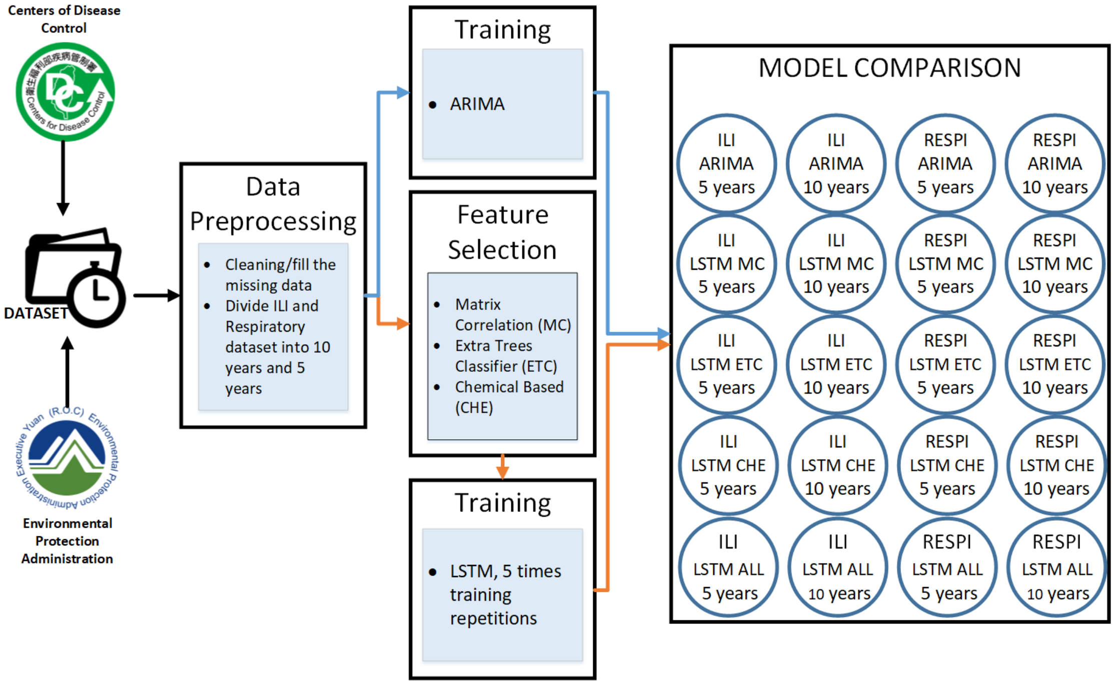


**Figure 2.** System Workflow.

### 3.3. *Experimental Procedure*

The experimental procedures are described in this section from data preprocessing, Feature Selection, ARIMA, and LSTM parameters.

#### Data Preprocessing

Two datasets are extracted from EPA [31] and CDC [32] from the year 2009 to 2018. There are two kinds of dataset, air quality/pollutants, and the number of ILI [33] and respiratory cases [34] in weekly from insurance resource. All subjects gave their informed consent for inclusion before they participated in the study. From these two datasets, we integrated the data and did data preprocessing, as follows:

- Convert the air quality dataset from hourly into daily by averaging all the parameters.
- Combine the air quality dataset daily with ILI and Respiratory dataset then convert onto weekly data for training process.
- Replace the missing data and the outliers by zero value.

In this case, we do not use all AQI parameters as three parameters of several stations are not available. Therefore, we removed three parameters, THC, CH4, and NMHC, to make the training process equally. All 15 parameters of air pollutant are described in Section 3.4.4. The detailed extracted dataset is 261 records for 5 years and 520 records for 10 years for the training process.

### 3.4. *Feature Selection*

Before the training process of LSTM, the feature selection is applied for dropping the number of input parameters when developing a predictive model. In this case, we implemented three categories of feature selection and 15 air pollutant parameters as follows:

#### 3.4.1. Matrix Correlation

A matrix correlation is a table representing the association coefficients for several variables. Any random variable in the table is strongly connected to each of the other values. It allows us to see which couples have the uppermost correlation [35]. Table 1 shows the 5 parameters feature selection of matrix correlation of ILI and Respiratory diseases.

**Table 1.** Matrix Correlation of ILI and Respiratory Disease.

| Order | ILI | Respiratory |
|---|---|---|
| 1 | AMB_TEMP | AMB_TEMP |
| 2 | NOx | NO2 |
| 3 | NO2 | NOx |
| 4 | NO | CO |
| 5 | CO | NO |

### 3.4.2. Extra Trees Classifier

Extremely Randomized Trees Classifier is a method of ensemble learning strategy which consolidates the results of multiple de-correlated decision trees obtained in a forest to generate its classification outcome. In concept, it is very similar to a Random Forest Classifier and only differs from it in the way decision trees are built in the forest. Table 2 shows the 5 parameters feature selection of Extra Trees Classifier of ILI and Respiratory diseases.

### 3.4.3. Feature Selection of Chemical Base

In general, PM2.5 can be separated into primary and secondary aerosols inside the atmosphere. Primary aerosols are generated directly by natural and anthropogenic contaminants, while the chemical reactions are occurring in the atmosphere release secondary aerosols. Secondary aerosols influence existing primary aerosols and relevant gaseous precursors. The development of secondary PM2.5 is closely linked to precursor gases NO2 and SO2. This association is illustrated in Lee et al. [36]'s 2005–2015 Taiwan Air Quality Studies. The U.S. Environmental Protection Agency also suggested using SO2, NO2/NOx, to assess the formation of secondary fine particulate matter (PM2.5) for modeling air dispersion. Table 3 listed the chemical correlation of feature selection for ILI and Respiratory disease.

**Table 2.** Extra Trees Classifier of ILI and Respiratory.

| Order | ILI | Respiratory |
|---|---|---|
| 1 | NO2 | NOx |
| 2 | O3 | WIND_SPEED |
| 3 | WIND_DIREC | AMB_TEMP |
| 4 | SO2 | PM10 |
| 5 | NOx | WS_HR |

**Table 3.** Chemical Base of Feature Selection.

| Order | Correlation |
|---|---|
| 1 | PM2.5 |
| 2 | PM10 |
| 3 | NO2 |
| 4 | SO2 |

### 3.4.4. All Parameters of Air Pollutant Dataset

The dataset extracted from the Taiwan Environmental Protection Administration (EPA) for the years 2009–2018 (10 years data). It consists of five stations, Xitun, Chungming, Fengyuan, Dali, and Shalu in Taichung City Taiwan. The features include 15 parameters that consists of 8 chemical and 7 meteorological pollutant factors [23], as described in Table 4.

