Exploring Machine Learning Utilization using Real-Life Dataset for Influenza Pandemic

Shahid Hussain
Department of Computer Science
Mohammad Ali Jinnah University, Karachi

Ubaida Fatima, PhD
Department of Mathematics
NED University of Engineering and Technology, Karachi

ABSTRACT

There must be an exact system for monitoring the influenza outbreaks to have an optimum solution for the recovery of infected people's health. For reducing the spread of future outbreaks of influenza virus, forecasting plays an important role. Influenza a is type of disease which is transferred to human beings through pigs, found in animals. It became pandemic in Spain, approximately, 1/3rd of human population died and 1/4th of pig population. Again in 2009, influenza "A" caused millions of deaths, and spread like a pandemic rapidly. Variety of researches inspected data obtained from World Health Organization and local hospitals at country level. This research work is based on mathematical biology using data science techniques in the domain of machine learning. This research suggests a modeling scheme for influenza pandemic predictions, its different classifications and types such as H1N1, B-Victoria etc. via machine learning prediction and regression as well as classification algorithms such as Logistic Regression (LR), Support Vector Machines (SVM) using Linear, Polynomial and RBF kernels; Naïve Bayes (NB) and Random Forest (RF) method for the prediction of influenza disease and its outbreak, the influenza kind became pandemic with the infected populated area. After using various kernels in SVM algorithm, it is observed that Polynomial and Linear kernels have approximately the same accuracy scores, while RBF kernel was not best-fitted for the considered influenza datasets. As far as the overall performance is concerned, at average, RF has the highest accuracy score as 74% while the LR had also the better average score as 72% after RF. After applying the considered ML algorithms, Random Forest algorithm performed in well-effective manner and comparatively it was analyzed as the best-fitted algorithm for the considered datasets.

Keywords

Influenza, pandemic, Forecasting model, H1N1, Influenza, Data science, Biology, Logistic Regression, SVM, Linear SVM, Polynomial SVM, RBF kernel, Naïve Bayes, Random Forest.

1.1. INTRODUCTION

In the domain of scientific literature, this research work embarks on a comprehensive exploration of the influenza pandemic through the lens of machine learning. The study delves into the application of advanced computational techniques to analyze real-life datasets, offering a nuanced understanding of the intricate patterns and dynamics within the context of the influenza outbreak [1]. Through this research, valuable visions have been unraveled that contribute to the evolving landscape of public health and data-driven decision-making through findings. Influenza pandemic has been an important factor to be diagnosed at the global level. Different studies made on this pandemic using different methods. Machine learning methods have been considered the great source to forecast the future outbreaks and the types of influenza are to be discussed along with their mathematical and machine learning algorithms. After 2009, influenza caused millions of deaths [2]. Scientists and doctors worked on it to diagnose it carefully to find out the optimal solution.

1.2.Machine Learning Methods for Influenza Prediction

In Machine Learning, various methods are used for the analysis, prediction and interpretation of different parameters of Influenza pandemic such as number of the cases, types, outbreaks as well as the binomial and multinomial data related to the parameters and symptoms. Each method has different domains required to get the required optimum result depending on the type of data [3]. ***Logistic Regression*** in the analysis of presence and absence of influenza symptoms such as *vaccinated vs not-vaccinated, male vs female, wash-hands vs non-wash-hands,* is another version of multiple linear regression analysis with extension of categorical outcome variable [4]. For fitting the data in well-effective manner, it is expected that the predictors are uncorrelated with one another and are meaningfully related to the answer. There is also uncorrelation in between the observations or data elements of model [5]. ***Support Vector Machine*** is an algorithm used to classify and predict regression monitored by machine learning theory to have better and improved predictive accuracy [6]. These are the systems which utilize the linear functions space in a higher feature space. For a certain dataset, it provides a better sketch of studying the problem of gaining knowledge, making forecasting or predictions as well as the decision making operations [7]. ***Naïve Bayes algorithm*** is a type of Supervised learning algorithm and it works on the concept of Bayes Theorem contained in the statistical theory. It is usually considered when there exists a dataset based on classification. It works for n-dimensional influenza training dataset and mostly used in text classification. In the prediction and classification of medicines specially pandemics, NB plays an important role as it works faster than other machine learning algorithms for the better predictions [8].

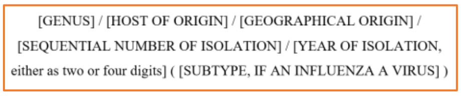
1.3.Influenza Disease

The Influenza illness is instigated by the elements of the orthomyxovirus family in which there four sorts of influenza viruses categorised as A – D. from different studies, it is observed that orthomyxoviruses such as

Influenza virus contains a segmented genome including 8 segments for Influenza A and B viruses while 7 segments are belonging to Influenza C and D, the segmented genome imitates in the nucleus [9].

1.4.Classification of Influenza Virus:

Basically, influenza germs are contained in Ribonucleic acid (RNA) viruses' category. Clinically, they are categorized into three types A, B and C with their further subtypes. Influenza A has 8 generic fragments containing 18 hemagglutinin (HA) and 11 neuraminidase (NA) subtypes. For instances: **B/Florida/06/2009** is the 6th influenza Type-B virus strain in 2009 in Florida. **C/Paris/3/74** is the 3rd influenza Type-C virus strain in 1974 in Paris [9]. Influenza virus classification is prescribed in its generalized form as in Figure 1. With H1N1, H2N1 and H3N2, humans are affected. Animals and birds are infected by mixtures of H5N1, H7N9 and H9N2. While, H17N10 and H18N11 are found in bats [10]. In three different categories A, B and C. Influenza A produces periodic pandemic globally. Influenza A is the most dangerous and infectious caused dangerous and undying simple respiratory diseases. While influenza C causes minor diseases and it appeared in mostly animals and birds [10].



[GENUS] / [HOST OF ORIGIN] / [GEOGRAPHICAL ORIGIN] / [SEQUENTIAL NUMBER OF ISOLATION] / [YEAR OF ISOLATION, either as two or four digits] ([SUBTYPE, IF AN INFLUENZA A VIRUS])

Figure 1: Influenza Virus Strain Pseudocode Presented in Literature

2. RELATED WORK

Globally, pandemic forecasting / prediction is the most common research area for scientists and researchers in the field of mathematical biology just to get the optimum solution and the best outcome within the time [11]. The reason is to take in-time medicine suitable decisions. Influenza occurrence can occur in unexpected times, so for future disasters, the forecasting of this influenza disease in very much crucial [12].

Marco Cacciabue et al. (2023) developed model based on an advanced machine learning algorithms for having accurate predictions, some influenza type A and B types were taken into consideration for analysis and future outbreaks and predictions. The models have the 99.5% and 99.3% accuracy for FULL HA and HA1 representations respectively [13].

