# A Novel Data-driven Methodology for Influenza Outbreak Detection and Prediction

**Lin Du**
Zuellig Pharma Holdings Pte. Ltd.
**Yan Pang** ( jamespang@nus.edu.sg )
Department of Analytics and Operations, National University of Singapore, 119613, SG

**Research Article**

**Keywords:** Influenza outbreak detection, Influenza outbreak prediction, Public health, Machine learning, Data-driven method

**Posted Date:** November 24th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-109171/v1

# A novel data-driven methodology for influenza outbreak detection and prediction

**LIN DU**[1,3] **and YAN PANG**[1,2,*]

[1]Business Analytics Centre, National University of Singapore, 119613, SG email: dulin@u.nus.edu
[2]Department of Analytics and Operations, National University of Singapore, 119613, SG
[3]Data Analytics, Zuellig Pharma Holdings Pte. Ltd., 228233, SG
[*]Corresponding author email: jamespang@nus.edu.sg
This work was supported by Zuellig Pharma Analytics and NUS Business Analytics Centre.

## ABSTRACT

Influenza is an infectious disease that leads to an estimated 5 million severe illness cases and 650,000 respiratory deaths worldwide each year. Early detection and prediction of influenza outbreaks are crucial to efficient resource planning to save patients' lives and healthcare costs. This paper proposes a novel data-driven methodology for influenza outbreaks detection and prediction. The doctor's diagnosis-based prescription dataset of Influenza-Like Illness (ILI) from more than 3,000 clinics in Malaysia is used in this study because the prescription data are reliable and can be captured timely. A new Region Index (RI) of the influenza outbreak is proposed based on the prescription dataset. With the newly proposed RI metric, statistical and machine learning models are developed to detect and predict influenza outbreaks. Cross-validation is conducted to evaluate the prediction model performance. The proposed methods are also validated by real-world evidence. It is proved to be sensitive and accurate in influenza outbreak prediction with $80-90\%$ accuracy, $70-80\%$ recall, and $70-80\%$ precision scores.

keywords: Influenza outbreak detection, Influenza outbreak prediction, Public health, Machine learning, Data-driven method

## Introduction

The World Health Organization (WHO) released the top 10 issues demanding attention in 2019, of which Infection diseases dominate the list. Influenza is one of them[1]. Influenza is a highly contagious respiratory tract infection that causes diseases ranging from mild respiratory tract infection (RTI) to severe pneumonia and even death. Worldwide, seasonal influenza leads to estimated up to 5 million severe illness cases and 650,000 respiratory deaths every year[2]. It also causes a significant burden on hospitalization, workplace absences, and productivity. For example, based on a study conducted at the University of Malaya Medical Centre in Malaysia in 2009, the direct healthcare costs for each hospitalized H1N1 patient was USD 510, which was 60% higher than the year 2007's per capita national expenditure on health of USD 318[3]. Given the circumstances, people are looking into detecting and predicting the influenza outbreaks early. This would bring tremendous value to the world's healthcare systems. Firstly, the early detection of influenza outbreaks is crucial to the healthcare system to enable efficient resource planning and save healthcare costs. Secondly, early detection can potentially help save people's lives. Thirdly, we can control the spreading of influenza if we can predict it early.

Traditional surveillance is widely used to monitor the anomalies of influenza-like illness (ILI) cases in selected hospitals or clinics. For example, the Malaysia Ministry of Health designs a system for monitoring influenza. One to two clinics are selected per state as sentinel sites, to conduct both clinical-based and laboratory-based surveillance. The system is designed to monitor influenza efficiently with low cost, to provide national trend data for ILI[4]. However, traditional surveillance usually requires weeks even months, to gather, process and release surveillance data[5]. In addition, the ILI trend may not be captured accurately due to a limited number of clinics monitored in the system.

In recent years, there were more research papers on ILI cases or influenza outbreak prediction. Many papers utilized data of historical ILI cases from traditional surveillance or WHO reports. These data had its limitation of low geographic coverage and small sample size. Some other papers utilized simulated data or Google search data. Google Flu Trends (GFT) was launched in 2008 to provide influenza estimated activities using google searches[6]. GFT provided near real-time estimates of seasonal influenza activity each day and stimulated many innovative research works in influenza outbreak detection using new data sources. In 2013, Andrea investigated the use of GARMA and developed an influenza forecast model incorporating a real-time influenza surveillance system and GFT[7]. The model provided individual medical centers with a warning of the expected number of influenza cases. It was shown that with GFT as an external variable, the model's forecast confidence was improved from 81% to 83%. However, it was tested on only one medical center. Hence geographic generalizability needs to be further evaluated.

In 2015, García YE utilized Bayesian model selection and Bayesian regression to detect outbreaks for ILI using surveillance data[8]. The method was applied to both the Spanish influenza outbreaks in USA San Francisco in 1918 and the acute respiratory illnesses (ARI) from Mexico San Luis Potosí for validation. The paper claimed to have accurate and consistent predictions. However, the model performance evaluation was based on observation, it lacked reporting of statistical measures.

Bédubourg G compared different statistical methods for early temporal detection of outbreaks from R package surveillance on a simulated data generated using a negative binomial model[9]. Among all the models, CUSUM GLM gave the best Recall at 79.5%, but with a very low precision value at 9.9%. Periodic Neg Binomial GLM gave the best precision value at 68.4%, but with a very low recall value at 20.7%. All the tested models could reach a balanced high score for both precision and recall, therefore they were either insensitive to miss out on real outbreaks or over-reacting to give a massive number of false alarms.

Ali Darwish investigated the performance of three different feature spaces in different models to forecast the weekly ILI rate in Syria using EWARS data from WHO[10]. In the same year, Yuzhou Zhang combined GFT together with WHO published data and developed a multivariate seasonal autoregressive integrated moving average model to track influenza epidemics in Australia, China, US, and UK. Both papers showed promising results[11]. However, similar to traditional surveillance conducted by the government, WHO published data could be several months of delay.

In this paper, we use a prescription dataset from over 3000 clinics in Malaysia, which can validate the geographic generalizability following 2013 Andrea's concept. A regional index (RI) is proposed based on this prescription dataset to capture the ILI trend in the regions. With RI and machine learning techniques, we propose a new data-driven methodology to detect and predict influenza outbreaks. The proposed methodology is proved to be sensitive and accurate in influenza outbreak prediction with $80-90\%$ accuracy, $70-80\%$ recall, and $70-80\%$ precision scores across five regions in Malaysia.

## Methods

### Data

In this study, a prescription dataset from over 3000 clinics in Malaysia is used. It includes weekly patient prescription and diagnosis details from 4 Jan 2016 to 21 July 2019. The pre-processing of prescription data is required to filter the relevant ILI data as the raw dataset contains data from all types of illnesses. In this study, the ILI data are identified by ICD10 codes, the international classification of diseases codes used by the WHO[12]. In order to give an early alarm of the influenza outbreaks, both confirmed cases and early symptoms are considered. Table 1 shows the ICD10 codes used to filter to ILI data from the original prescription dataset.

**Table 1.** ICD10 Code and ILI Diagnosis

| ICD10 | Diagnosis | Selected Reason |
|---|---|---|
| J09, J10, J11 | Influenza | Patients diagnosed as influenza are included in the analysis to ensure specificity. |
| R50 | Fever | According to 2017 Julia[13], ILI is defined by WHO as "An acute respiratory illness with a measured temperature of $> 38\ ^\circ$ and cough, with onset within the past 10 days", where fever and cough are the two key diagnoses. |
| R05 | Common Cough | |
| R06.7 | Sneezing | 2015 Yang[14] studied the key diagnosis associated with influenza. Fever + cough had the best sensitivity and fever + cough + sneezing had the best specificity at 77%. Therefore, sneezing is selected to increase specificity. |
| J00, J30 | Common Cold | 2016 Charles[15] points out that common cold and influenza normally share similar symptoms . Influenza patient might have been diagnosed as common cold at the beginning. |

Similar to past surveillance systems, weekly data are used in this study. We aggregate the data by the clinics and count the number of ILI prescriptions weekly. The data can be aggregated into a lower frequency, e.g., biweekly or monthly. However, we use weekly data because it provides timely results. We can detect the weekly outbreak status and predict whether the next coming week will be an outbreak. This provides outbreak alerts several weeks early than the official reports from traditional surveillance methods.

### Region Index

To study the regional influenza outbreak, we will introduce RI, a metric that normalizes the impact of the size and the number of clinics at the regional level. Using the prescription data from clinics, we have the flexibility to decide the granularity of regions by grouping the clinics based on geographic location. In this paper, we define five regions in Malaysia following the definition used by the Malaysia Federal Department of Town and Country (Table 2)[16]. Over 3000 clinics are segmented into five regions: Central, East Coast, East Malaysia, Northern, and Southern (Figure 1).

**Table 2.** Definition of the Five Regions in Malaysia[16]

| Region | States |
|---|---|
| Central | Selangor |
| East Coast | Kelantan, Pahang, Terengganu |
| East Malaysia | Sabah, Sarawak |
| Northern | Kedah, Penang, Perak, Perlis |
| Southern | Johor, Melaka, Negeri Sembilan |

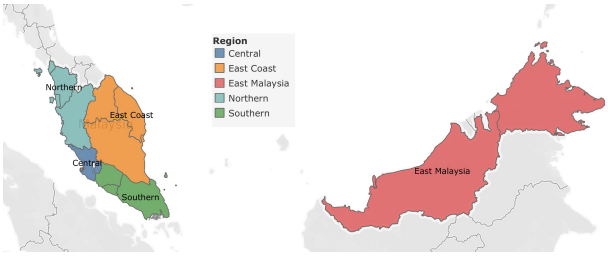



**Figure 1.** Five Regions in Malaysia

The mathematical definition of RI is shown Equation (1). Table 3 shows an example of regional RI.

$$R_{j,r} = \frac{\sum_{i=1}^{n_{j,r}} N_{i,j,r}}{\sum_{i=1}^{n_{j,r}} S_{i,j,r}}, \; j \geq 1 \tag{1}$$

where:

| | | |
|---|---|---|
| $R_{j,r}$ | is | Region Index (RI) at week $j$ of Region $r$ |
| $r$ | $\in$ | {Central, East Coast, East Malaysia, Northern, Southern} (Table 2) |
| $N_{i,j,r}$ | is | number of ILI prescriptions of clinic $i$ at week $j$ of Region $r$ |
| $n_{j,r}$ | is | number of clinics at week $j$ of Region $r$ |
| $S_{i,j,r}$ | is | average size of the clinic $i$ at week $j$ of Region $r$ as defined in Equation (2) |

$$S_{i,j} = \begin{cases} \frac{\sum_{J=1}^{J=j-1} N_{i,J}}{j-1} & \text{if } j > 1 \\ N_{i,1} & \text{if } j = 1 \end{cases} \tag{2}$$

**Table 3.** Regional Index (RI) of Prescription Data

| Region | Date | Region Index |
|---|---|---|
| Central | Week 4 - 10 Jan 2016 | 1 |
| Central | Week 11 - 17 Jan 2016 | 1.2 |
| Central | ... | ... |
| Central | Week 15 - 21 July 2019 | 1.5 |
| East Coast | Week 4 - 10 Jan 2016 | 1 |
| ... | ... | ... |
| Southern | Week 15 - 21 July 2019 | 1.6 |

## Influenza Outbreaks Detection Method

RI had normalized the original ILI cases for each week and each region. Next, we will apply anomaly detection models to label the regional outbreak on a weekly basis. From the past research papers[17–19], the 70th and 90th percentile are often used on normalized ILI cases to identify weak and strong indications of influenza outbreaks. Applying these thresholds on the data, $RI \geq 1.05$ and $RI \geq 1.2$ gives a weak and strong indication of influenza outbreaks at the 70th and 90th percentile, respectively

(Table 4). In the example illustration plot of the Southern region, the weeks in the pink range represent a strong indication of influenza outbreaks at above 90th percentile; the weeks in the light pink range represent a weak indication of influenza outbreaks between 70th and 90th percentile (Figure 2).