**Table 4.** Dataset Air Pollutant Parameters.

| No. | Parameter | Unit | Name |
|---|---|---|---|
| **Chemical pollutant factors** | | | |
| 1 | CO | ppm | Carbon Monoxide |
| 2 | NO | ppb | Nitrogen Monoxide |
| 3 | NO2 | ppb | Nitrogen Dioxide |
| 4 | NOx | ppb | Nitrogen Oxide |
| 5 | O3 | ppb | Ozone |
| 6 | PM2.5 | $\mu g/m^3$ | fine aerosol |
| 7 | PM10 | $\mu g/m^3$ | aerosol |
| 8 | SO2 | ppb | Sulfur Dioxide |
| **Meteorogical factors** | | | |
| 1 | AMB_TEMP | Celsius | ambient temperature |
| 2 | RAINFALL | mm | Rainfall |
| 3 | RH | % | Relative Humidity |
| 4 | WD_HR | degrees | Wind Direction Hourly |
| 5 | WIND_DIREC | degrees | Wind Direction |
| 6 | WIND_SPEED | m/s | Wind Speed |
| 7 | WS_HR | m/s | Wind Speed Hourly |

### *3.5. Training Parameters*

There are two kinds of model training methods, that is ARIMA and LSTM. In this case, we intend to compare the model accuracy of these two methods in a particular environment.

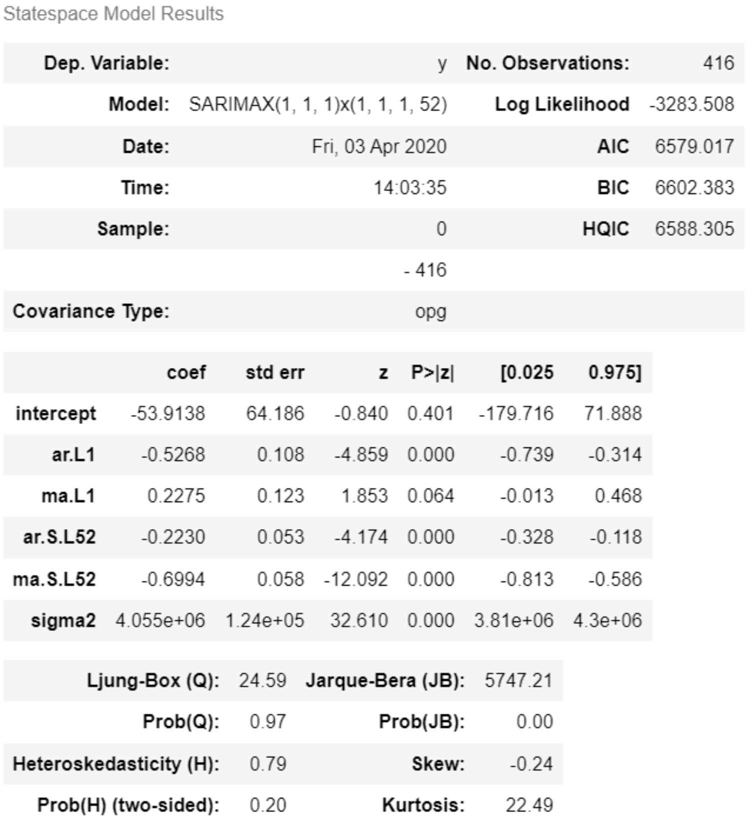
#### 3.5.1. ARIMA Model Training

An ARIMA model is used to determine if ILI and respiratory disease numbers could be predicted in advance. The case of the weekly illness is included in the dataset to predict the trend in the time series. In setting up the ARIMA model, the first 80% of data collected was utilized as training data, while the rest was used as validation data for creating the ARIMA model. The plots of autocorrelation and partial autocorrelation are then made. A *p*-value of less than 0.05 indicates that the null hypothesis of non-stationarity is rejected when a Dickey-Fuller test is performed. The pyramid library's auto-arima is then used to execute an ARIMA model. It was used to find the ARIMA model's best (p,d,q). ARIMA Training Procedures are described in Algorithm 1.

**Algorithm 1** ARIMA Training Methods

```
Require: ILI, Respiratory, and AQI Dataset integration
Ensure: Sum up the number of disease cases per week and order them
$tseriesdt = pd.Series(tseriesdt)$
$size = int(len(tseriesdt) * 0.8)$
$train, val = tseriesdt[0 : size], tseriesdt[size : len(tseriesdt)]$ ← Training and Validation data partition
$p - value < 0.05$ ← Execute Dickey-Fuller test to create p-value
$auto\_arima(train,$start_p=0, start_q=0, max_p=10, max_q=10, start_P=0, start_Q=0, max_P=10, max_Q=10, m=52, seasonal=True, trace=True, d=1, D=1, error_action='warn', suppress_warnings=True, random_state = 20, n_fits=30) ← Running the ARIMA model
$AIC$ ← Get Akaike's Information Criterion (AIC)
$RMSE$ ← Calculate the RMSE
```

Figure 3 describes the SARIMAX(1, 1, 1) × (1, 1, 1, 52) configuration was identified as the most optimal for modeling the time series based on the lowest AIC.



Statespace Model Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **No. Observations:** | 416 |
| **Model:** | SARIMAX(1, 1, 1)x(1, 1, 1, 52) | **Log Likelihood** | -3283.508 |
| **Date:** | Fri, 03 Apr 2020 | **AIC** | 6579.017 |
| **Time:** | 14:03:35 | **BIC** | 6602.383 |
| **Sample:** | 0 | **HQIC** | 6588.305 |
| | - 416 | | |
| **Covariance Type:** | opg | | |

| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **intercept** | -53.9138 | 64.186 | -0.840 | 0.401 | -179.716 | 71.888 |
| **ar.L1** | -0.5268 | 0.108 | -4.859 | 0.000 | -0.739 | -0.314 |
| **ma.L1** | 0.2275 | 0.123 | 1.853 | 0.064 | -0.013 | 0.468 |
| **ar.S.L52** | -0.2230 | 0.053 | -4.174 | 0.000 | -0.328 | -0.118 |
| **ma.S.L52** | -0.6994 | 0.058 | -12.092 | 0.000 | -0.813 | -0.586 |
| **sigma2** | 4.055e+06 | 1.24e+05 | 32.610 | 0.000 | 3.81e+06 | 4.3e+06 |

| | | | |
|---|---|---|---|
| **Ljung-Box (Q):** | 24.59 | **Jarque-Bera (JB):** | 5747.21 |
| **Prob(Q):** | 0.97 | **Prob(JB):** | 0.00 |
| **Heteroskedasticity (H):** | 0.79 | **Skew:** | -0.24 |
| **Prob(H) (two-sided):** | 0.20 | **Kurtosis:** | 22.49 |

**Figure 3.** ARIMA model  statespace.

### 3.5.2. LSTM Model Training

LSTMs are sequential neural networks that assume dependence between the observations in a given series. Such as, they have been widely used for forecasting purposes of the time series. In configuring the LSTM model, the first 100 observation data from the time series dataset was called. After that, a dataset matrix is produced. The data is then adjusted using MinMaxScaler so that the neural network can appropriately perceive it. With the previous option set to 5, the data is divided into training and test sets. When the "previous" parameter is set to this, the values $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$ (all under X train) are used to forecast the value at time $t$ (Y train for the training data). Following that, 150 epochs are completed. LSTM Training Procedures are described in Algorithm 2.