Daniel Palomar et al. (2023) presented a machine learning model for influenza A Virus by using hemagglutination inhibition (HI) that exactly predicted outputs of HI comprising of human IAV H3N2 viruses by taking the sample data of their subtype HA1 categorizations and related metadata. The Random Forest algorithm was used to generate the assumptions or predictions as well as future outbreaks. It is also recommended in the paper that interpretable artificial intelligence, such as Sharpley Additive exPlanations (SHAP) may be the great source to study the exchanges between amino acid sites. [14]

Steven Riley et al. (2023) evaluated the performance of XGBoost forecast models for the influenza data of 30 countries between 2010 and 2018 by comparing with another model in null state and a antique regular model using MZE and mMAE. Their results suggest that machine learning framework for predicting influenza cases may be accepted as a valuable health tool worldwide in the upcoming scenarios [15]

Jie Zhang et al. (2023) suggested the forecasting of Influenza pandemic using current and historical values by proposing the Dynamic Virtual Graph Significance Networks (DVGSN). Visualized algorithm having graphs could easily determine the knowledge from similar infection situations without limitation of time window. It was the initial effort to learn dynamic virtual graph for time-series estimate tasks and is suitable in the arenas of community well-being, natural sciences and so on [1].

Avishai Halev et al. (2023) suggested a model to forecast the outbreaks of swine farms throughout the production process. It predicts the future disease outbreak in two production systems, which yields the good ability of forecasting outbreaks with 0.798 accuracy in initial system and accuracies of 63&, 71% and 70% on piggy generative and respirational disorder, influenza A virus and Mycoplasma hyopneumoniae in the second system [16].

Shang-Kai Hung et al. (2023) conducted a comprehensive study on the dataset obtained from the five emergency departments in United States and Taiwan from 2015 to 2020. The seven different machine learning algorithms were utilized for the predictions and classifications. The algorithms include SVM, Xtreme Gradient Boosting, Conditional RF, RF and ANN. By comparing all the algorithms applied, extreme gradient boosting achieved superior performance with an area under the 82% receiver operating characteristic curve with 92% sensitivity, 89% specificity and 72% accuracy [17].

Chengbing Huang et al. (2023) worked for Hemagglutinin HA and suggested that HA is the main symptom for spreading the viral infections by imposing the combination amongst the host membrane and the disease. This study designed a computational model to recognize the existence of HA. After applying the particular algorithms, the model attained the accuracy of 95.85%. it may help the biochemical scholars for the study of HA for future predictions [18].

Edna Marquez et al. (2023) facilitated the deciding in the medical distinction between affected and not-affected influenza patients by applying an appropriate machine learning algorithm for a large dataset created on their indications and demographic features. On applying different algorithms, it was suggested that Random Forest method had a high level accuracy specially in the clinical diagnosis where it was difficult and challenging to execute the molecular tests [19] .

3. INFLUENZA DATA COLLECTION

Data includes monthly wise number of cases for different types of Influenza with a combined set of sentinel and non-sentinel surveillances. The data types for the selected regions include the dataset from 2009 to mid of 2023 to evaluate the optimum change in the number of cases. Types of Influenza are related to type A and B, such as: A-H1N1, A-H3, B-Victoria and B-Yamagata. Another dataset is obtained in the form of outbreaks classification of China for the same period i.e., 2009-22 including Global outbreak, Local outbreak, Sporadic and Widespread outbreak Influenza dataset [20], [21].

4. RESEARCH METHODS

4.1.Logistic Regression

It is one of the most famous ML methods which is operated on the data which has categorical dependent variable [5, 22]. the elementary equation for linear regression with numerous independent variables is:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_\rho X_\rho + \sigma(Y) \qquad \textbf{(1)}$$

In above Equation (1) , β_0 is the intercept, or the point at which the regression line touches the vertical Y axis. This is considered a constant value.

Equation

(2) represents the mathematical format of Logistic regression [23] while Equation (3) is known as ***logistic transformation (logit)*** by applying the logarithmic function.

$$Y = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_\rho X_\rho}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_\rho X_\rho}} \qquad \textbf{(2)}$$

$$\boldsymbol{OR}$$

$$ln\left(\frac{Y}{1 - Y}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_\rho X_\rho \qquad \textbf{(3)}$$

4.2. Support Vector Machine (SVM)

For a certain dataset, it provides a better sketch of studying the problem of gaining knowledge, making forecasting or predictions as well as the decision making operations [7].

4.2.1. Kernel functions in SVM

SVM algorithm had utilized kernel-based techniques for the first time in machine learning techniques. The purpose was to remove barrier in classification of data as well as the scattering of the data in various direction [24]. ***Linear Kernel*** is the basic type of kernel used in SVM algorithm. It is the faster than other kernel functions used in machine learning algorithms involving classification. It can be mathematically expressed as in Equation (4).

$$K(x_i, y_i) = x_i \, . \, y_i \qquad \textbf{(4)}$$

Polynomial kernel is favored when the dataset has the non-linearity [25]. Due to its low precision and efficiency, it is not mostly used as compared to other kernels. Mathematically it is written as in Equation (5).

$$K(x_i, y_i) = (x_i \, . \, y_i + 1)^d \; ; \qquad \textbf{(5)}$$

$$\mathrm{d} = (1,2,3, \ldots)$$

In equation (5) , "**d**" denotes the degree of polynomial.

Radial basis function (RBF) Kernel is mostly used for Non-linear dataset. Since it has a large convergent region and hence it is considered as the best classification kernel. It can be mathematically expressed as in Equation (6).

$$K(x_i, y_i) = \exp\left(- \frac{\left\|x_i - x_j\right\|^2}{2\,\sigma^2}\right) \qquad \textbf{(6)}$$

4.3.Naïve Bayes Algorithm

It works for n-dimensional influenza training dataset and mostly used in text classification. In the prediction and classification of medicines specially pandemics, NB plays an important role as it works faster than other machine learning algorithms for the better prediction scores [8].

4.4.Random Forest Algorithm

When the classification and regression is involved in data sciences, Random Forest method works efficiently as compared to other machine learning algorithms. In this algorithm, there are various forests generated based on the number of trees. Its accuracy becomes more higher, when the number of trees is greater [26]. It chooses the observations and creates a decision tree and the decision is made based on the mainstream or the large number of the outcomes having the common characteristics [27, 28].

4.5.Train-Test Split Evaluation

It is a technique for defining the performance of a machine learning algorithm [29]. This method is helpful for classification problems and is widely used for supervised learning algorithm [30]. *Train Dataset* refers to the subset, in which the machine learning model is being fit. *Test Dataset* refers to the subset, in which the fitted machine learning model is then evaluated.