**Table 4.** Weekly RIs statistics Summary

| | Number of Weeks % | | | | |
|---|---|---|---|---|---|
| RI Range | Central | East Coast | East Malaysia | Northern | Southern |
| RI $< 1.05$ | 71% | 71% | 71% | 61% | 82% |
| RI in $[1.05, 1.2)$ | 23% | 19% | 21% | 26% | 13% |
| RI $\geq 1.2$ | 6% | 10% | 8% | 14% | 5% |

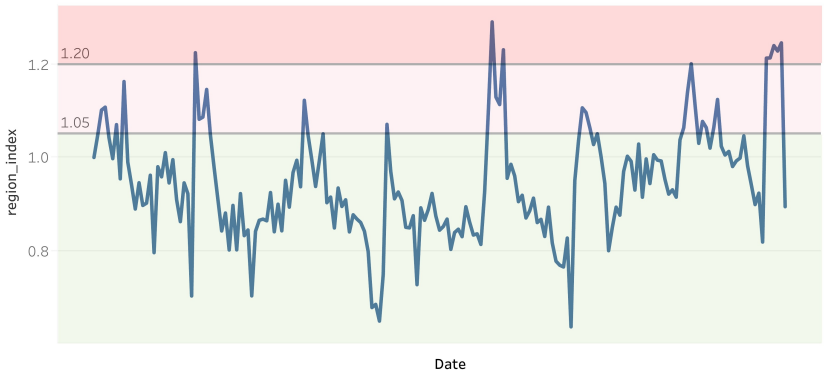



**Figure 2.** Example: Apply 70% and 90% threshold to Southern region's RIs

These thresholds are used in the anomaly detection model design. The majority (70%) of the RIs are below RI=1.05, representing a normal situation. RI=1.05 will be used as the minimum requirement for a week to be labeled as an outbreak. (Equation (10)). RI=1.2 will be applied in one of the anomaly detection models. (Equation (3)).

### *Anomaly Detection Models*

We consider influenza outbreaks as anomalies in weekly RI value. To identify the anomalies, we use two types of statistical methods, including five statistical models (Table 5). The type A method focuses on detecting the anomalies over the statistical upper bound of the dataset, and the type B method identifies the abrupt growth in time-series data. The type B method complements the type A method when the base values are low. The details of the anomaly detection models are described below. Here we use $O_{j,r,i}$ as the anomaly label at week $j$ in region $r$ using model $i$. The Southern region is used as an example, where labeled weeks are plotted in red triangles for each of the five models (Figure 3).

- Model 1: Simple Threshold
  The first model is Simple Threshold. The anomaly label using Simple Threshold is given in Equation (3). RI=1.2 is selected as the threshold to give a 90% confidence interval (Table 4). It means the labeled weeks have at least 20% more ILI cases than the historical average.

$$O_{j,r,1} = \begin{cases} 1 & \text{if } R_{j,r} \geq 1.2 \\ 0 & \text{else} \end{cases}, \; j \geq 1 \tag{3}$$

- Model 2: Z-score Model
  In Z-score model, the anomaly label is given in Equation (4). In each region, $\mu$ and $\sigma$ are the mean and standard deviation of the RIs every half-year. p=1.3 is used for 90th percentile (Equation (5)).

$$O_{j,r,2} = \begin{cases} 1 & \text{if } R_{j,r} \geq \mu + p * \sigma \\ 0 & \text{else} \end{cases}, \; j \geq 1 \tag{4}$$

**Table 5.** Five Statistics Models used for Anomaly Detection

| Method | Explanation | Model | Feature |
|---|---|---|---|
| A. Outliers over Upper Bound | Statistical models which tries to identify anomalies which have value outside of norm band | 1) Simple Threshold: $RI >= 1.2$ which is equivalent to above 90th percentile (Table 4) | Interpretable and captures all extreme High RI |
| | | 2) Z-score Model: $RI > 90\%$ Confidence Interval's upper bound | Captures higher than upper bound points using mean and standard deviation |
| | | 3) Tukey's Model: $RI > 90\%$ IQR upper bound | Captures higher than upper bound points using quantiles |
| B. Abrupt Growth | Statistical models which tries to identify anomalies which have abrupt growth | 4) Growth Value: RI growth value > median (positive weekly growth value) | Captures abrupt growth in value |
| | | 5) Growth Rate: RI growth rate > 10% | Captures abrupt growth in percentage |

$$\Pr(O_{j,r,2} = 1) = 1 - Z_{score}(p) \tag{5}$$

where:

| | | |
|---|---|---|
| $\mu$ | is | the mean of the RIs for each of the half-year (26 weeks) windows |
| $\sigma$ | is | the standard deviation of the RIs for each of the half-year (26 weeks) windows |
| $p$ | is | a constant. In this paper $p = 1.3$ is used to get 90% confidence interval by Equation (5) |

- Model 3: Tukey's Model
The anomaly label using Tukey's Model is given in Equation (6). The confidence interval is computed using quantiles of the RIs every half-year. Here, we take q=0.4 for 90% confidence interval (Equation (7)).

$$O_{j,r,3} = \begin{cases} 1 & \text{if } R_{j,r} \geq Q75 + q \times IQR \\ 0 & \text{else} \end{cases}, \; j \geq 1 \tag{6}$$

$$\Pr(O_{j,r,3} = 1) \xrightarrow{\text{normal approximation}} 1 - Z_{score}(0.6745 + q * 1.35) \tag{7}$$

where:

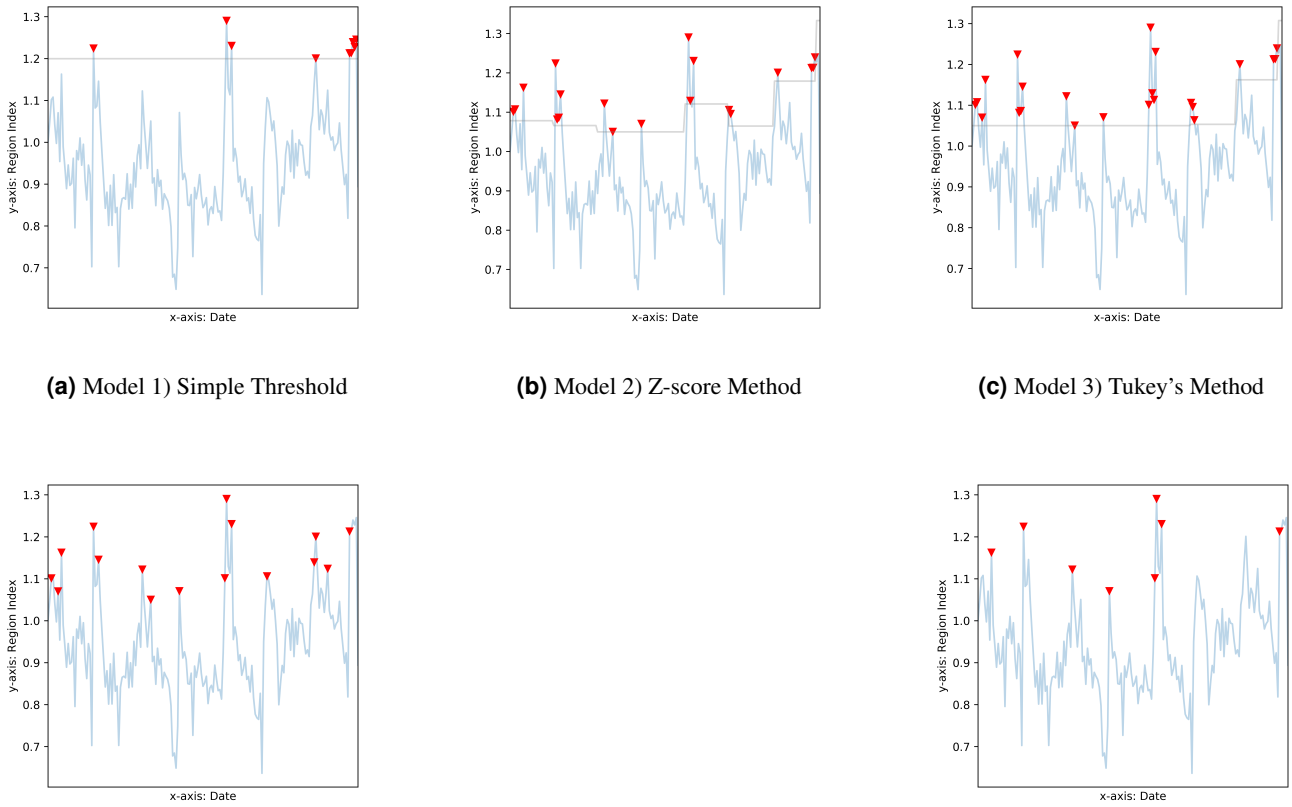
| | | |
|---|---|---|
| $Q75$ | is | 75th-Percentile of the RIs of the half-year (26 weeks) windows |
| $IQR$ | is | Inter-Quantile Range of the RIs of the half-year (26 weeks) windows |
| $q$ | is | a constant. In this paper $q = 0.4$ is used to get 90% Confidence Interval by Equation (7) |

- Model 4: Growth Value
The fourth model is using RI growth value as a measurement to identify abrupt growing RIs. The anomaly label using growth value is given in Equation (8). It means the labeled weeks have RI growth value exceeding the median of the positive growth values.

$$O_{j,r,4} = \begin{cases} 1 & \text{if } R_{j,r} - R_{j-1,r} \geq median(R_{J,r} - R_{J-1,r}) \text{ where } R_{J,r} > R_{J-1,r} \\ 0 & \text{else} \end{cases} \tag{8}$$

- Model 5: Growth Rate
The fifth model uses RI growth rate as a measurement to identify abrupt growing RIs. The anomaly label using the growth rate is given in Equation (9). It means the labeled weeks have RI growth rate exceeding 10%.

$$O_{j,r,5} = \begin{cases} 1 & \text{if } \frac{R_{j,r} - R_{j-1,r}}{R_{j-1,r}} \geq 10\% \\ 0 & \text{else} \end{cases}, \; j > 1 \tag{9}$$




**(a)** Model 1) Simple Threshold

**(b)** Model 2) Z-score Method

**(c)** Model 3) Tukey's Method

**(d)** Model 4) Growth Value

**(e)** Model 5) Growth Rate

**Figure 3.** Outlier lables from five Statistics Models

### *Outbreak Labelling*

The weekly outbreak labelling in each region is done in two steps, given by Equation (10) and (11) respectively. An illustration of this labeling process is shown in Figure 4.