Figure 4 describes the layer's structure of LSTM model used in this paper.

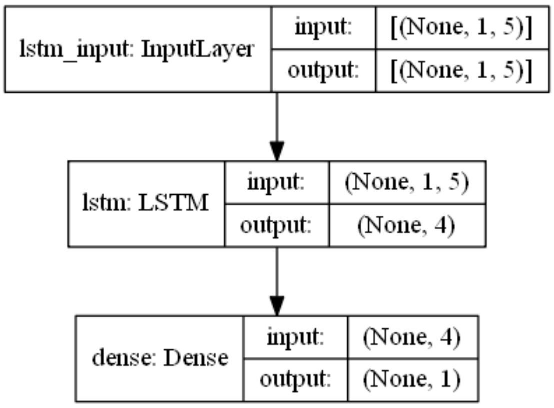


**Figure 4.** LSTM model architecture.

The training models are divided by four conditions of feature selection, as follows:

- Matrix Correlation
- Extra Trees Classifier
- Chemical reactions happening in the atmosphere
- All parameters (15 parameters of air pollutant)

**Algorithm 2** LSTM Training Methods

```
Require: ILI, Respiratory, and AQI Dataset integration
Ensure: Sum up the number of disease cases per week and order them
$df = df[:100]$ ← Call the first 100 observation data
$MinMaxScaler$ ← Normalize dataset with MinMaxScaler
$trainsize = int(len(df) * 0.8)$
$valsize = len(df) - trainsize$
$train, val = df[0 : trainsize, :], df[trainsize : len(df), :]$ ← Training and validation data partition
$previous = 5$ ←Number of previous parameter
$X\_train, Y\_train = create\_dataset(train, previous)$
$X\_val, Y\_val = create\_dataset(val, previous)$
$model = tf.keras.Sequential()$ ← Create LSTM model
$model.add(LSTM(4, input\_shape = (1, previous)))$ ← Add LSTM layers with 4 hidden layers
$model.add(Dense(1))$ ← Add Dense Layer
$model.compile$(loss='mean_squared_error', optimizer='adam') ← Compile the model using Adam optimizer
$model.fit$(X_train, Y_train, epochs=150, batch_size=1, verbose=2) ← Train the model in 150 epochs
$RMSE$ ← Calculate the RMSE
```

## 4. Results

In this section, the experimental results are presented which is comparing the forecasting result between ARIMA and LSTM.

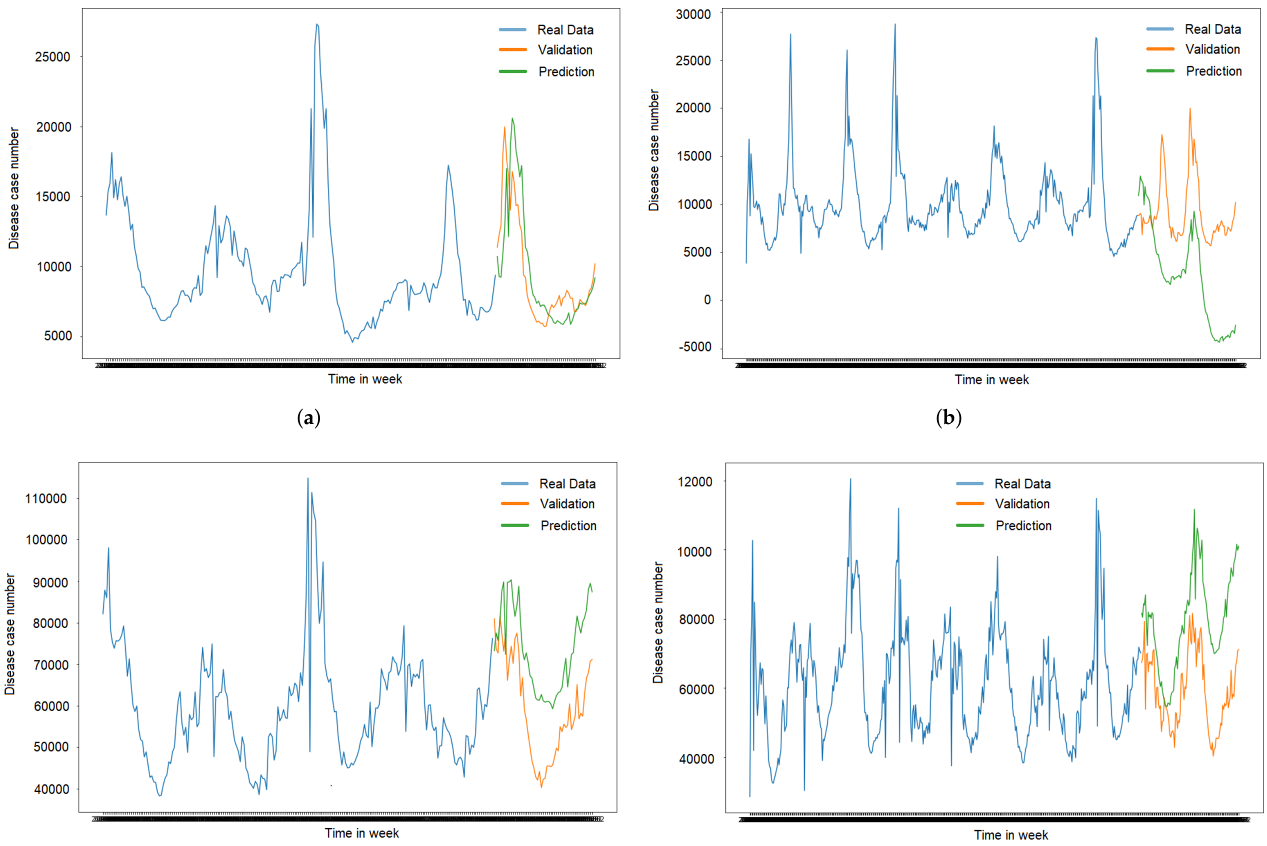
### 4.1. ARIMA Prediction Results

With 80% of the time series dataset used as the training data to build the ARIMA model, the remaining 20% is used to test the predictions of the model. Table 5 describes the comparison of RMSE value of four ARIMA models. The table shows that ARIMA models using five years data have a lower RMSE value at 2174.8 and 5581.6, compared to 10 years of data usage at 8173.6 and 15,093.1.

**Table 5.** The RMSE of ARIMA Model Comparison.

| ARIMA Model | RMSE |
|---|---|
| ILI ARIMA 5 years | 2174.8 |
| ILI ARIMA 10 years | 8173.6 |
| Respiratory ARIMA 5 years | 5581.6 |
| Respiratory ARIMA 10 years | 15,093.1 |

The following tables describe the actual data compared to validation and prediction. The blue line represented the real data used for the training. The orange line is the validation for the forecast, and the green line shows the prediction of the validation data. It can be seen from the graph's trends, The trend's line in Figure 5a has the closed prediction compared to Figure 5b–d. It is because of the model in Figure 5a has smallest RMSE value.