4.6.Confusion matrix

The Confusion Matrix is an N×N matrix practised for determining the performance of classification model for assumed set of test data [31]. In the matrix, there are two divisions which are **predicted values** and **actual values** along with the total number of predictions [32]. For a binary classification problem, we would have a **2×2 matrix**, as shown in Figure 2.

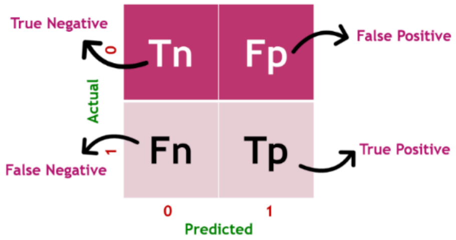


Figure 2: Confusion Matrix as an example

5. IMPLEMENTATION AND RESULTS

There have been for algorithms implemented on four different influenza datasets and they are explained in this section along with their comparison.

5.1. Dataset 1 : Influenza A-H1N1 and A-H3 in Pakistan (2009-22)

As far as the types of influenza are concerned along with their number of cases in Pakistan, here Figure 3 shows the number of cases for the type AH1N1 and AH3. The vertical axis shows the number of cases while the horizontal axis represents number of months for total 168 months in 14 years i.e., 2009-22.

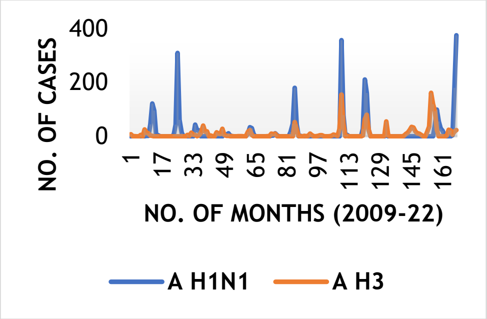
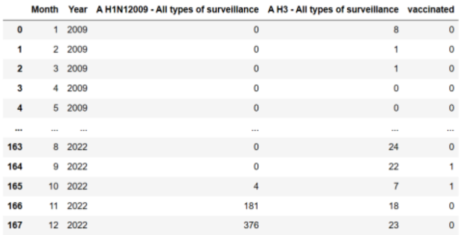


Figure 3: Visualization of Influenza Dataset 1

The whole data has been fragmented using Train-test split method as 60% (0.6) for Train dataset and 40% (0.4) Test Dataset from the Figure 4 shown for the vaccinated column as Target variable as well as AH3 and AH1N1 as the feature variables.



	Month	Year	A H1N12009 - All types of surveillance	A H3 - All types of surveillance	vaccinated
0	1	2009	0	8	0
1	2	2009	0	1	0
2	3	2009	0	1	0
3	4	2009	0	0	0
4	5	2009	0	0	0
...	...	...	...	...	...
163	8	2022	0	24	0
164	9	2022	0	22	1
165	10	2022	4	7	1
166	11	2022	181	18	0
167	12	2022	376	23	0

Figure 4: Influenza AH1N1 and AH3 Cases in Pakistan

Using *Logistic Classifier*, in Figure 5, the confusion matrix shows the accuracy score as 0.84, Precision score 0.79, Recall score 0.84 and F1 score 0.79. Hence the model score is 0.84 or 84%.

Using *SVM (Linear Kernel)*, in Figure 6, the confusion matrix shows the accuracy score as 0.78, Precision score 0.61, Recall score 0.78 and F1 score 0.68. Hence the model score is 0.78 or 78%.

Using *SVM (Polynomial Kernel)*, in Figure 7, the confusion matrix shows accuracy score as 0.82, Precision score 0.68, Recall score 0.82 and F1 score 0.74. Hence the model score is 0.82 or 82%.

Using *SVM (RBF Kernel)*, in Figure 8, the confusion matrix shows accuracy score as 0.69, Precision score 0.51, Recall score 0.69 and F1 score 0.59. Hence the model score is 0.69 or 69%.

Using *Gaussian Naïve Bayes*, in Figure 9, the confusion matrix shows the accuracy score as 0.76, Precision score 0.60, Recall score 0.76 and F1 score 0.68. Hence the model score is 0.76 or 76%.

Using *Random Forest*, In Figure 10, the confusion matrix shows the accuracy score as 0.75, Precision score 0.73, Recall score 0.75 and F1 score 0.74. Hence the model score is 0.75 or 75%.

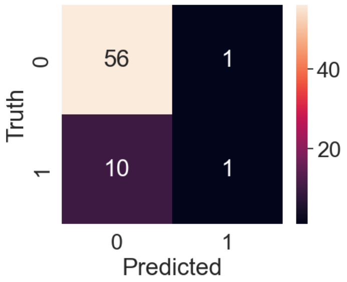


Figure 5: Confusion matrix of considered dataset 1 using LR

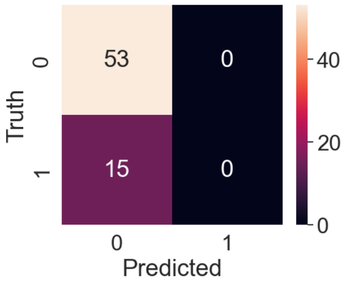


Figure 6: Confusion matrix of Dataset 1 using Linear SVM

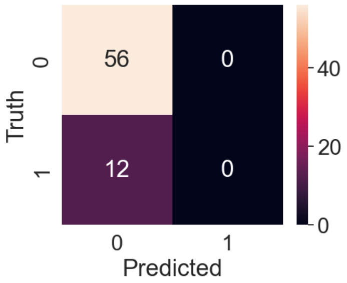


Figure 7: Confusion matrix of Dataset 1 using Polynomial SVM

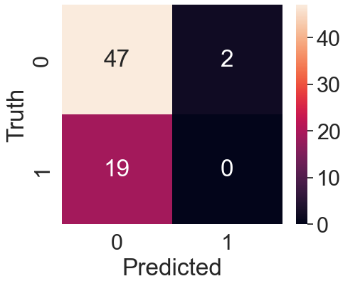


Figure 8: Confusion matrix of Dataset 1 using RBF SVM

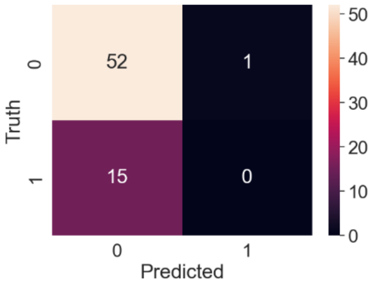


Figure 9: Confusion matrix of Dataset 1 using Gaussian NB

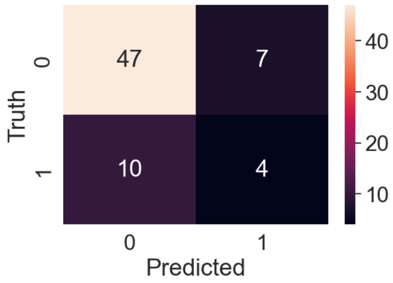


Figure 10: Confusion matrix of considered dataset 1 using RF

5.2. Dataset 2: Influenza AH1N1, AH3 in China (2009-22)

As far as the types of influenza are concerned along with their number of cases in China, here Figure 11 shows the number of cases for the type AH1N1 and AH3. The vertical axis shows the number of cases while the horizontal axis represents number of months for total 168 months in 14 years i.e., 2009-22.