In the first step (Equation (10)), as long as one of the five anomaly detection models detect the anomaly, the week will be labelled as an outbreak. Because the five models complemented each other in anomaly detection, it increases the sensitivity of influenza outbreak detection. Note that $R_{j,r}$ greater than 1.05 from $70th$ percentile is imposed, so that the weeks labeled as outbreaks were at least 5% more in ILI cases than historical average to avoid over-labeling.

In the second step (Equation (11)), a two-week outbreak window is proposed in this paper. This is because the development of an infectious disease outbreak takes some time. Based on our study of the prescription historical dataset, for any region $r$ that starts to show a strong indication of an outbreak in week $j-1$, the next week $j$ will be considered as a continuity of the previous outbreak. Following the empirical observation, this paper defines the start of an influenza outbreak to be a two-week period.

$$I_{j,r} = \begin{cases} 1 & \text{if } R_{j,r} \geq 1.05 \text{ and } \sum_{m=1}^{m=5} O_{j,r,m} \geq 1 \\ 0 & \text{else} \end{cases}, \; j \geq 1 \tag{10}$$

where:

$I_{j,r}$ is the influenza outbreak indicator at week $j$ of Region $r$,
1 means outbreak, 0 means non-outbreak

$$I_{j,r} = \begin{cases} 1 & \text{if } I_{j-1,r} = 1 \\ I_{j,r} & \text{else} \end{cases} \tag{11}$$

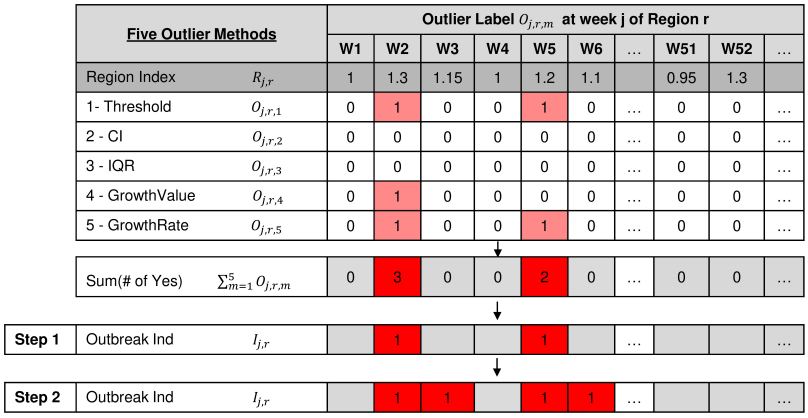



**Figure 4.** Influenza outbreak detection for historical data illustration

## Influenza Outbreaks Prediction Method

With the labeled influenza outbreak data (Table 6), we develop an ensemble machine learning method to predict future outbreaks.

**Table 6.** Prescription Data with influenza outbreaks labeled

| **Region** ($r$) | **Date** ($j$) | **RI** ($R_{j,r}$) | **Influenza outbreaks** ($I_{j,r}$) |
|---|---|---|---|
| Central | Week 4 - 10 Jan 2016 | 1 | 0 |
| Central | Week 11 - 17 Jan 2016 | 1.2 | 1 |
| Central | ... | ... | ... |
| Central | Week 15 - 21 July 2019 | 1.5 | 1 |
| East Coast | Week 4 - 10 Jan 2016 | 1 | 0 |
| ... | ... | ... | ... |
| Southern | Week 15 - 21 July 2019 | 1.6 | 1 |

### *Feature Generation - Focus on Prior Outbreak Pattern*

In supervised learning, feature X to response Y relationship needs to be constructed to train the model using historical data. And then, given a new X, the model can predict the corresponding Y. In this paper, response Y is the outbreak indicator of the next week. Feature X is constructed as the RI patterns of w weeks prior. Here w is a parameter. This feature-response construction allows the model to study patterns before an outbreak.

Assuming there are n weeks of historical data available, Table 6 can be reformatted into Table 7 for each region r. The feature set X to response Y construction is shown in Equation (12) for historical data. In each region r, X takes the past w weeks' RI, and Y is the influenza outbreak indicator. The same construction works to predict future outbreaks, shown in Equation (13). Given known $X_{n-w+1}$, i.e., the most recent $w$ weeks' RI from historical data, the classification model predicts unknown $Y_{n-w+1}$, i.e., the outbreak indicator of next week.

Note that the (X, Y) pairs are mutually independent. It has assumed the outbreak indicator of week j only depends on the RI pattern of the week j-w to week j-1. In other words, it is the week j-w to week j-1's RI pattern that decides whether week j is an outbreak. That is why parameter w needs to be carefully selected. We will discuss the use of cross-validation in Section  to select the optimal value for parameter w.

$$(X_u, Y_u) = ((R_u, R_{u+1}, ..., R_{w+u-1}), (I_{w+u})) \text{ for } 1 \leq u \leq n-w \tag{12}$$

**Table 7.** Reformat Table 6 for Each Region r

| Date Week ($j$) | $W_1$ | $W_2$ | $W_3$ | ... | $W_{n-1}$ | $W_n$ |
|---|---|---|---|---|---|---|
| RI ($R_{j,r}$) | $R_1$ | $R_2$ | $R_3$ | ... | $R_{n-1}$ | $R_n$ |
| Outbreak Indicator ($I_{j,r}$) | $I_1$ | $I_2$ | $I_3$ | ... | $I_{n-1}$ | $I_n$ |

where:

$X_u$ is the feature constructed at week $w+u$, composed of RI from $w$ weeks prior
$Y_u$ is the response at week $w+u$, which is the outbreak indicator

$$(X_{n-w+1}, Y_{n-w+1}) = (R_{n-w+1}, R_{n-w+2}, ..., R_n), (I_{n+1})) \tag{13}$$

***Model Design***

Response Y, the outbreak indicator, is a 1/0 binary variable, where 1 represents outbreak and 0 represents non-outbreak. This paper uses an ensemble model with Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB) (Figure 5). The pseudocode of the ensemble model is shown in Algorithm 1.

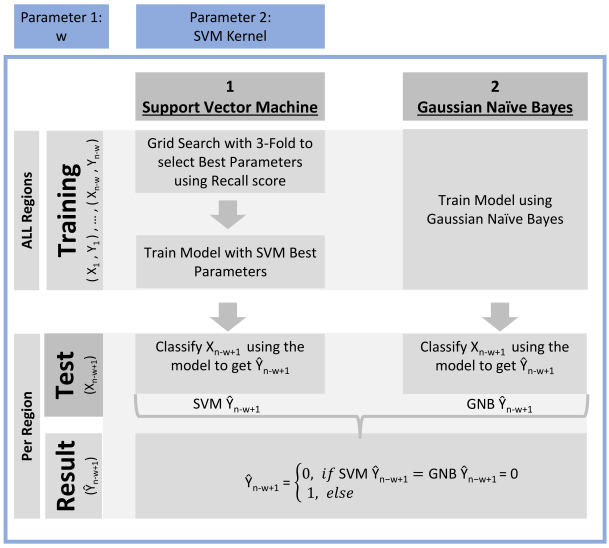



**Figure 5.** Influenza outbreak prediction model architecture

SVM is selected because it fits perfectly with the paper's problem setup. As it is widely known, SVM can be used in supervised learning, which plots each example X as points in space. It aims to find a hyperplane to separate the points by categories Y as wide as possible. With the hyperplane, new example X entering the space will fall to one side of the hyperplane, therefore being predicted to belong to a category Y. Apply the same concept to the data structure defined in Part A. For features constructed using historical data as described in Equation (12), each example X is a point in space. SVM aims to find a hyperplane to separate Y=1 outbreaks from Y=0 non-outbreaks as wide as possible. With the hyperplane, Equation (13) new X (X is the last W weeks RI) entering the space will be categorized to be either Y=1 or Y=0 (Y is the prediction for next week's outbreak indicator).

SVM might fail to separate outbreak cases from non-outbreak cases if the pattern for Y is not so distinct. This would lead to many false negative predictions by SVM and therefore missing out on capturing outbreaks correctly. To capture outbreaks as

**Algorithm 1** Ensemble Machine Learning Model for Influenza Outbreak Prediction

**Input:** $(X_u, r, Y_u, r)$ for u of $1 \leq u \leq n-w$; and $X_{n-w+1,r}$ in region r denoted as $X'_r$;
**Output:** Prediction of outbreak indicator $Y'_r$;

**for all** $r \in Regions$ **do**
  $X_{train} \leftarrow Union(X_{train}, X_r)$
Train SVM using $X_{train}$ denoted as $SVM_{trained}$
Train GNB using $X_{train}$ denoted as $GNB_{trained}$
**for all** $r \in Regions$ **do**
  $Y'_{SVM,r} \leftarrow SVM_{trained}(X'_r)$
  $Y'_{GNB,r} \leftarrow GNB_{trained}(X'_r)$
  $Y'_r \leftarrow Or(Y'_{SVM,r}, Y'_{GNB,r})$
**return** $Y'_r$ for all $r \in Regions$

much as possible, i.e., increasing model recall score, this paper complements SVM with one more classification model, the GNB.

Inspired from García YE's work in 2015[8], by assuming RI follow Gaussian distribution, GNB can be applied. It follows Bayes theorem to predict using conditional probability function. GNB is selected to capture outbreaks that SVM might miss out from a different angle.

#### *Model Evaluation and Parameter Tuning using Cross-Validation*

To determine parameter $w$ and SVM *kernel* and evaluate the stability and trustworthiness of prediction results, a cross-validation process is designed in Algorithm 2.