(**a**) (**b**)

(**c**) (**d**)

**Figure 5.** ILI and Respiratory ARIMA models. (**a**) ILI from 5 years dataset training; (**b**) ILI from 10 years dataset training; (**c**) Respiratory from 5 years dataset training; (**d**) Respiratory from 10 years dataset training.

### *4.2. LSTM Prediction Results*

In the LSTM training models, the feature selections are applied with five times repetitions. For example, Table 6 presented the ILI RMSE in 5 and 10 years data. The data in 10 years has more low error than five years data. The mean of all data, MC, ETC, and CHE for ten years dataset have the RMSE values at 506.2, 473.5, 573.6, and 515.1, respectively. While in 5 years dataset, the RMSE increased about 30% at 730.6, 707.9, 728.7, and 713.6, respectively.

Table 7 described the Respiratory RMSE in 5 and 10 years data. Like in ILI case, The data in 10 years has more low error than five years data. The mean of all data, MC, ETC, and CHE for ten years dataset have the RMSE values at 3258, 3064.7, 3186.7, and 3507.7, respectively. While in 5 years dataset, the RMSE increased about 25% at 4752.1, 4750.3, 4798.1, and 4773.7, respectively.

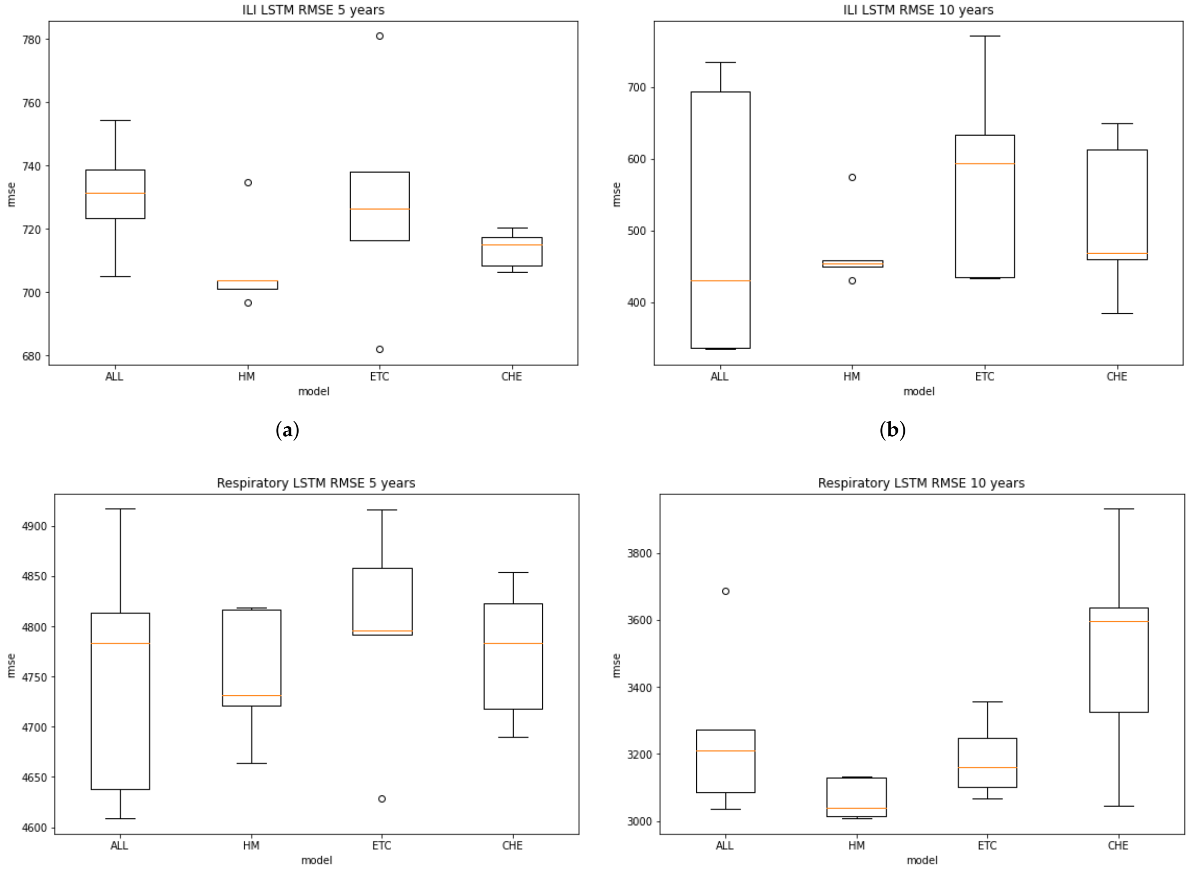
**Table 6.** ILI RMSE using 5 and 10 years data.

| Training Repetitions | ILI LSTM RMSE 5 years | | | |
|---|---|---|---|---|
| | ALL | MC | ETC | CHE |
| 1 | 754.3 | 703.6 | 780.9 | 706.4 |
| 2 | 705.1 | 701.0 | 682.1 | 717.4 |
| 3 | 738.8 | 703.6 | 738.0 | 715.2 |
| 4 | 723.3 | 696.7 | 726.3 | 708.5 |
| 5 | 731.5 | 734.8 | 716.3 | 720.4 |
| **mean** | **730.6** | **707.9** | **728.7** | **713.6** |
| **std** | **18.3** | **15.3** | **35.9** | **5.9** |
| **Training Repetitions** | **ILI LSTM RMSE 10 years** | | | |
| | **ALL** | **MC** | **ETC** | **CHE** |
| 1 | 336.19 | 574.67 | 433.57 | 459.57 |
| 2 | 734.85 | 449.82 | 594.45 | 385.61 |
| 3 | 431.1 | 458.10 | 633.23 | 612.43 |
| 4 | 335.31 | 430.23 | 770.90 | 649.33 |
| 5 | 693.32 | 454.50 | 435.67 | 468.72 |
| **mean** | **506.2** | **473.5** | **573.6** | **515.1** |
| **std** | **194.3** | **57.6** | **142.8** | **111.2** |