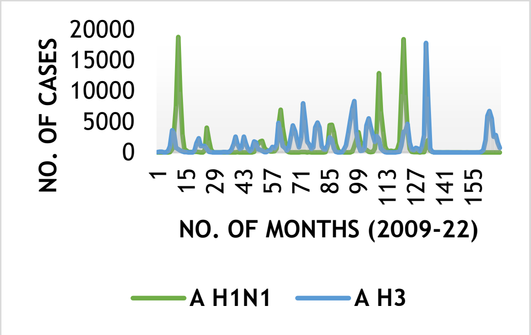
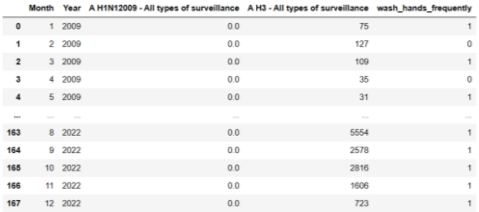


Figure 11: Visualization of Influenza Dataset 2

The whole data has been fragmented using Train-test split method as 60% (0.6) for Train dataset and 40% (0.4) for Test Dataset from the Figure 12 shown for the “Washing_hands_frequently” column as Target variable as well as AH3 and AH1N1 as the feature variables.



	Month	Year	A H1N12009 - All types of surveillance	A H3 - All types of surveillance	wash_hands_frequently
0	1	2009	0.0	75	1
1	2	2009	0.0	127	0
2	3	2009	0.0	109	1
3	4	2009	0.0	35	0
4	5	2009	0.0	31	1
...	...	...	...	...	...
163	8	2022	0.0	5554	1
164	9	2022	0.0	2578	1
165	10	2022	0.0	2816	1
166	11	2022	0.0	1606	1
167	12	2022	0.0	723	1

Figure 12: Influenza AH1N1 and AH3 Cases in China

Using *Logistic Classifier*, in Figure 13, the confusion matrix shows the accuracy score as 0.85, Precision score 0.76, Recall score 0.85 and F1 score 0.80. Hence the model score is 0.85 or 85%.

Using *SVM (Linear Kernel)*, in Figure 14, the confusion matrix shows the accuracy score as 0.85, Precision score 0.73, Recall score 0.85 and F1 score 0.78. Hence the model score is 0.85 or 85%.

Using *SVM (Polynomial Kernel)*, in Figure 15, the confusion matrix shows accuracy score as 0.78, Precision score 0.68, Recall score 0.78 and F1 score 0.73. Hence the model score is 0.85 or 85%.

Using *SVM (RBF Kernel)*, In Figure 16, the confusion matrix shows accuracy score as 0.82, Precision score 0.68, Recall score 0.82 and F1 score 0.73. Hence the model score is 0.85 or 85%.

Using *Gaussian Naïve Bayes*, in Figure 17, the confusion matrix shows the accuracy score as 0.80, Precision score 0.66, Recall score 0.80 and F1 score 0.72. Hence the model score is 0.80 or 80%.

Using *Random Forest*, in Figure 18, the confusion matrix shows the accuracy score as 0.87, Precision score 0.81, Recall score 0.87 and F1 score 0.84. Hence the model score is 0.87 or 87%.

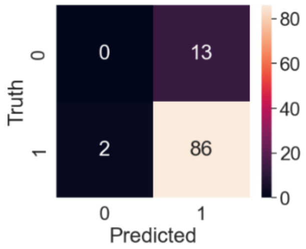


Figure 13: Confusion matrix of considered dataset 2 using LR

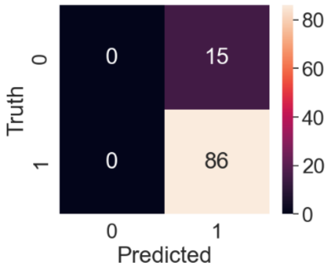


Figure 14: Confusion matrix of Dataset 2 using Linear SVM

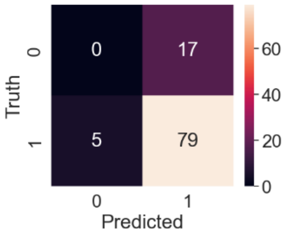


Figure 15: Confusion matrix of Dataset 2 using Polynomial SVM

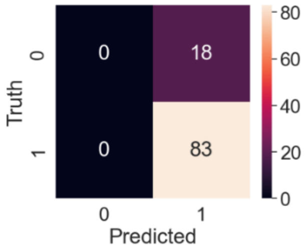


Figure 16: Confusion matrix of Dataset 2 using RBF SVM

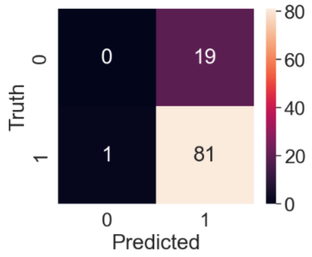


Figure 17: Confusion matrix of Dataset 2 using Gaussian NB

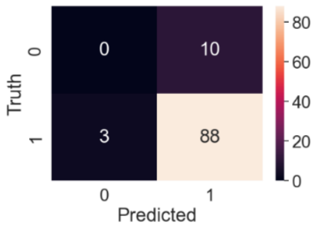


Figure 18: Confusion matrix of considered dataset 2 using RF

5.3. Dataset 3: Influenza B-Type in China (2015-19)

For influenza Type-B, here Figure 19 represents the graph of number of cases found in China for the time period 2015-19 showing type B-Yamagata and B-Victoria. The vertical axis shows the number of cases while the horizontal axis represents number of weeks for total five years i.e., 2015-19. Figure 21 is showing the type of Outbreak and the number of cases occurred in each outbreak categorically.