**Algorithm 2** Cross-Validation Process

**Input:** $(X, Y)$;
**Output:** Recall, Precision, Accuracy for Cross Validation;

**for** $i = 0$ to 100 **do**
  **for all** $r \in Regions$ **do**
    $(X_{i,r,train}, Y_{i,r,train}), (X_{i,r,test}, , Y_{i,r,test}) \leftarrow split(X_r, Y_r)$ {randomly select 70% of $(X_r, Y_r)$ to be train, the remaining 30% to be test}
    $(X_{i,train}, Y_{i,train}) \leftarrow Union((X_{i,train}, Y_{i,train}), (X_{i,r,train}, Y_{i,r,train}))$
  Train SVM using training data $(X_{i,train}, Y_{i,train})$, as $SVM_{trained}$
  Train GNB using training data $(X_{i,train}, Y_{i,train})$, as $GNB_{trained}$
  **for all** $r \in Regions$ **do**
    $Y'_{i,r,SVM} \leftarrow SVM_{trained}(X_{i,r,test})$
    $Y'_{i,r,GNB} \leftarrow GNB_{trained}(X_{i,r,test})$
    $Y'_{i,r} \leftarrow Or(Y'_{i,SVM,r}, Y'_{i,GNB,r})$
    $recall_{i,r} \leftarrow Recall(Y_{i,r,test}, Y'_{i,r})$
    $precision_{i,r} \leftarrow Precision(Y_{i,r,test}, Y'_{i,r})$
    $accuracy_{i,r} \leftarrow Accuracy(Y_{i,r,test}, Y'_{i,r})$
$\mu(recall_r) \leftarrow mean(recall_{i,r})$
$\sigma(recall_r) \leftarrow stdev(recall_{i,r})$
$\mu(precision_r) \leftarrow mean(precision_{i,r})$
$\sigma(precision_r) \leftarrow stdev(precision_{i,r})$
$\mu(accuracy_r) \leftarrow mean(accuracy_{i,r})$
$\sigma(accuracy_r) \leftarrow stdev(accuracy_{i,r})$
**return** $\mu(recall_r), \sigma(recall_r), \mu(precision_r), \sigma(precision_r), \mu(accuracy_r), \sigma(accuracy_r)$

Three evaluation metrics are selected to be reported for model performance evaluation in this paper.

1. Recall as the primary metrics aiming to find all real outbreaks
   Recall measures how sensitive the model is in reporting actual outbreaks, i.e., how many real outbreaks are being predicted correctly by the model. The main purpose of the project is to detect potential outbreaks early and not miss any actual outbreaks. The historical data is imbalanced in outbreak indicator labeling, and there are much more 0s

(non-outbreak weeks) than 1s (outbreak weeks). Therefore, the recall score is the optimal option to evaluate how sensitive the model is in identifying real outbreaks. It is used in the cross-validation process for parameter tuning as well to increase model sensitivity.

2. Precision to ensure predicted outbreaks are real outbreaks
   Precision measures how precise the model is in reporting outbreaks, i.e., for all the weeks predicted as an outbreak by the model, how many are real outbreaks. It is reported together with Recall during cross-validation to avoid over-labeling of the outbreaks. It is also reported as the confidence level of the prediction result, indicating the probability of the predicted outbreak being a real outbreak.

3. Accuracy for reference
   Accuracy is the most intuitive performance measure. Due to the nature of imbalanced data, accuracy is quite high in general. Therefore, it is reported just for reference.

## Results

### Result for Influenza Outbreaks Detection Method

We validate the outbreak labeling results by comparing the labels against GFT. GFT data is selected for comparison due to three reasons. Firstly, earlier research had proven GFT's usage in outbreak detection[6]. Secondly, government official reports are usually only for large-scale influenza outbreaks, for example, seasonal influenza outbreaks or any pandemic caused by new or unknown viruses. Besides pandemic, our paper also detects small-scale outbreaks at the early stage, which would not be reported by the government. GFT, which uses search results, could serve a similar purpose as our paper. Thirdly, GFT can specify search terms and geographic granularity to align with our model. We can use ILI relevant search terms and choose the same cities from GFT to provide the closest comparison with our model result.

Figure 6 shows an example of the comparison using the Southern region. As defined in Table 2, the region is composed of three cities: Johor, Melaka, and Negeri Sembilan. The upper graph shows the Southern region's influenza outbreak labeled using the approach proposed in this section (red color indicates outbreaks). The lower graph shows the GFT search index of ILI relevant terms for the same region. From this comparison, GFT shows a similar outbreak period as our model. Our detection model is further supported by GFT to give good outbreak labeling.

From Figure 6, we highlighted three localized outbreaks where GFT shows spikes in the search index. We compared the time our paper gives the signal with that of GFT for each outbreak, as summarized in Table 8. From the comparison, our method could detect the same outbreak in the same week or earlier than GFT.

**Table 8.** First Spike Date Comparison of This Paper's Outbreak Labelling vs. GFT for the Southern Region

| Outbreak ID | Our Paper | GFT | Conclusion |
|---|---|---|---|
| 1 | Week 30 Jan 2017 | Week 12 March 2017 | Our paper detects earlier than GFT |
| 2 | Week 15 Jan 2018 | Week 14 Jan 2018 | Our paper detects at the same week as GFT |
| 3 | Week 10 Jun 2019 | Week 7 July 2019 | Our paper detects earlier than GFT |

### Result for Influenza Outbreaks Prediction Method

In this study, using the historical data from 4 Jan 2016 to 21 Jul 2019, we may predict whether next week 22 to 28 Jul 2019 is an outbreak for each of the five regions in Malaysia (Table 2).

Firstly, we use Algorithm 2 to select the best parameter. The cross-validation result of recall score is shown in Table 9. SVM $kernel = rbf$ and $w = 3$ are selected as the best parameter, as it gives the highest recall score. Table 10 reports the precision and accuracy scores on top of the recall scores for the selected best parameter. For all regions, there is high recall and precision scores and low standard deviation. The model is proved to be sensitive, accurate, and robust in predicting influenza outbreaks.

Next, we apply these selected best parameters to Algorithm 1. The model predicts that East Malaysia will not be in an influenza outbreak next week, and all rest of the regions will be in an influenza outbreak next week (Table 11).

The performance of the proposed influenza outbreaks prediction model is summarized below.

- from cross-validation, the model is proven to be reliable and stable, with 70 to 80% of recall, 70 to 80% of precision and 80 to 90% of accuracy scores for most regions, and small variation in each runs (Table 10).

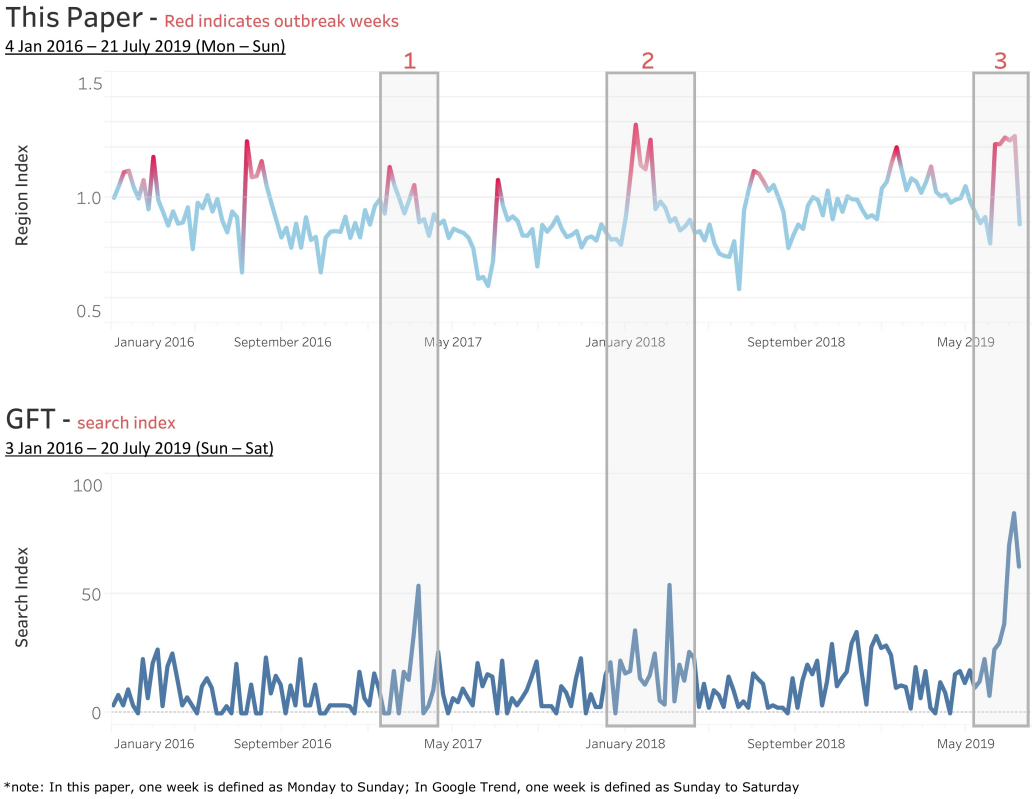



**Figure 6.** Outbreak labelling result vs GFT for the Southern region

- the model is sensitive to capture the real outbreaks, where 68 to 77% of the real outbreaks can be predicted as outbreak correctly by the model across the five regions (Table 10. Recall).
- the model predicted that for next week 22 to 28 Jul 2019, East Malaysia would not be in an influenza outbreak. There is 60% of probability East Coast will be an outbreak; 70% of probability Northern will be an outbreak; 75% of probability Southern will be an outbreak; 76% of probability Central will be an outbreak. (Table 10, Table 11).