**Table 7.** Respiratory RMSE using 5 and 10 years data.

| Training Repetitions | RESPIRATORY LSTM RMSE 5 years | | | |
|---|---|---|---|---|
| | ALL | MC | ETC | CHE |
| 1 | 4813.9 | 4731.2 | 4629.1 | 4690.1 |
| 2 | 4637.6 | 4817.0 | 4916.0 | 4783.7 |
| 3 | 4608.9 | 4720.7 | 4858.2 | 4853.7 |
| 4 | 4916.9 | 4818.8 | 4795.5 | 4718.4 |
| 5 | 4783.1 | 4663.6 | 4791.6 | 4822.4 |
| **mean** | **4752.1** | **4750.3** | **4798.1** | **4773.7** |
| **std** | **128.0** | **66.9** | **107.4** | **68.7** |
| **Training Repetitions** | **RESPIRATORY LSTM RMSE 10 years** | | | |
| | **ALL** | **MC** | **ETC** | **CHE** |
| 1 | 3273.5 | 3015.2 | 3355.9 | 3046.1 |
| 2 | 3084.6 | 3130.4 | 3160.6 | 3596.6 |
| 3 | 3210.2 | 3131.0 | 3247.6 | 3326.3 |
| 4 | 3687.0 | 3007.9 | 3066.8 | 3931.4 |
| 5 | 3034.6 | 3038.8 | 3102.6 | 3638.3 |
| **mean** | **3258.0** | **3064.7** | **3186.7** | **3507.7** |
| **std** | **258.1** | **61.4** | **116.7** | **335.6** |

From the five times repetition, we can draw the boxplot to present the distribution of RMSE values. A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can also tell about the outliers and what their values are. Figure 6 describes the distribution of the data in boxplot.



(**a**) (**b**)

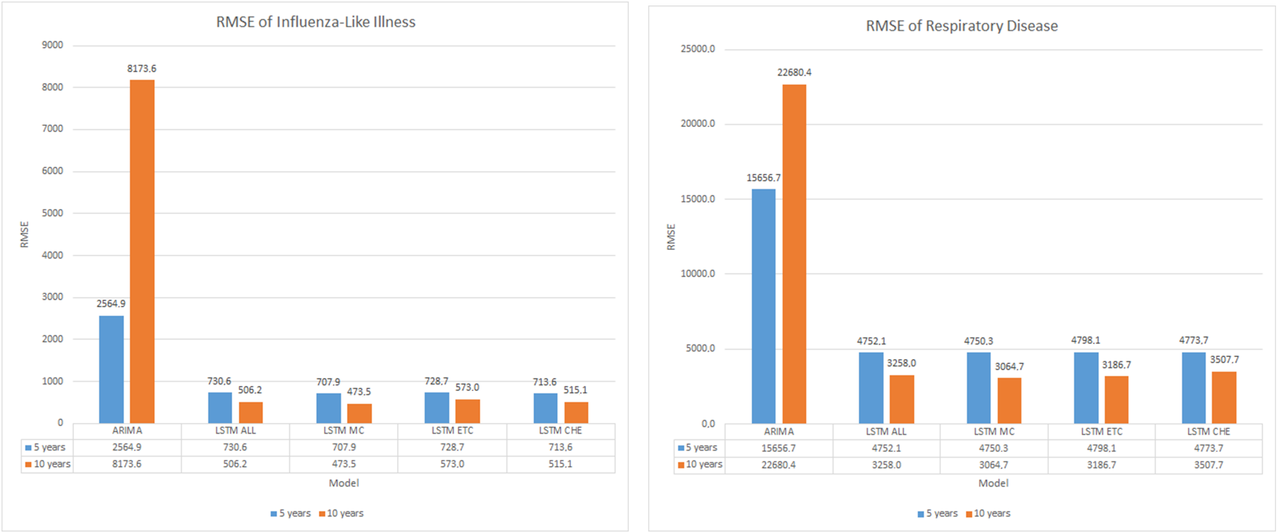
(**c**) (**d**)

**Figure 6.** The bloxplot prediction distribution in 5 times repetitions. (**a**) ILI LSTM boxplot from 5 years dataset training; (**b**) ILI LSTM boxplot from 10 years dataset training; (**c**) Respiratory LSTM boxplot from 5 years dataset training; (**d**) Respiratory LSTM boxplot from 10 years dataset training.

### *4.3. Model Comparison*

The entire models comparison are presented in the graph Figure 7a,b, as follows.

It can be seen from both graphs, Univariate ARIMA have significant error rate than multivariate LSTM models. ARIMA RMSE values have 4 to 10 times more high compared to LSTM.

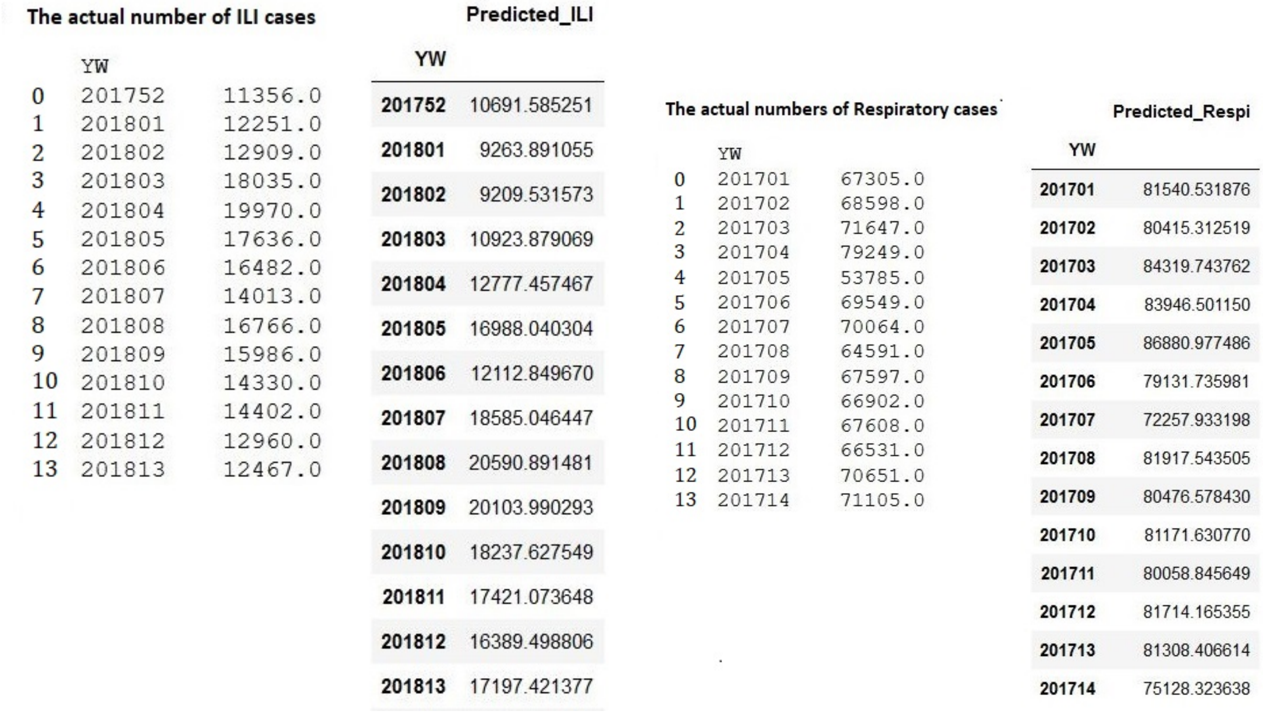


(**a**) (**b**)

**Figure 7.** The model comparison results in graphs. (**a**) The RMSE of Influenza-like Illness; (**b**) The RMSE of Respiratory Disease.