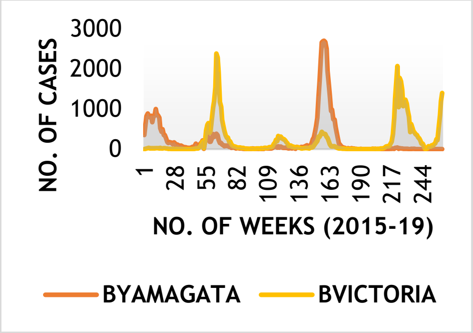
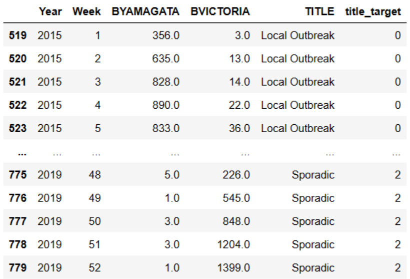


Figure 19: Visualization of Influenza Dataset 3



	Year	Week	BYAMAGATA	BVICTORIA	TITLE	title_target
519	2015	1	356.0	3.0	Local Outbreak	0
520	2015	2	635.0	13.0	Local Outbreak	0
521	2015	3	828.0	14.0	Local Outbreak	0
522	2015	4	890.0	22.0	Local Outbreak	0
523	2015	5	833.0	36.0	Local Outbreak	0
...	...	...	...	...	...	...
775	2019	48	5.0	226.0	Sporadic	2
776	2019	49	1.0	545.0	Sporadic	2
777	2019	50	3.0	848.0	Sporadic	2
778	2019	51	3.0	1204.0	Sporadic	2
779	2019	52	1.0	1399.0	Sporadic	2

Figure 20: B-Type Cases found in China in Period 2015-19

The whole data is divided using Train-test split method as 70% (0.7) for Train dataset and 30% (0.3) Test Dataset from Figure 20 shown for TITLE column as Target variable as well as BYAMAGATA and BVICTORIA as the feature variables. For the confusion matrix, **Local Outbreak** as **0**, **Regional Outbreak** as **1**, **Sporadic** as **2**, **Widespread Outbreak** as **3** values are set.

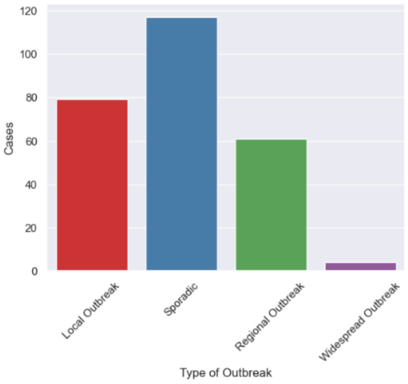


Figure 21: Type-B Influenza Cases Showing Type of Outbreak in China 2015-19

Using *Logistic Classifier,* in Figure 22, the confusion matrix shows the accuracy score as 0.63, Precision score 0.63, Recall score 0.63 and F1 score 0.51. Hence the model score is 0.63 or 63%.

Using *SVM (Linear Kernel),* in Figure 23, the confusion matrix shows the accuracy score as 0.62, Precision score 0.68, Recall score 0.62 and F1 score 0.56. Hence the model score is 0.62 or 62%.

Using *SVM (Polynomial Kernel),* in Figure 24Figure 15, the confusion matrix shows accuracy score as 0.61, Precision score 0.57, Recall score 0.55 and F1 score 0.49. Hence the model score is 0.61 or 61%.

Using *SVM (RBF Kernel),* in Figure 25, the confusion matrix shows accuracy score as 0.53, Precision score 0.58, Recall score 0.53 and F1 score 0.48. Hence the model score is 0.53 or 53%.

Using *Gaussian Naïve Bayes*, in Figure 26, the confusion matrix shows the accuracy score as 0.57, Precision score 0.49, Recall score 0.50 and F1 score 0.46. Hence the model score is 0.57 or 57%.

Using *Random Forest*, in Figure 27, the confusion matrix shows the accuracy score as 0.70, Precision score 0.69, Recall score 0.70 and F1 score 0.70. Hence the model score is 0.70 or 70%.

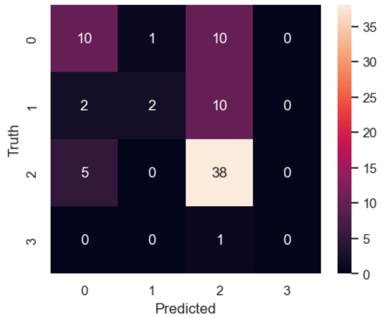


Figure 22: Confusion matrix of considered dataset 3 using LR

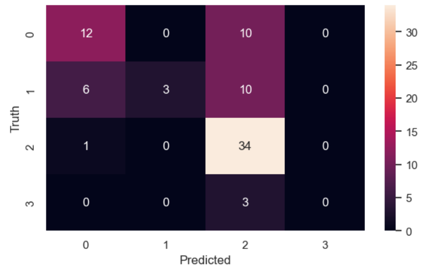


Figure 23: Confusion matrix of Dataset 3 using Linear SVM

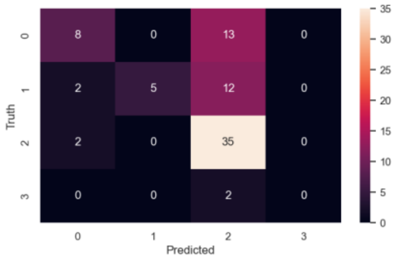


Figure 24: Confusion matrix of Dataset 3 using Polynomial SVM

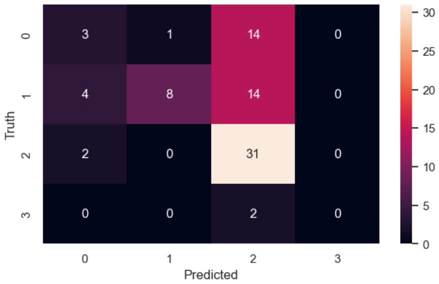


Figure 25: Confusion matrix of Dataset 3 using RBF SVM

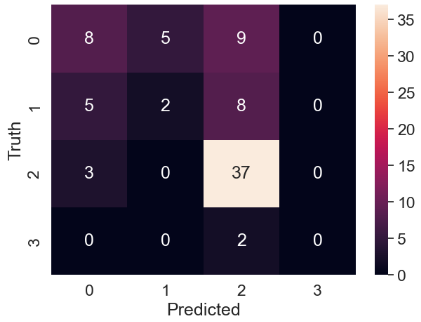


Figure 26: Confusion matrix of Dataset 3 using Gaussian NB

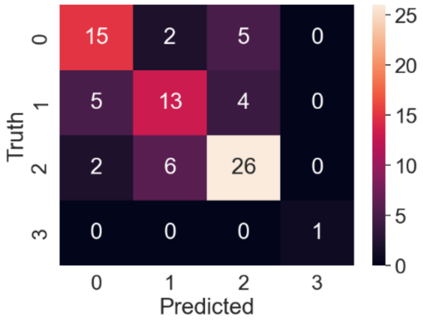


Figure 27: Confusion matrix of considered dataset 3 using RF

5.4. Influenza Type "A" and "B" in China (2010-19)

For influenza Type A and B, here Figure 28 represents the graph of number of cases found in China for the time period 2010-19 showing type A and B. The vertical axis shows the number of cases while the horizontal axis represents number of weeks for total ten years i.e., 2010-19.