**Table 9.** Cross-Validation Result on Historical data 4 Jan 2016 to 21 Jul 2019

| Parameters | | Recall score: mean(standard deviation) | | | | |
|---|---|---|---|---|---|---|
| Kernel | w | Central | East Coast | East Malaysia | Northern | Southern |
| poly | 2 | 0.59 (0.10) | 0.71 (0.10) | 0.70 (0.11) | 0.67 (0.08) | 0.66 (0.12) |
| poly | 3 | 0.61 (0.11) | 0.75 (0.11) | 0.75 (0.09) | 0.68 (0.08) | 0.67 (0.12) |
| poly | 4 | 0.62 (0.11) | 0.76 (0.09) | 0.75 (0.10) | 0.66 (0.09) | 0.67 (0.11) |
| poly | 5 | 0.59 (0.11) | 0.71 (0.10) | 0.73 (0.12) | 0.62 (0.09) | 0.66 (0.12) |
| poly | 6 | 0.58 (0.12) | 0.68 (0.11) | 0.74 (0.11) | 0.61 (0.09) | 0.65 (0.12) |
| rbf | 2 | 0.69 (0.10) | 0.76 (0.09) | 0.73 (0.10) | 0.73 (0.09) | 0.70 (0.13) |
| rbf | 3 | 0.68 (0.12) | 0.75 (0.11) | 0.77 (0.09) | 0.73 (0.08) | 0.71 (0.13) |
| rbf | 4 | 0.69 (0.09) | 0.76 (0.09) | 0.73 (0.10) | 0.71 (0.09) | 0.69 (0.12) |
| rbf | 5 | 0.69 (0.10) | 0.75 (0.10) | 0.76 (0.10) | 0.71 (0.09) | 0.67 (0.13) |
| rbf | 6 | 0.68 (0.11) | 0.74 (0.11) | 0.75 (0.10) | 0.69 (0.09) | 0.66 (0.12) |

To further validate the prediction model performance, we also compared the prediction result with real-world evidence.

The Strait Times reported in July that ILI cases soar in Negri Sembilan[20], which is a city in the Southern region (Table 2). It is aligned with the model prediction result, as specified in Table 11.

**Table 10.** Cross-Validation Result on Historical Data 4 Jan 2016 to 21 Jul 2019 with Optimal Parameter: Kernel = rbf, w=3

| | Central | East Coast | East Malaysia | Northern | Southern |
|---|---|---|---|---|---|
| Recall | 0.68 (0.12) | 0.75 (0.11) | 0.77 (0.09) | 0.73 (0.08) | 0.71 (0.13) |
| Precision | 0.76 (0.10) | 0.60 (0.10) | 0.75 (0.09) | 0.70 (0.09) | 0.75 (0.14) |
| Accuracy | 0.83 (0.05) | 0.79 (0.05) | 0.85 (0.04) | 0.77 (0.05) | 0.89 (0.03) |

**Table 11.** Predicted Influenza Outbreaks for Next Week 22 to 28 Jul 2019 with Optimal Parameter: Kernel = rbf, w=3
1 represents outbreak, 0 represents non-outbreak

| | Central | East Coast | East Malaysia | Northern | Southern |
|---|---|---|---|---|---|
| Predicted | 1 | 1 | 0 | 1 | 1 |

From Google Trend data (Figure 7)[21], in week 22 to 28 Jul 2019, we can see that for Malaysia overall, there is a distinct spike in search of 'Influenza', which is consistent with the prediction result from our model (Table 11). And more specifically at the region level,

- Northern, Southern, and Central regions have the top GFT Index indicating high influenza searches, which is consistent with the model result that these regions will have an outbreak with >= 70% probability (Table 10, Table 11).
- East coast has a lower GFT index indicating lower influenza searches, which is consistent with the model result that it will have an outbreak with a lower probability of 60% (Table 10, Table 11).
- East Malaysia has the lowest GFT index with the least likelihood of an outbreak, which is consistent with the model result showing 0, no outbreak (Table 11).

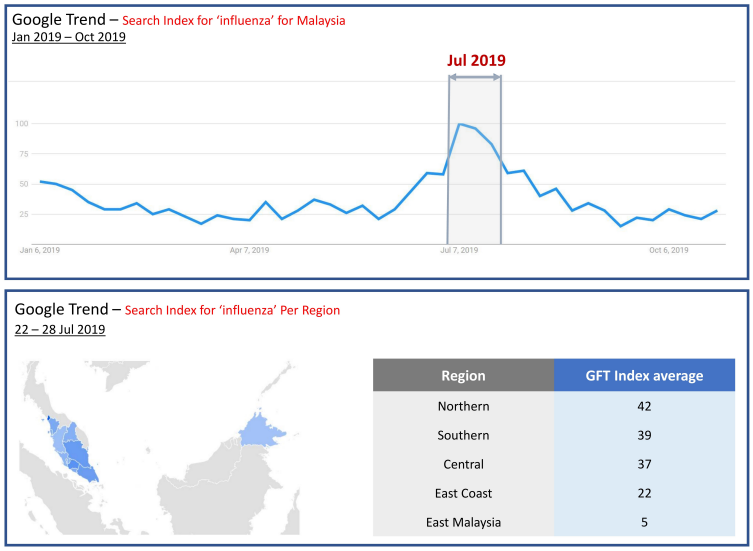



**Figure 7.** Google Flu Trend Result (week 22 – 28 Jul 2019)

## Discussion

This paper proposes a data-driven methodology using prescription dataset from over 3000 clinics to effectively detect and predict influenza outbreaks. In a typical outbreak life-cycle as plotted in Figure 8, there are three types of datasets that can be used for the influenza outbreak study: Google trend dataset, prescription dataset and official reports.

Prescription data stands out because of the following reasons. First, it monitors ILI cases based on licensed doctor diagnoses, which is usually more reliable compared to Google Trend search-based data. Secondly, prescription data provide earlier

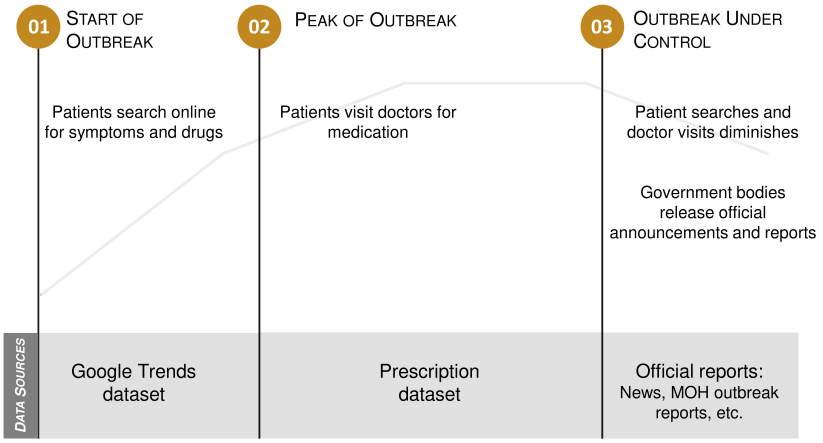



**Figure 8.** A typical outbreak lifecycle

detection insights of influenza outbreaks compared to the official reports. Andrea validated that prescription data from one medical center works well for the influenza case forecast[7]. In this study, we use a prescription data from over 3000 clinics, covering half of clinics in Malaysia[22]. We can extend Andrea's work to address geographic generalizability.

A new measurement metrics of influenza outbreaks, i.e., RI, is proposed in this paper. RI compares the number of prescriptions at the current week with the historical weekly average in the region. Unlike previous papers[11,17,19], which simply use the total number of prescriptions from all hospitals or clinics, RI uses the average size of individual clinics and handles cases when new clinics are added into or removed from the dataset during the sample period. In statistics terms, RI eliminates the biases of different clinic sizes and a varying number of clinics. It gives a good indication of whether the current week shows an anomaly of ILI prescription in the region.

We design an influenza outbreak detection method based on RI using statistical outlier detection models and validate the method with GFT results. In the real-world scenario, the sensitivity of the outbreak detection model is crucial because we try to detect as many as possible outbreaks. To improve the outbreak detection methods' sensitivity, we introduce five complemented statistical models in this paper. These models are used to label the weekly regional outbreaks to train the prediction model.

The paper emphasizes the study of RI patterns before an outbreak and develops a machine learning model to predict future outbreaks. There are usually two types of methodologies when detecting outbreaks, i.e., regression models and classification models. Regression models focus on seasonal or periodical outbreaks and better fit long-term predictions. In contrast, classification models capture dynamic patterns and better fit short-term predictions. In this paper, we decide to use classification models because Malaysia, located in South East Asia, does not have distinctive seasons; therefore, there are no clear seasonal trends for ILI cases[4]. Moreover, we can give early alerts using classification models by learning the patterns right before the outbreaks.

Cross-Validation of the model is applied to evaluate the performance. With $80-90\%$ accuracy, $70-80\%$ recall, and $70-80\%$ precision scores, the methodology is proved to be sensitive and accurate in predicting influenza outbreaks. Compared to previous similar research works, the proposed methodology is more reliable, effective, and scalable in influenza outbreak detection and prediction.

In the future, the proposed methodology introduced in this paper can be easily adapted to other diseases that prescription data covers, such as hand foot mouth disease, dengue, and COVID-19, etc. Moreover, the methodology, currently predicting at the region level, can be extended easily to the city level, or any granularity by grouping clinics based on geographic location. Besides the short-term prediction results presented in this paper, the methodology can be extended for long-term outbreak prediction by restructuring the data and incorporate other machine learning models. In addition, we can apply more complex nonlinear models like LSTM or other Neural Networks to the data set to see if we can have better prediction accuracy.

## References

1. World Health Organization. 2020. "Ten Health Issues WHO Will Tackle This Year" [online] Available at:<https://www.who.int/news-room/feature-stories/ten-threats-to-global-health-in-2019>
2. World Health Organization. 2020. "Influenza (Seasonal)". [online] Available at: <https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)>