### *4.4. Sampling and Result Analysis*

In this section, we demonstrate the several prediction results of ARIMA and LSTM models. For ARIMA, we take the example of two different models. First, Arima ILI model prediction of five years dataset with RMSE value at 2564.9 in Figure 8a. Second, ARIMA Respiratory model prediction of ten years dataset with RMSE value at 22680.4 in Figure 8b.



**The actual number of ILI cases**

```
    YW
0   201752    11356.0
1   201801    12251.0
2   201802    12909.0
3   201803    18035.0
4   201804    19970.0
5   201805    17636.0
6   201806    16482.0
7   201807    14013.0
8   201808    16766.0
9   201809    15986.0
10  201810    14330.0
11  201811    14402.0
12  201812    12960.0
13  201813    12467.0
```

| | Predicted_ILI |
|---|---|
| **YW** | |
| **201752** | 10691.585251 |
| **201801** | 9263.891055 |
| **201802** | 9209.531573 |
| **201803** | 10923.879069 |
| **201804** | 12777.457467 |
| **201805** | 16988.040304 |
| **201806** | 12112.849670 |
| **201807** | 18585.046447 |
| **201808** | 20590.891481 |
| **201809** | 20103.990293 |
| **201810** | 18237.627549 |
| **201811** | 17421.073648 |
| **201812** | 16389.498806 |
| **201813** | 17197.421377 |

**The actual numbers of Respiratory cases**

```
    YW
0   201701    67305.0
1   201702    68598.0
2   201703    71647.0
3   201704    79249.0
4   201705    53785.0
5   201706    69549.0
6   201707    70064.0
7   201708    64591.0
8   201709    67597.0
9   201710    66902.0
10  201711    67608.0
11  201712    66531.0
12  201713    70651.0
13  201714    71105.0
```

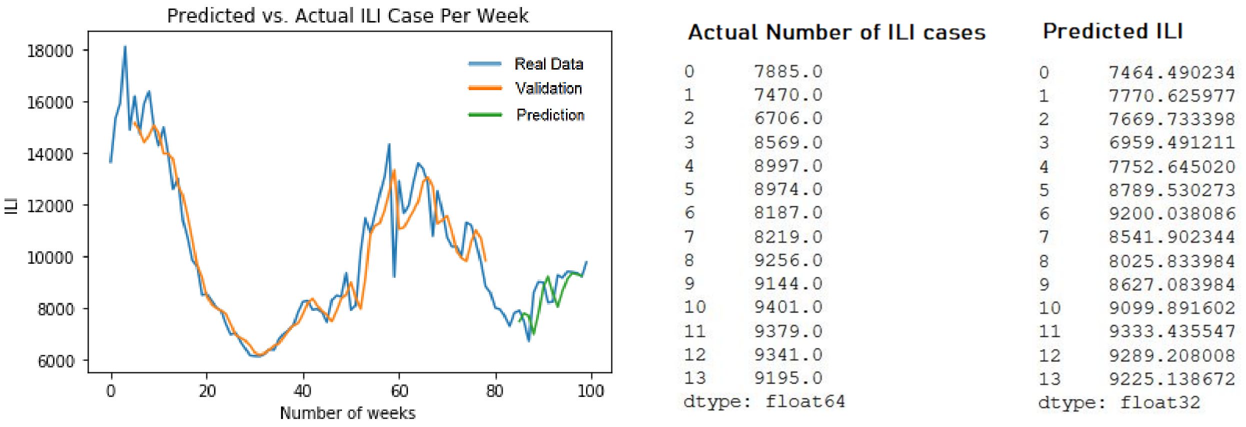
| | Predicted_Respi |
|---|---|
| **YW** | |
| **201701** | 81540.531876 |
| **201702** | 80415.312519 |
| **201703** | 84319.743762 |
| **201704** | 83946.501150 |
| **201705** | 86880.977486 |
| **201706** | 79131.735981 |
| **201707** | 72257.933198 |
| **201708** | 81917.543505 |
| **201709** | 80476.578430 |
| **201710** | 81171.630770 |
| **201711** | 80058.845649 |
| **201712** | 81714.165355 |
| **201713** | 81308.406614 |
| **201714** | 75128.323638 |

(**a**) (**b**)

**Figure 8.** The Prediction sampling of ARIMA. (**a**) ARIMA ILI 5 years; (**b**) ARIMA Respiratory 10 years.

For the LSTM model, we pick up two different models. First, LSTM ILI model prediction of five years dataset with RMSE value at 473.5 in Figure 9b. Second, the LSTM

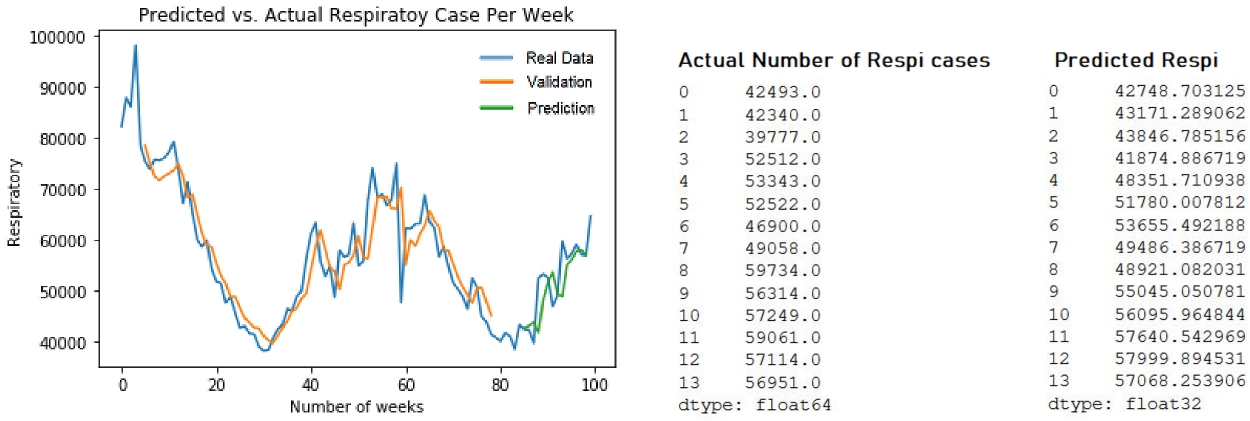
Respiratory model prediction of five years dataset with RMSE value at 4798.1 in Figure 9b. The training and evaluation of each models are described in Figures 9a and 10a.



```
Actual Number of ILI cases          Predicted ILI
0      7885.0                       0      7464.490234
1      7470.0                       1      7770.625977
2      6706.0                       2      7669.733398
3      8569.0                       3      6959.491211
4      8997.0                       4      7752.645020
5      8974.0                       5      8789.530273
6      8187.0                       6      9200.038086
7      8219.0                       7      8541.902344
8      9256.0                       8      8025.833984
9      9144.0                       9      8627.083984
10     9401.0                       10     9099.891602
11     9379.0                       11     9333.435547
12     9341.0                       12     9289.208008
13     9195.0                       13     9225.138672
dtype: float64                      dtype: float32
```