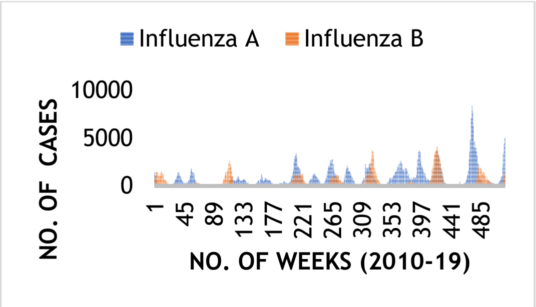
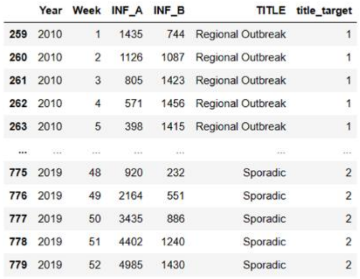


Figure 28: Visualization of Influenza Dataset 4

The whole data has been divided using Train-test split method as 70% (0.7) for Train dataset and 30% (0.3) Test Dataset from Figure 29 shown for TITLE column as Target variable as well as INF_A and INF_B as the feature variables. For the confusion matrix, **Local Outbreak** as **0**, **Regional Outbreak** as **1**, **Sporadic** as **2**, **Widespread Outbreak** as **3** values are set.



	Year	Week	INF_A	INF_B	TITLE	title_target
259	2010	1	1435	744	Regional Outbreak	1
260	2010	2	1126	1087	Regional Outbreak	1
261	2010	3	805	1423	Regional Outbreak	1
262	2010	4	571	1456	Regional Outbreak	1
263	2010	5	398	1415	Regional Outbreak	1
...	...	...	...	...	...	...
775	2019	48	920	232	Sporadic	2
776	2019	49	2164	551	Sporadic	2
777	2019	50	3435	886	Sporadic	2
778	2019	51	4402	1240	Sporadic	2
779	2019	52	4985	1430	Sporadic	2

Figure 29: Type A and B Cases found in China (2010-19)

Using *Logistic Classifier*, in Figure 30, the confusion matrix shows the accuracy score as 0.57, Precision score 0.57, Recall score 0.56 and F1 score 0.56. Hence the model score is 0.57 or 57%.

Using *SVM (Linear Kernel),* in Figure 31, the confusion matrix shows the accuracy score as 0.59, Precision score 0.58, Recall score 0.59 and F1 score 0.58. Hence the model score is 0.59 or 59%.

Using *SVM (Polynomial Kernel),* in Figure 32, the confusion matrix shows accuracy score as 0.61, Precision score 0.63, Recall score 0.61 and F1 score 0.58. Hence the model score is 0.61 or 61%.

Using *SVM (RBF Kernel),* in Figure 33, the confusion matrix shows accuracy score as 0.59, Precision score 0.58, Recall score 0.59 and F1 score 0.58. Hence the model score is 0.59 or 59%.

Using *Gaussian Naïve Bayes*, in Figure 34, the confusion matrix shows the accuracy score as 0.55, Precision score 0.57, Recall score 0.55 and F1 score 0.54. Hence the model score is 0.55 or 55%.

Using *Random Forest*, in Figure 35, the confusion matrix shows the accuracy score as 0.64, Precision score 0.65, Recall score 0.63 and F1 score 0.63. Hence the model score is 64%.

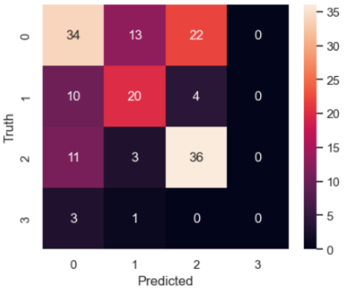


Figure 30: Confusion matrix of considered dataset 4 using LR

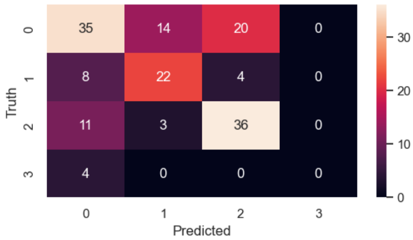


Figure 31: Confusion matrix of Dataset 4 using Linear SVM

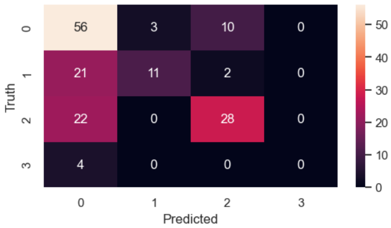


Figure 32: Confusion matrix of Dataset 4 using Polynomial SVM

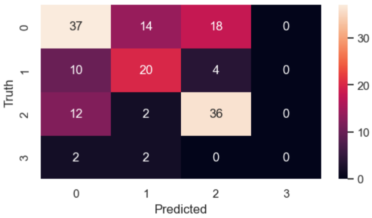


Figure 33: Confusion matrix of Dataset 4 using RBF SVM

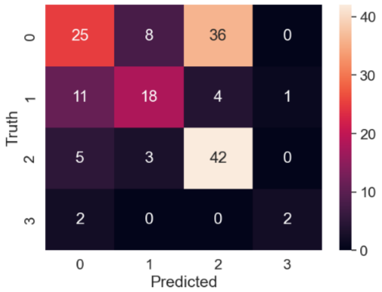


Figure 34: Confusion matrix of Dataset 4 using Gaussian NB

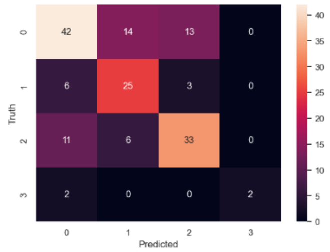


Figure 35: Confusion matrix of considered dataset 4 using RF

6. COMPARATIVE ANALYSIS OF ALL DATASET MODELS:

6.1. DATASET 1: Influenza A-H1N1 and A-H3 in Pakistan (2009-22)

In all the algorithms, there are better reading accuracies are observed but here Logistic regression algorithm yields the best accuracy of 86% among all the algorithms, hence this dataset can be best utilized for classification using Support Vector Machines. Its visualization can be observed in Figure 36.