3. Ong MP, et al. 2010. "High direct healthcare costs of patients hospitalized with pandemic (H1N1) 2009 influenza in Malaysia" [online] Available at:<https://www.researchgate.net/publication/45648809_High_direct_healthcare_costs_of_patients_hospitalised_with_pandemic_H1N1_2009_influenza_in_Malaysia>
4. Selvanesan Sengol, et al. 2018. "Malaysia Influenza Surveillance" Protocol [online] Available at: <https://www.researchgate.net/publication/329023936_MALAYSIA_INFLUENZA_SURVEILLANCE_PROTOCOL>
5. World Health Organization. 2020. "Influenza update" [online] Available at: <https://www.who.int/influenza/surveillance_monitoring/updates/latest_update_GIP_surveillance/en/>
6. Ginsberg J, et. al. 2009. "Detecting influenza epidemics using search engine query data". Nature 457, 1012–1014. <https://static.googleusercontent.com/media/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf>
7. Andrea Freyer Dugas, et. al. 2013. "Influenza Forecasting with Google Flu Trends" [online] Available at: <https://www.researchgate.net/publication/235778448_Influenza_Forecasting_with_Google_Flu_Trends>
8. García YE, Christen JA, Capistrán MA. "A Bayesian outbreak detection method for Influenza-Like Illness". BioMed Research International;2015. [online] Available at: <https://www.hindawi.com/journals/bmri/2015/751738/>
9. Bédubourg G, Le Strat Y, 2017. "Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study". PLoS ONE 12(7): e0181227. <https://doi.org/10.1371/journal.pone.0181227>
10. Ali Darwish, et. al. 2020. "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" [online] Available at: <https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-020-4889-5>
11. Yuzhou Zhang, et. al. 2019. "Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data" [online] Available at: <https://www.nature.com/articles/s41598-019-39871-2>
12. World Health Organization. 2016. "ICD10 code for Diagnosis" [online] Available at: <https://icd.who.int/browse10/2016/en#>
13. Julia Fitzner, et. al. 2017. "Revision of clinical case definitions: influenza-like illness and severe acute respiratory infection" [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5791775/>
14. Yang JH, et.al. 2015. "Predictive Symptoms and Signs of Laboratory-confirmed Influenza: A prospective Surveillance Study of Two Metropolitan Areas in Taiwan" [online] Available at: <https://www.ncbi.nlm.nih.gov/pubmed/26554802>
15. Charles Patrick Davis, 2016. "Cold vs. Flu" [online] Available at: <https://www.medicinenet.com/cold_vs_flu/article.htm#cold_vs_flu_facts>
16. MLIT, 2015. "An Overview of Spatial Policy in Asian and European Countries - Malaysia" [online] Available at: <https://www.mlit.go.jp/kokudokeikaku/international/spw/general/malaysia/index_e.html>
17. Rachael Pung, Vernon Jian Ming Lee. 2019. "Implementing the World Health Organization Pandemic Influenza Severity Assessment framework—Singapore's experience" [online] Available at: <https://onlinelibrary.wiley.com/doi/full/10.1111/irv.12680>
18. Basma AbdElGawad, et. al. 2020. "Evaluating tools to define influenza baseline and threshold values using surveillance data, Egypt, season 2016/17" [online] Available at: <https://www.sciencedirect.com/science/article/pii/S1876034119301728>
19. Pi Guo, et. al. 2017. "Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model" [online] Available at: <https://www.nature.com/articles/srep46469.pdf>
20. Raja Noraina Rja Rahim, 2019. "Straits Times ILI cases soar in Negri Sembilan" [online] Available at: <https://www.nst.com.my/news/nation/2019/07/503434/ili-cases-soar-negri-sembilan>
21. "Google Trend data – search term 'influenza"' of [online] Available at: <https://trends.google.com/trends/explore?q=influenza>

**22.** Makmor Tumin, et. al. 2018. "Demographic and socioeconomic factors associated with access to public clinics" [online] Available at: <https://www.researchgate.net/publication/327249275_Demographic_and_socioeconomic_factors_associated_with_access_to_public_clinics>

## Acknowledgements

We thank Zuellig Pharma Analytics for providing the prescription data for this research work. We thank Tristan Tan, Vice President of Zuellig Pharma Analytics for the topic.

## Author contributions statement

L.D. and P.Y. conceived and designed the study. L.D. prepared and analyzed the data, built and validated the models, wrote the manuscript. Both authors reviewed and approved the final manuscript.

## Additional information

**Competing Interests:** The authors declare no competing financial interests.
**Correspondence** and requests for materials should be address to P.Y.

# Figures

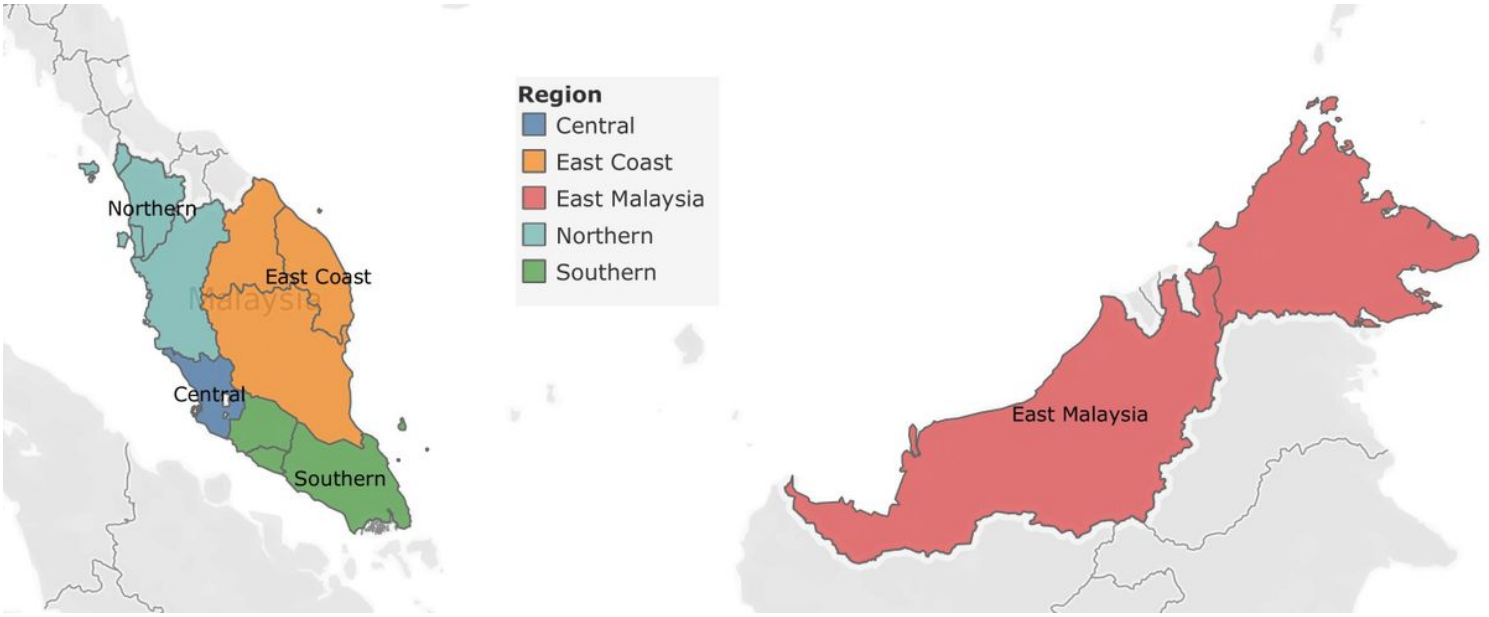



**Figure 1**

Five Regions in Malaysia Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

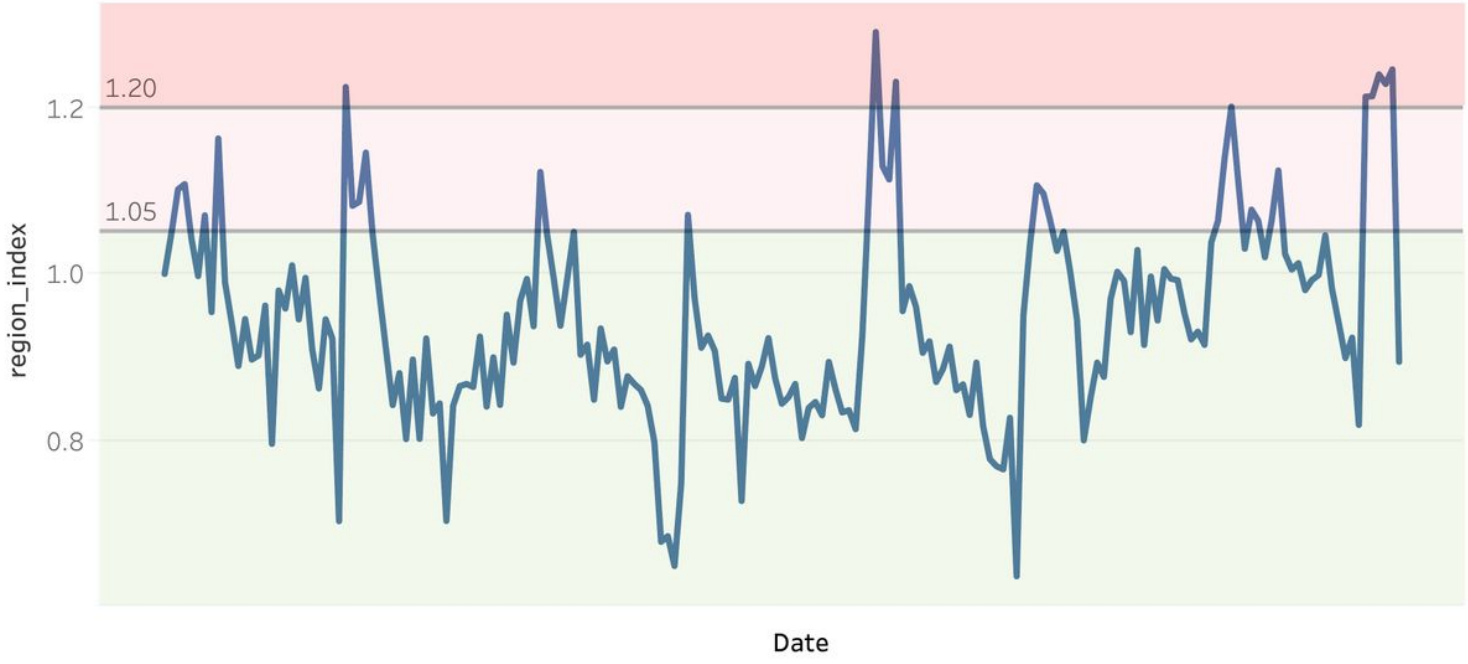
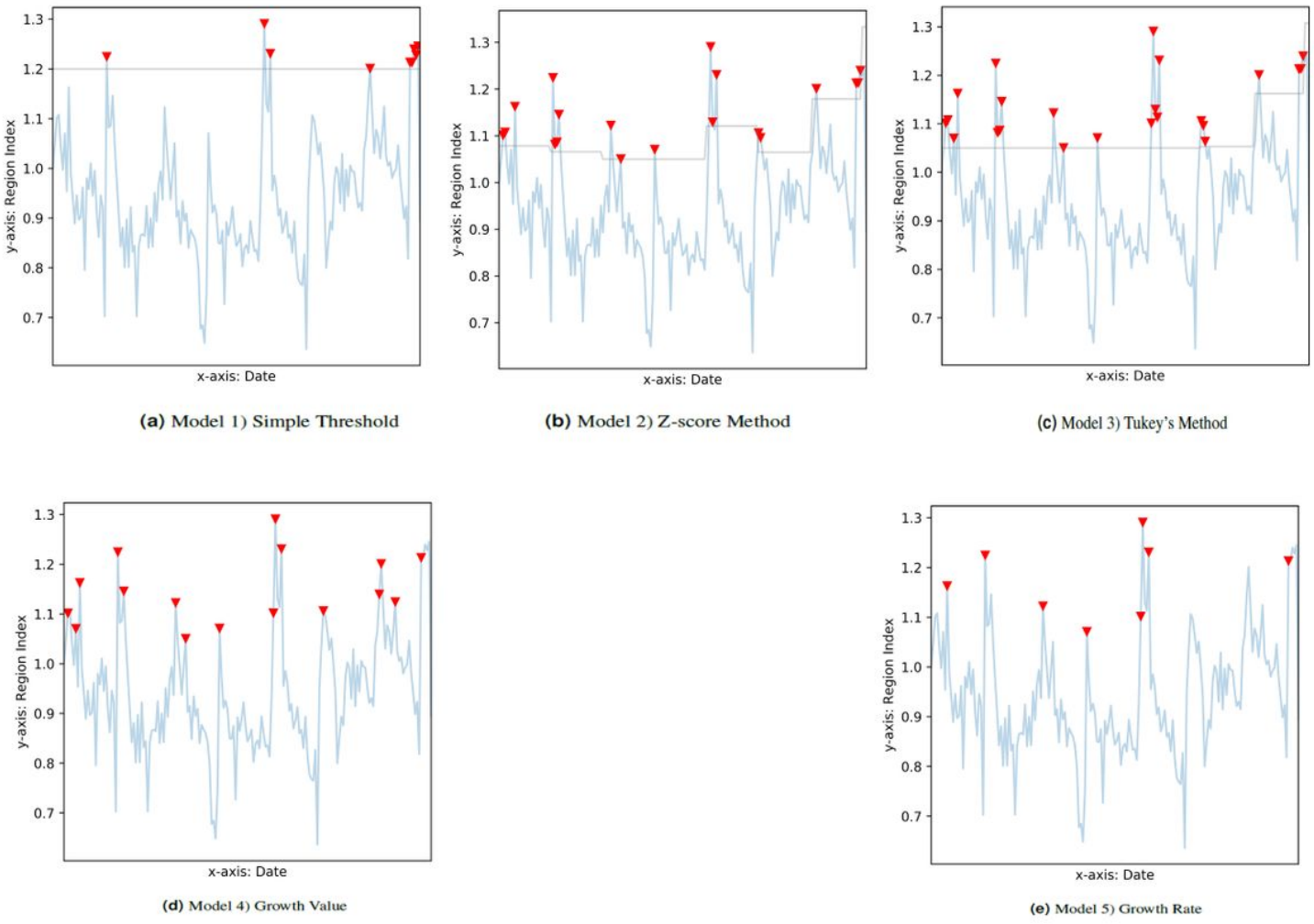