(**a**) (**b**)

**Figure 9.** ILI LSTM sampling predictions. (**a**) ILI LSTM MC model for 10 years dataset training; (**b**) The Prediction Result of LSTM MC 10 years.



```
Actual Number of Respi cases         Predicted Respi
0      42493.0                       0      42748.703125
1      42340.0                       1      43171.289062
2      39777.0                       2      43846.785156
3      52512.0                       3      41874.886719
4      53343.0                       4      48351.710938
5      52522.0                       5      51780.007812
6      46900.0                       6      53655.492188
7      49058.0                       7      49486.386719
8      59734.0                       8      48921.082031
9      56314.0                       9      55045.050781
10     57249.0                       10     56095.964844
11     59061.0                       11     57640.542969
12     57114.0                       12     57999.894531
13     56951.0                       13     57068.253906
dtype: float64                       dtype: float32
```

(**a**) (**b**)

**Figure 10.** Respiratory LSTM sampling predictions. (**a**) Respiratory LSTM ALL model for 5 years dataset training; (**b**) The Prediction Result of LSTM ALL 5 years.

## 5. Discussion

From the training model experiments, it can be seen that ARIMA has an excellent model in the five-year dataset of ILI at 2564.9 compared to ten years at 8173.6 of RMSE value. This accuracy is similar to the Respiratory dataset, which gets 15,656.7 in the five-year dataset and 22,680.4 of RMSE value in the ten-year dataset. On the contrary, LSTM has better accuracy in the ten-year dataset compared to the five-year dataset. On average RMSE in the ILI dataset, LSTM has 720.2 RMSE value in five years and 517.0 in ten years dataset. Also, in the Respiratory disease dataset, LSTM gets 4768.6 of five years of data and 3254.3 of the ten-year dataset. In terms of the LSTM various parameters to be trained using feature selection, there are no significant differences among the RMSE values. These analyses revealed that the LSTM model typically outperforms ARIMA by three to seven times higher accuracy.

From the ILI boxplot of 5-year and 10-year period data, it can be seen that the MC model in the two periods has a minor interquartile range showing the most stable prediction compared with the other models. The trend also happens in the respiratory model, in which the MC boxplot indicates the smallest range of interquartile. However, the models will keep changing as the new data arrive. Therefore, for production implementation, it needs dynamics retraining. The methods can vary, such as triggering the retraining by regular scheduling or a particular RMSE threshold calculation.

## 6. Conclusions and Future Works

The influenza-like illness and respiratory disease were predicted by comparative models of ARIMA and LSTM for 5 and 10 year periods. While ARIMA captured a variety of typical temporal structures of the diseases that change over time in the training model, LSTM carried out some feature selection techniques, including matrix correlation, extra trees classifier, and chemical-based before training input data. As a result, three LSTM-based training types were formed, two of which selected the top five highest-ranking data. The remaining type, chemical-based, used four parameters, which are the main contributors to the diseases. Thus, the study conducted comparisons of the disease's prediction between ARIMA, ordinary LSTM, and three LSTM models associated with feature selection techniques during five years and ten years. As a result, there is a highly consistent trend of the prediction models for ILI and respiratory cases. Besides, all four LSTM prediction models present a higher accuracy than ARIMA's outcome in both 5 and 10 year periods of ILI and respiratory diseases. Comparing two periods of time, the prediction model of ARIMA for five years is more accurate than one for ten years, whereas the ten-year prediction models using LSTM got the opposite result, the result of ten-year prediction models reach a higher accuracy than five-year models'. Among the four LSTM-based models, the matrix correlation technique's outcomes achieve the most accuracy, and chemical-based feature selection is also a promising approach. The chemical-based show is not much different from the matrix correlation's results, and its four selected factors are based on fundamental knowledge.

In the future, the model's implementation can be applied in further development such as ARIMAX to provide the balance input parameters for a statistically-based model. Also, the models can be integrated along with the sensor. Therefore, a real-time response system can be activated to mitigate disease outbreaks, such as the recent COVID-19 pandemic infection in advanced time.

**Author Contributions:** Conceptualization, Y.-T.T., C.-T.Y. and D.-Y.C.; methodology, Y.-T.T., C.-T.Y. and D.-Y.C.; software, E.K. and K.L.P.N.; validation, Y.-T.T. and P.-Y.L.; formal analysis, Y.-T.T. and P.-Y.L.; investigation, C.-T.Y. and D.-Y.C.; resources, Y.-T.T. and P.-Y.L.; data curation, E.K. and K.L.P.N.; writing—original draft preparation, E.K. and K.L.P.N.; writing—review and editing, C.-T.Y., D.-Y.C., Y.-T.T. and P.-Y.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was sponsored by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. 110-2221-E-029-020-MY3, 110-2621-M-029-003, 110-2811-E-029-003, 110-2622-E-029-003, TCVGH-T1087804, TCVGH-T1097801, TCVGH-T1107803, TCVGH-1107201C, and TCVGH-T1117803.

**Conflicts of Interest:** The authors declare that there is no conflict of interest.

## References

1. Yoon, Y.K.; Park, C.S.; Kim, J.W.; Hwang, K.; Lee, S.Y.; Kim, T.H.; Park, D.Y.; Kim, H.J.; Kim, D.Y.; Lee, H.J.; et al. Guidelines for the antibiotic use in adults with acute upper respiratory tract infections. *Infect. Chemother.* **2017**, *49*, 326–352. [CrossRef]
2. Zhu, D.; Cai, C.; Yang, T.; Zhou, X. A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. *Big Data Cogn. Comput.* **2018**, *2*, 5. [CrossRef]
3. Yang, C.T.; Chen, C.J.; Tsan, Y.T.; Liu, P.Y.; Chan, Y.W.; Chan, W.C. An implementation of real-time air quality and influenza-like illness data storage and processing platform. *Comput. Hum. Behav.* **2018**, *100*, 266–274 . [CrossRef] [PubMed]
4. Kristiani, E.; Chen, Y.A.; Yang, C.T.; Huang, C.Y.; Tsan, Y.T.; Chan, W.C. Using deep ensemble for influenza-like illness consultation rate prediction. *Future Gener. Comput. Syst.* **2021**, *117*, 369–386. [CrossRef]
5. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
6. Panda, D.; Ray, R.; Abdullah, A.A.; Dash, S.R. Predictive Systems: Role of Feature Selection in Prediction of Heart Disease. *J. Phys. Conf. Ser.* **2019**, *1372*, 012074. [CrossRef]