6.2. DATASET 2: Influenza AH1N1, AH3 in China (2009-22)

In all the algorithms, there are better reading accuracies are observed but here Random Forest algorithm yields the best accuracy of 87% among all the algorithms, hence this dataset can be best utilized for classification using Random Forest Algorithm as observed in Figure 37.

6.3. DATASET 3: Influenza B-Type in China (2015-19)

In all the algorithms, there are better reading accuracies are observed but here Random Forest algorithm yields the best accuracy of 70% among all the algorithms, hence this dataset can be best utilized for classification using Random Forest Algorithm as observed in Figure 38.

6.4. DATASET 4: Influenza Type “A” and “B” in China (2010-19)

In all the algorithms, there are better reading accuracies are observed but here Random Forest algorithm yields the best accuracy of 64% among all the algorithms, hence this dataset can be best utilized for classification using Random Forest Algorithm. Its visualization can be observed in Figure 39.

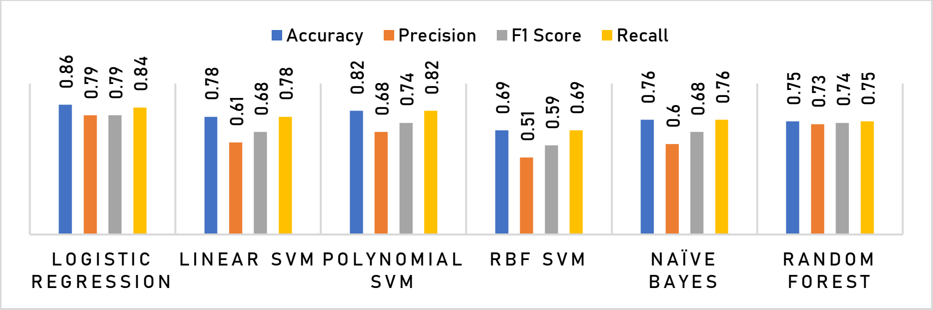


Figure 36: Comparative Analysis of ML Algorithms on Pakistan Dataset 2009-22

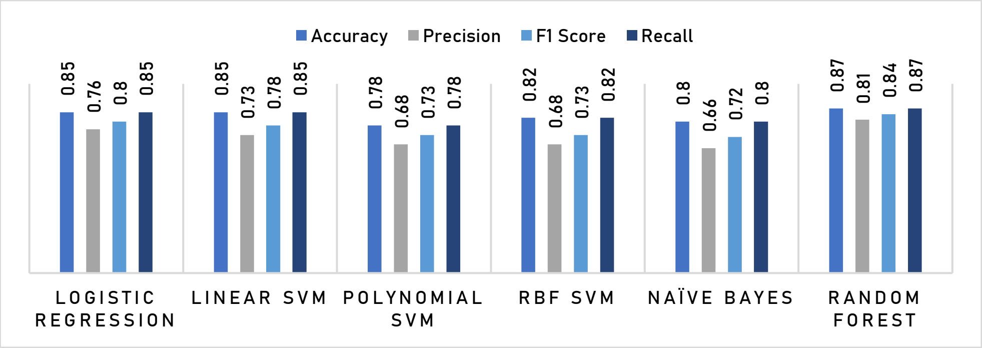


Figure 37: Comparative Analysis of ML Algorithms on China Dataset 2009-22

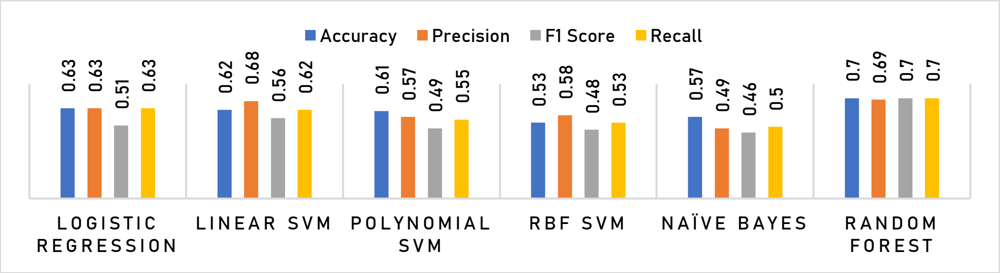


Figure 38: Comparative Analysis of ML Algorithms on B-Type China Dataset 2015-19

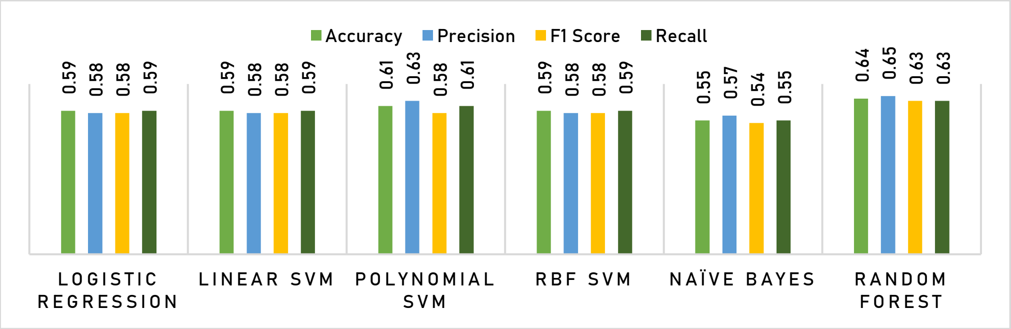


Figure 39: Comparative Analysis of ML Algorithms on Influenza China Dataset 2010-19

7. CONCLUSION AND FUTURE WORK

As we reflect on the outcomes, it becomes evident that the amalgamation of diverse machine learning approaches provides a robust framework for comprehensively understanding and addressing the complexities inherent in real-life datasets associated with infectious diseases.

We obtained average accuracy scores as 73% in Logistic Regression, 71% in Linear SVM, 70% in Polynomial SVM, 66% in RBF SVM, 67% in Gaussian NB and 74% in random forest algorithm. Hence, Figure 40 shows that Random Forest yields the best performance among all the classifiers for the considered datasets. Following are some points which can be considered for having better results in prediction as well as for outspreading the research work:

1. The rich data obtained from a particular region with huge symptoms.
2. Data can be more accurately visualized by having the best train-test split ratios and ML methods' kernels.

3. It can be extended to the deep learning algorithms such as Artificial Neural Networks.

8. DATA AVAILABILITY:

Considered datasets links are shared in this manuscript and will be available on request. The datasets were obtained from [20, 21].