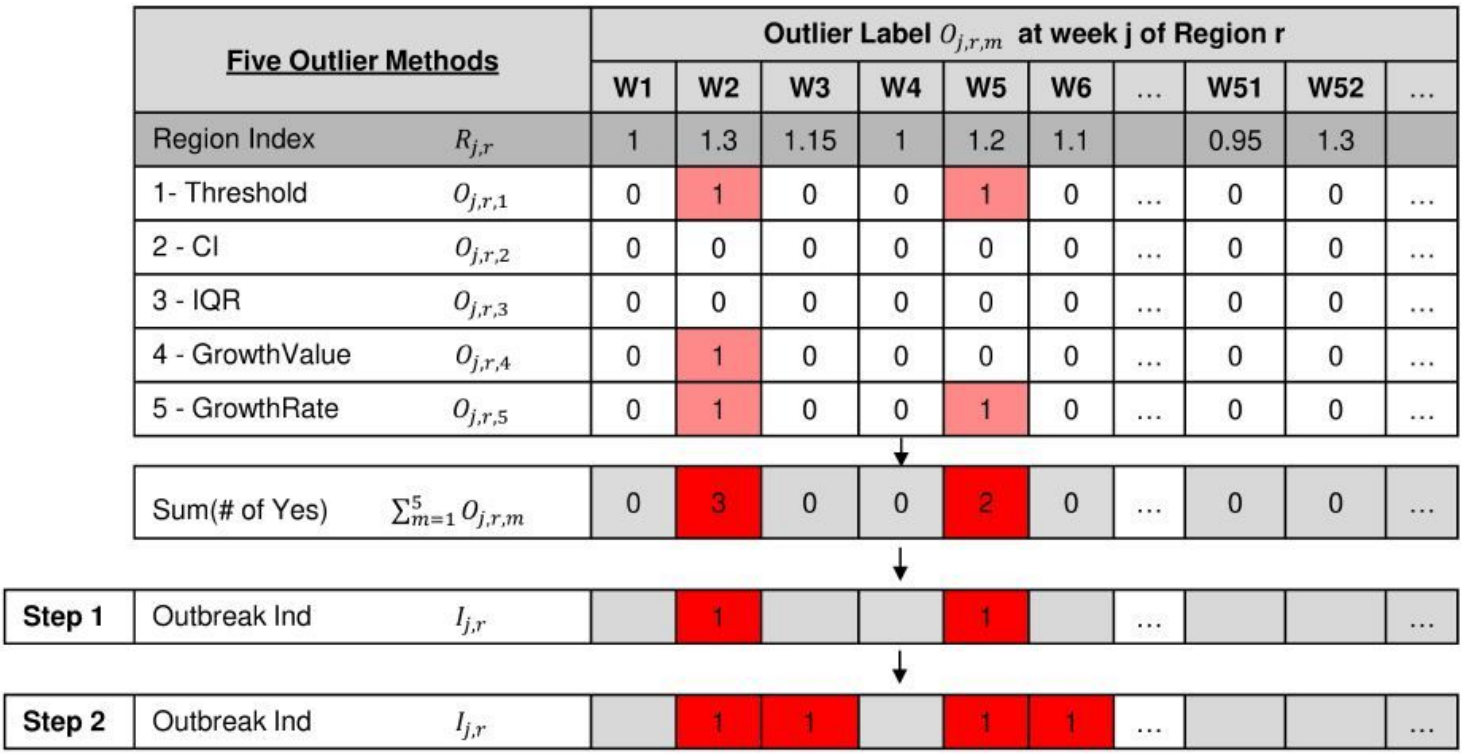

**Figure 2**

Example: Apply 70% and 90% threshold to Southern region's RIs




## Figure 3

Outlier lables from five Statistics Models




## Figure 4

Influenza outbreak detection for historical data illustration

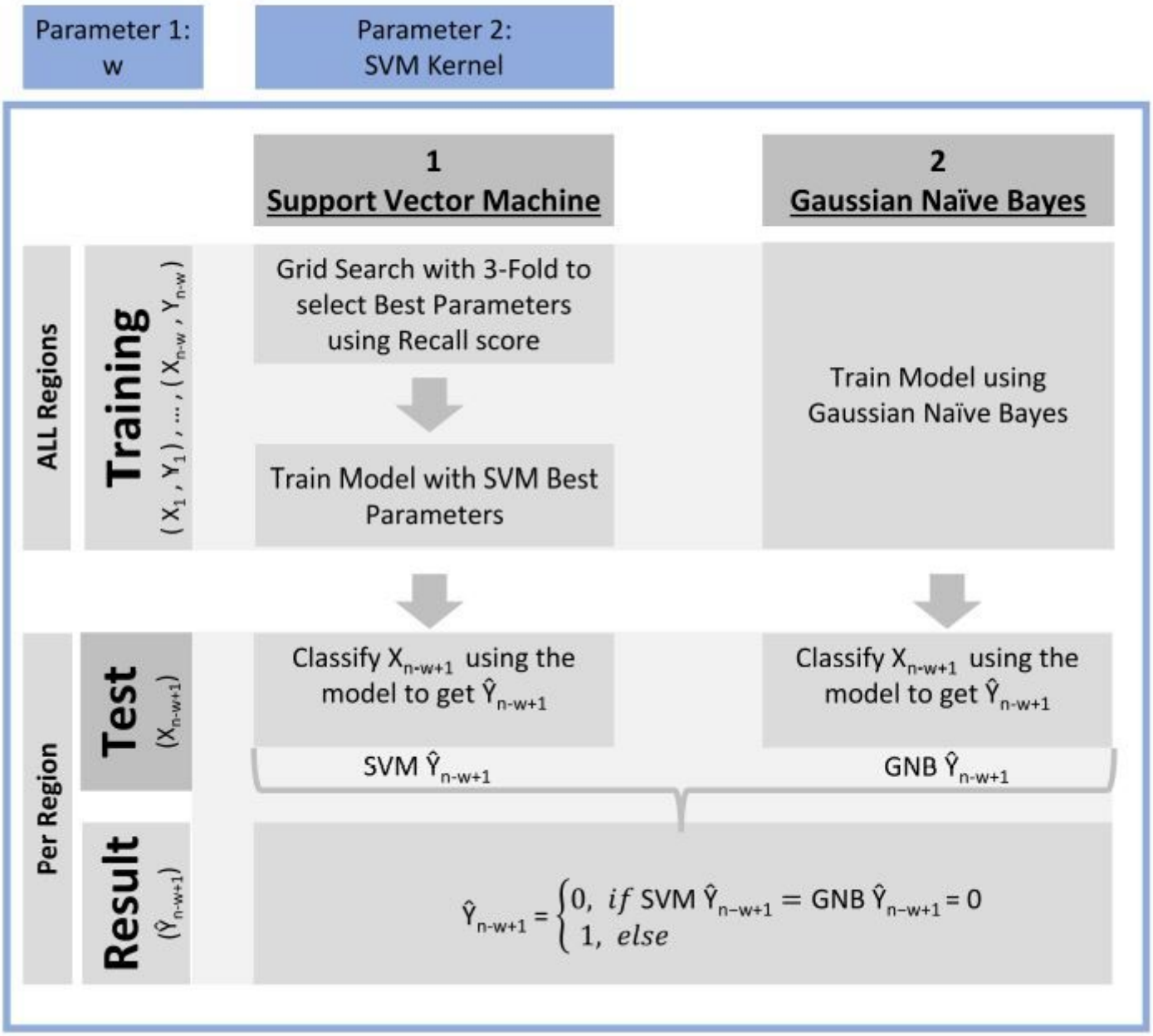



**Figure 5**

Influenza outbreak prediction model architecture