7. Zhai, B.; Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci. Total Environ.* **2018**, *635*, 644–658. [CrossRef] [PubMed]
8. Sethi, J.K.; Mittal, M. A new feature selection method based on machine learning technique for air quality dataset. *J. Stat. Manag. Syst.* **2019**, *22*, 697–705. [CrossRef]
9. Xie, Y.; Zhao, B.; Zhang, L.; Luo, R. Spatiotemporal variations of $PM_{2.5}$ and $PM_{10}$ concentrations between 31 Chinese cities and their relationships with $SO_2$, $NO_2$, CO and $O_3$. *Particuology* **2015**, *20*, 141–149. [CrossRef]
10. Chen, C.W.; Hsieh, Y.H.; Su, H.C.; Wu, J.J. Causality test of ambient fine particles and human influenza in Taiwan: Age group-specific disparity and geographic heterogeneity. *Environ. Int.* **2018**, *111*, 354–361. [CrossRef]
11. Liu, P.Y.; Tsan, Y.T.; Chan, Y.W.; Chan, W.C.; Shi, Z.Y.; Yang, C.T.; Lou, B.S. Associations of PM2.5 and aspergillosis: Ambient fine particulate air pollution and population-based big data linkage analyses. *J. Ambient. Intell. Humaniz. Comput.* **2018**, 1–11. [CrossRef]
12. World Health Organization. Ambient Air Pollution. 2016. Available online: https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/ambient-air-pollution (accessed on 20 February 2019).
13. Yang, C.T.; Chen, Y.A.; Chan, Y.W.; Lee, C.L.; Tsan, Y.T.; Chan, W.C.; Liu, P.Y. Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources. *J. Supercomput.* **2020**, *76*, 9303–9329. [CrossRef]
14. Genc, S.; Zadeoglulari, Z.; Fuss, S.H.; Genc, K. The adverse effects of air pollution on the nervous system. *J. Toxicol.* **2012**, *2012*, 782462. [CrossRef]
15. Kampa, M.; Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **2008**, *151*, 362–367. [CrossRef]
16. Özkaynak, H.; Baxter, L.K.; Dionisio, K.L.; Burke, J. Air pollution exposure prediction approaches used in air pollution epidemiology studies. *J. Expo. Sci. Environ. Epidemiol.* **2013**, *23*, 566–572. [CrossRef]
17. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **2018**, *178*, 1544–1547. [CrossRef]
18. Eymen, A.; Köylü, Ü. Seasonal trend analysis and ARIMA modeling of relative humidity and wind speed time series around Yamula Dam. *Meteorol. Atmos. Phys.* **2019**, *131*, 601–612. [CrossRef]
19. Chakraborty, T.; Chattopadhyay, S.; Ghosh, I. Forecasting dengue epidemics using a hybrid methodology. *Phys. Stat. Mech. Its Appl.* **2019**, *527*, 121266. [CrossRef]
20. Siami-Namini, S.; Namin, A.S. Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv* **2018**, arXiv:1803.06386.
21. Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* **2018**, *561*, 918–929. [CrossRef]
22. de Myttenaere, A.; Golden, B.; Grand, B.L.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* **2016**, *192*, 38–48. [CrossRef]
23. Kristiani, E.; Yang, C.T.; Huang, C.Y.; Lin, J.R.; Nguyen, K.L.P. PM2.5 Forecasting Using LSTM Sequence to Sequence Model in Taichung City. In *Information Science and Applications*; Springer: Singapore, 2020; pp. 497–507.
24. Yang, C.T.; Chan, Y.W.; Liu, J.C.; Lou, B.S. An implementation of cloud-based platform with R packages for spatiotemporal analysis of air pollution. *J. Supercomput.* **2017**, *76*, 1416–1437. [CrossRef]
25. Lee, C.F.; Yang, C.T.; Kristiani, E.; Tsan, Y.T.; Chan, W.C.; Huang, C.Y. Recurrent Neural Networks for Analysis and Automated Air Pollution Forecasting. In Proceedings of the International Conference on Frontier Computing, Kuala Lumpur, Malaysia, 8–10 May 2018; Springer: Singapore, 2018; pp. 50–59.
26. Liu, L.; Yang, X.; Liu, H.; Wang, M.; Welles, S.; Márquez, S.; Frank, A.; Haas, C.N. spatial–Temporal analysis of air Pollution, climate change, and Total Mortality in 120 cities of china, 2012–2013. *Front. Public Health* **2016**, *4*, 143. [CrossRef] [PubMed]
27. Yang, C.T.; Chen, S.T.; Den, W.; Wang, Y.T.; Kristiani, E. Implementation of an Intelligent Indoor Environmental Monitoring and management system in cloud. *Future Gener. Comput. Syst.* **2018**, *96*, 731–749. [CrossRef]
28. Boldo, E.; Linares, C.; Lumbreras, J.; Borge, R.; Narros, A.; García-Pérez, J.; Fernández-Navarro, P.; Pérez-Gómez, B.; Aragonés, N.; Ramis, R.; et al. Health impact assessment of a reduction in ambient PM2.5 levels in Spain. *Environ. Int.* **2011**, *37*, 342–348. [CrossRef]
29. Yazdi, M.D.; Wang, Y.; Di, Q.; Zanobetti, A.; Schwartz, J. Long-term exposure to PM2.5 and ozone and hospital admissions of Medicare participants in the Southeast USA. *Environ. Int.* **2019**, *130*, 104879. [CrossRef]
30. Verma, A.K.; Pal, S.; Kumar, S. Prediction of skin disease using ensemble data mining techniques and feature selection method—A comparative study. *Appl. Biochem. Biotechnol.* **2020**, *190*, 341–359. [CrossRef]
31. Taiwan Environmental Protection Administration. Taiwan Air Quality Dataset. Available online: https://data.epa.gov.tw/en/dataset (accessed on 4 January 2019).
32. Center for Diseases Control Ministry of Health and Welfare. Taiwan CDC Open Data Portal. 2016. Available online: https://data.cdc.gov.tw/en/dataset/ (accessed on 4 January 2019).
33. Center for Diseases Control and Welfare. Health Insurance Outpatient and Emergency Visits—Influenza. 2016. Available online: https://data.cdc.gov.tw/en/dataset/hi-outpatient-emergency-visit-influenza (accessed on 4 January 2019).
34. Center for Diseases Control and Welfare. Area, Age, and Gender Statistical Table-Influenza. 2016. Available online: https://data.cdc.gov.tw/en/dataset/aagstable-influenza (accessed on 4 January 2019).

35. Statistics How To. Correlation Matrix: Definition. 2016. Available online: https://www.statisticshowto.com/correlation-matrix/ (accessed on 21 May 2020).
36. Lee, C.S.; Chang, K.H.; Kim, H. Long-term (2005–2015) trend analysis of PM2.5 precursor gas $NO_2$ and $SO_2$ concentrations in Taiwan. *Environ. Sci. Pollut. Res.* **2018**, *25*, 22136–22152. [CrossRef]