9. ACKNOWLEDGEMENT:

I extend my deepest appreciation to my co-author Dr. Ubaida Fatima for her invaluable guidance and support and mentorship throughout the research process and the Department of mathematics, NEDUET Karachi.

10. CONFLICT OF INTEREST:

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

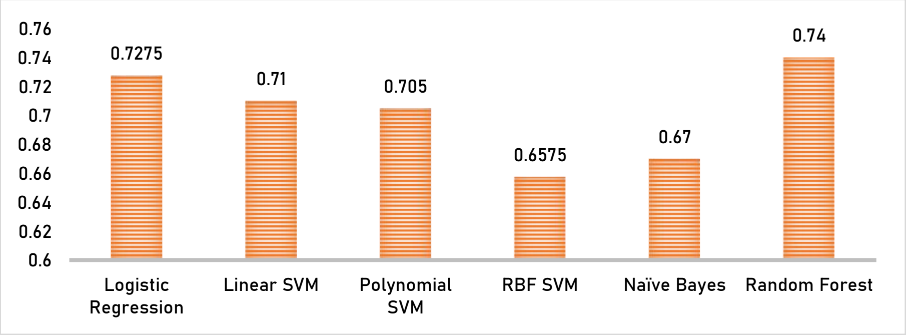


Figure 40: Average Accuracy Scores of Algorithms

11. REFERENCES

[1] Zhang, J., et al., *Predicting influenza with pandemic-awareness via Dynamic Virtual Graph Significance Networks.* Computers in Biology and Medicine, 2023. **158**: p. 106807.

[2] Khan, M., et al., *Forecast the Influenza Pandemic Using Machine Learning.* Cmc -Tech Science Press-, 2020. **66**: p. 331-340.

[3] Choi, R.Y., et al., *Introduction to Machine Learning, Neural Networks, and Deep Learning.* Translational Vision Science & Technology, 2020. **9**(2): p. 14-14.

[4] LaValley, M.P., *Logistic regression.* Circulation, 2008. **117**(18): p. 2395-2399.

[5] Hilbe, J.M., *Logistic regression models*. 2009: CRC press.

[6] Chauhan, V.K., K. Dahiya, and A. Sharma, *Problem formulations and solvers in linear SVM: a review.* Artificial Intelligence Review, 2019. **52**(2): p. 803-855.

[7] Jakkula, V., *Tutorial on support vector machine (svm).* School of EECS, Washington State University, 2006. **37**(2.5): p. 3.

[8] Bayes, T., *Naive bayes classifier.* Article Sources and Contributors, 1968: p. 1-9.

[9] Hutchinson, E.C. and Y. Yamauchi, *Understanding Influenza*, in *Influenza Virus: Methods and Protocols*, Y. Yamauchi, Editor. 2018, Springer New York: New York, NY. p. 1-21.

[10] NCIRD, *Types of Influenza Viruses.* CDC, 2023.

[11] Poirier, C., et al., *Real time influenza monitoring using hospital big data in combination with machine learning methods: comparison study.* JMIR public health and surveillance, 2018. **4**(4): p. e11361.

[12] Yin, Z., L.M. Sulieman, and B.A. Malin, *A systematic literature review of machine learning in online personal health data.* Journal of the American medical informatics association, 2019. **26**(6): p. 561-576.

[13] Cacciabue, M. and D.N. Marcone, *INFINITy: A fast machine learning-based application for human influenza A and B virus subtyping.* Influenza Other Respir Viruses, 2023. **17**(1): p. e13096.

[14] Shah, S., et al., *Seasonal antigenic prediction of influenza A H3N2 using machine learning*. 2023.

[15] Wang, H., K.O. Kwok, and S. Riley, *Forecasting influenza incidence as an ordinal variable using machine learning.* medRxiv, 2023: p. 2023.02.09.23285705.

[16] Halev, A., et al., *Outbreak Prediction in Swine Populations with Machine Learning.* 2023.

[17] Hung, S.-K., et al., *Developing and validating clinical features-based machine learning algorithms to predict influenza infection in influenza-like illness patients.* Biomedical Journal, 2023. **46**(5): p. 100561.

[18] Zou, X., et al., *Accurately identifying hemagglutinin using sequence information and machine learning methods.* Front Med (Lausanne), 2023. **10**: p. 1281880.

[19] Marquez, E., et al., *Supervised Machine Learning Methods for Seasonal Influenza Diagnosis.* Diagnostics, 2023. **13**(21): p. 3352.

[20] Saloni Dattani, F.S., Edouard Mathieu, Hannah Ritchie and Max Roser. *Influenza* [cited 2024 February 2024]; Influenza dataset]. Available from: https://ourworldindata.org/influenza.

[21] LACHMANN, A. *Weekly Influenza Reports by Country*. [cited 2024 February 2024]; Available from: https://www.kaggle.com/datasets/lachmann12/weekly-influenza-reports-by-country.

[22] He, Z., J. Camobreco, and K. Perkins, *How he won: Using machine learning to understand Trump's 2016 victory.* Journal of Computational Social Science, 2022. **5**(1): p. 905-947.

[23] Stoltzfus, J.C., *Logistic regression: a brief primer.* Academic emergency medicine, 2011. **18**(10): p. 1099-1104.

[24] Patle, A. and D.S. Chouhan. *SVM kernel functions for classification*. in *2013 International conference on advances in technology and engineering (ICATE)*. 2013. IEEE.

[25] Bodlaender, H.L., et al., *On problems without polynomial kernels.* Journal of Computer and System Sciences, 2009. 75(8): p. 423-434.

[26] Breiman, L., *Random forests.* Machine learning, 2001. 45: p. 5-32.

[27] Ziegler, A. and I.R. König, *Mining data with random forests: current options for real-world applications.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014. **4**(1): p. 55-63.

[28] Lokanan, M.E., *Incorporating machine learning in dispute resolution and settlement process for financial fraud.* Journal of Computational Social Science, 2023. **6**(2): p. 515-539.

[29] Salazar, J.J., et al., *Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy.* Journal of Petroleum Science and Engineering, 2022. **209**: p. 109885.

[30] Tan, J., et al., *A critical look at the current train/test split in machine learning.* arXiv preprint arXiv:2106.04525, 2021.

[31] Beauxis-Aussalet, E. and L. Hardman. *Visualization of confusion matrix for non-expert users*. in *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*. 2014.

[32] Maria Navin, J. and R. Pankaja, *Performance analysis of text classification algorithms using confusion matrix.* International Journal of Engineering and Technical Research (IJETR), 2016. **6**(4): p. 75-8.