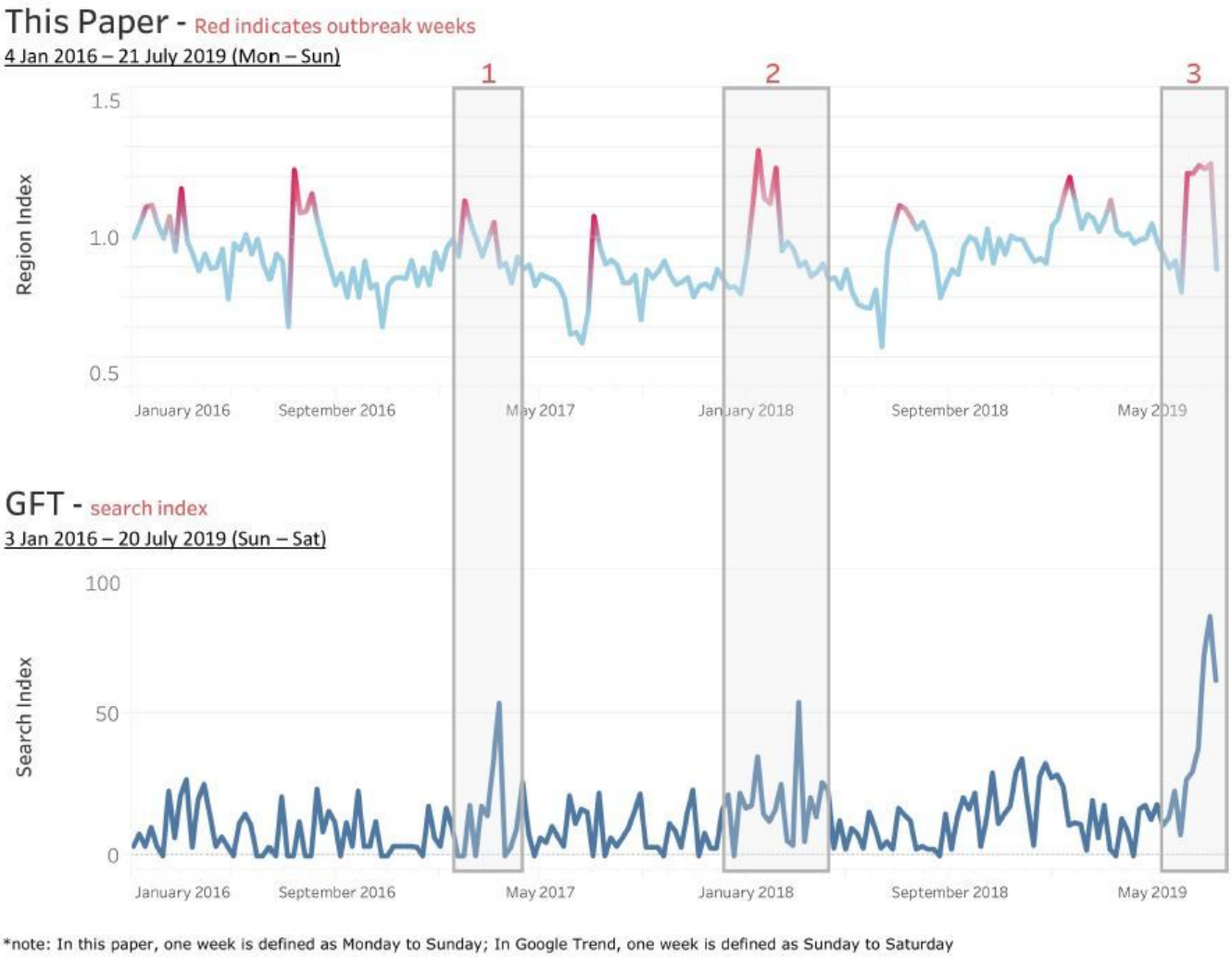



**Figure 6**

Outbreak labelling result vs GFT for the Southern region

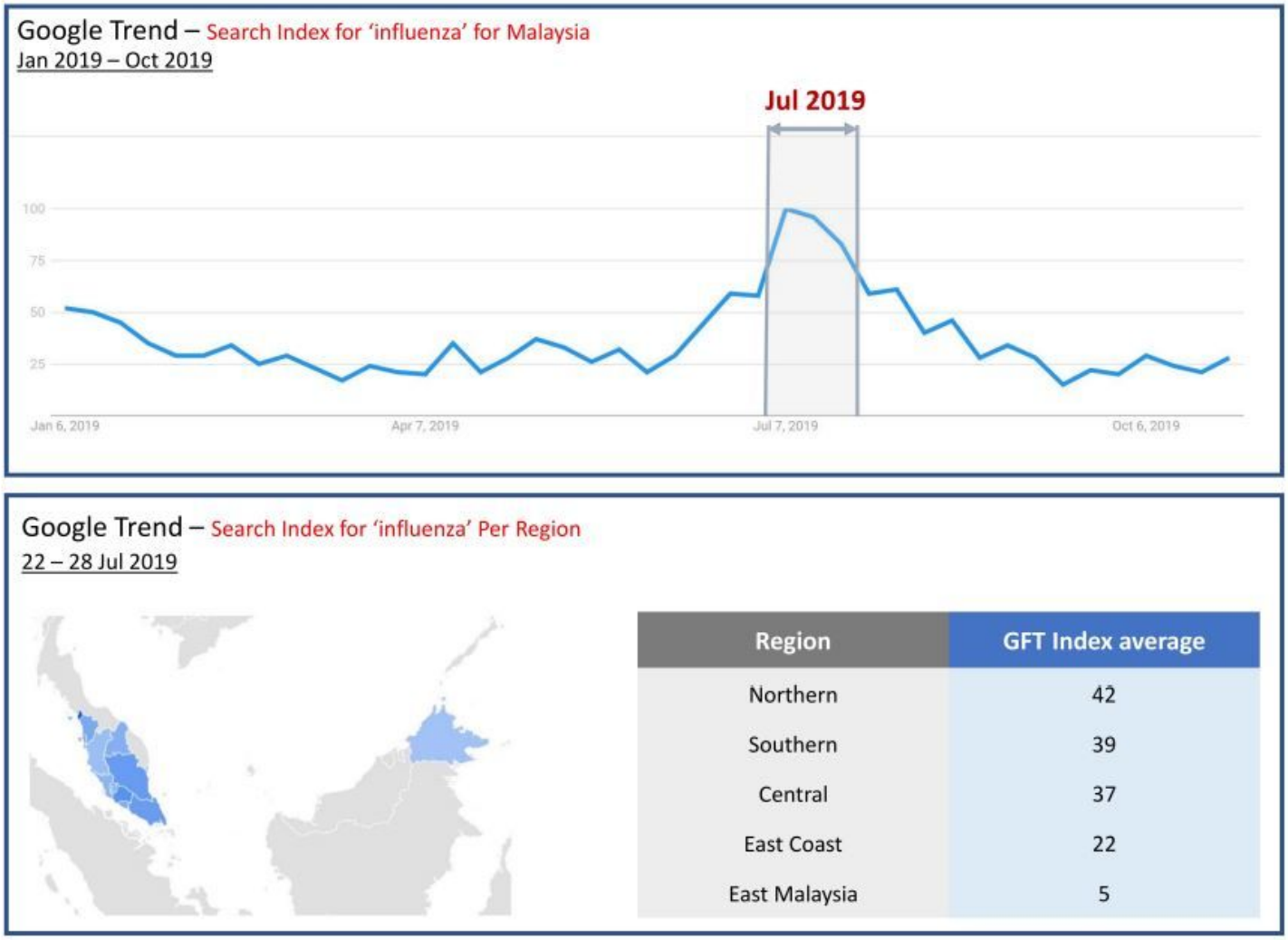



**Figure 7**

Google Flu Trend Result (week 22 – 28 Jul 2019) Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

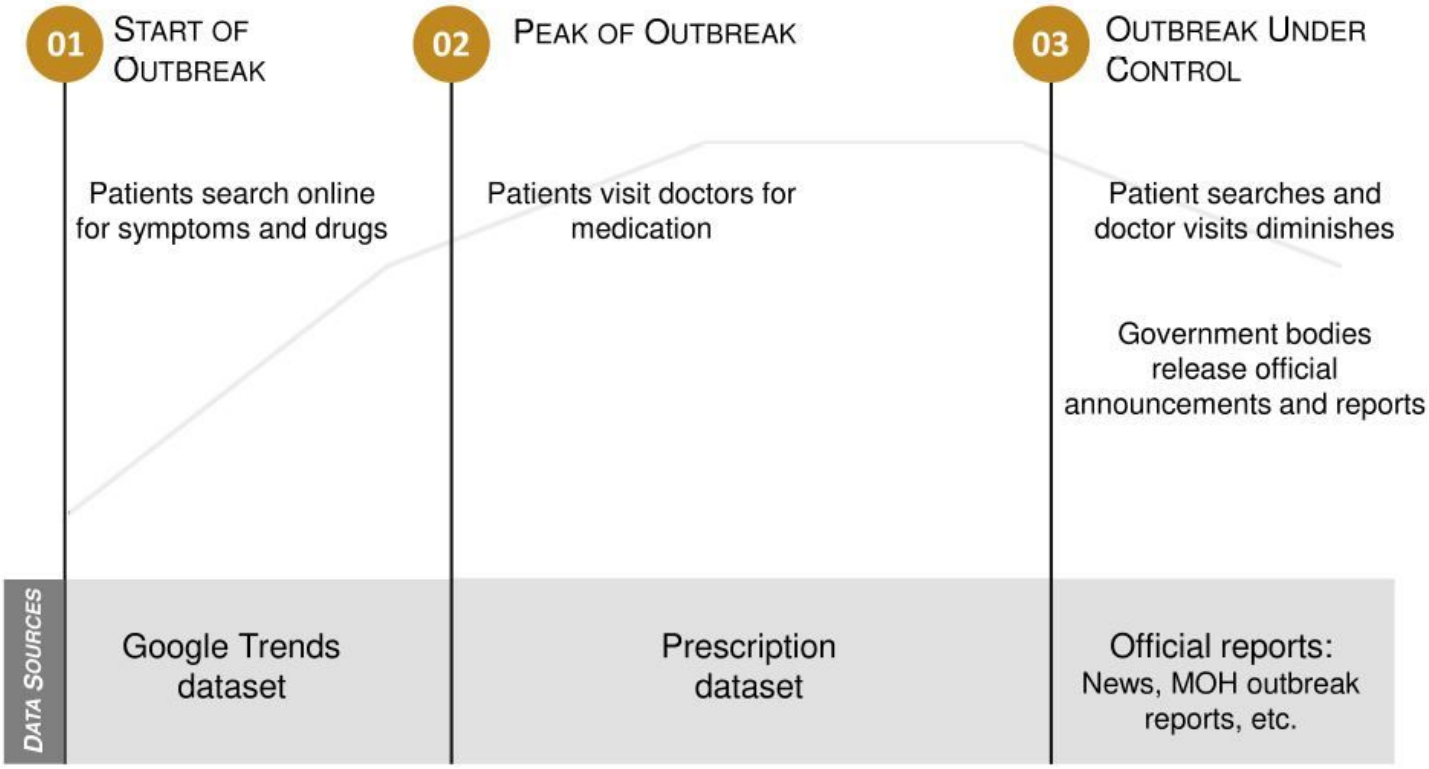



**Figure 8**

A typical outbreak lifecycle