Early Detection of Influenza Using Machine Learning Techniques

Sajal Maheshwari[1], Anushka Sharma[1], Ranjan Kumar[1], Pratyush[1]

[1] Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India
sajalmaheshwari21@gmail.com

Abstract. Influenza is an infectious disease that rapidly affects all living beings and the effect of this outbreak can be seen as infection in thousands of people yearly across the world. For example, the Spanish Flu is detected as one of the most devasting Influenza outbreak in 1918. Recently, the usage of Machine Learning techniques has been increased in medical research fields which include early diagnosis of a disease, pathology and classification of various diseases. In this study, the dataset is extracted Human Surveillance records from the Influenza Research database. Initially, 15651 records extracted over the period of 2006 to 2017 for data preprocessing and then 9548 records have been selected for further analysis. The next step follows the classification and analysis of data based on the four different machine learning methods, Support Vector Machines, K-nearest Neighbors, Artificial Neural Networks and Random Forest. Their performances are evaluated in terms of sensitivity, specificity and accuracy. It is evident from the experiments that Random Forest is one of the best machine learning techniques for early detection of the Influenza.

Keywords: Artificial Neural Network, classification, Influenza, KNN, Random Forest, Sensitivity, Specificity, SVM.

1 Introduction

Influenza is an infectious disease caused by RNA viruses from the Orthomyxoviridae family which infects the respiratory tract of animals, birds, and humans. It is found that usually Type A, Type B and Type C of the influenza viruses affect human beings, while the Type D influenza virus too has the potential to affect human beings. Influenza spreads around the world in yearly outbreaks, resulting in about three to five million cases of severe illness and causing about 290,000 to 650,000 deaths. “Spanish flu” resulted in 15 - 20 million deaths [1]. “Asian Influenza” caused 115,700 deaths, the “Hong Kong Influenza” pandemic result 112,000 deaths [2]. In 1997, some other Influenza strains, like H5N1 recorded 18 cases which caused 6 deaths in Hong Kong. Later, in 2003, it reappeared in Hong Kong and mainland China, causing great loss to poultry farms and other bird species. In 2003, another strain, H7N7, appeared in Holland with around 90 cases and 1 death in humans [3]. The 2009 H1N1 pandemic strain have around 30,000 cases reported across 74 countries by June 11, 2009[4].

During the 2019-20 influenza seasons, influenza was associated with 405,000 hospitalizations and 22,000 deaths [5]. The rapid evolution and mutation of influenza virus has created a generation of vaccine-resistant and antiviral medication-resistant variants; thereby the influenza virus infection remains a major public health threat [6].

Machine Learning Techniques are playing a significant role in detecting, predicting and classifying many medical issues [7-10]. Different machine learning techniques like- Decision Trees (DT), Random Forests (RF), Artificial Neural Networks (ANN), Bayesian Networks (Bayes Network), Support Vectors Machines (SVMs), and Gaussian Processes (GPs) have been used for many medical purposes [11-13]. Influenza virus can be predicted in a patient by implementing machine learning techniques on the patient's dataset which includes the symptoms.

Many sources provide access to the influenza human surveillance dataset. IRD's (Influenza Research Database) Influenza Human Surveillance Record [14] consists of 15,561 records out of which 9,611 records possess information regarding the Flu Test Status. There are several other databases available such as Health Data [15], DataWorld [16], WHO [17], and CDC [18]. IRD's Influenza Human Surveillance Record has been selected over these datasets as it has many attributes like Symptoms, Vaccination Status, Pre-Medical Conditions, Diagnosis Status, Post-Medication Status, and many other useful clinical attributes many of which are not present in the other databases.

Initially, four different machine learning techniques viz., Random Forests, Neural Networks, Support Vector Machine with linear kernel, and k-NN algorithm are used for predictive analysis of the influenza test status of the patient. This knowledge can be used by the physician/healthcare worker to diagnose the patient more accurately and economically. Although some Point-of-care (POC) test like RIDTs (Rapid influenza diagnostic tests), DIAs or NAATs (Nucleic acid amplification tests) tests give results in less than 15 minutes with high sensitivity percentage [19], however recently it has been observed that it is not enough to predict flu with high accuracy. This is because certain impediments which are economic, regulatory, and policy-related, as well as due to user's perceptions and other cultural barriers [20], prevent the use of many highly accurate POC tests.

2 Materials and Methods

Figure 1 describes the schematics of the methodology followed by the study. The data for this study is taken from Influenza Research Database (IRD) [14]. After preprocessing the data, symptoms and other clinical data are the main attributes. We have implemented four different machine learning techniques for predictive analysis on the preprocessed data.

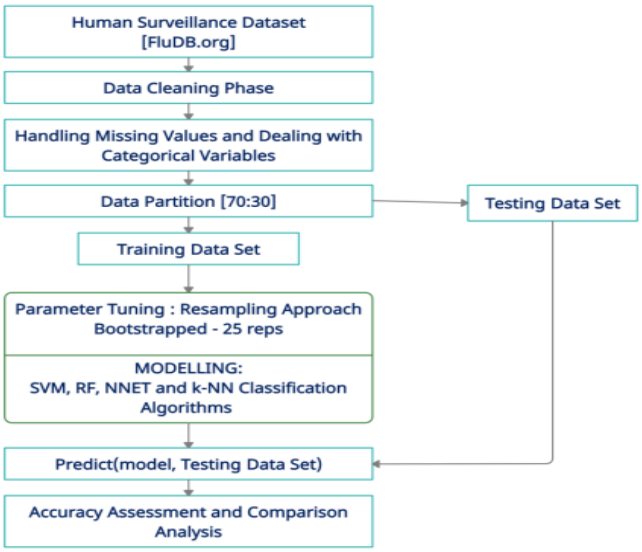

Fig.1. Flowchart of the study

2.1 Study Area

A stepwise logistic regression was performed by Monto et.al. [21] to determine the basic symptoms and patient characteristics which predicts an influenza infection. Their study determined that the best predictors of influenza infection were cough and fever, having a positive predictive value of 79% and patients with influenza like illness having both cough and fever within 48 hours of symptom onset are likely to have influenza. Further, in the paper by Marquez and Barrón [22] the goal was to create an artificial intelligence system to support a diagnosis of influenza on the basis of a data of the Mexican population wherein relevant factors such as clinical symptoms of influenza, patient's clinical record and other factors were utilized. It has been also determined in [23] that machine learning methods perform better than the expert-built Bayesian model for the detection of infectious diseases such as influenza. Proposed MSDII-FFNN, a feed-forward neural network based model, predicted with the 90 percent accuracy in predicting the influenza pandemic using different parameters [24].

The present study is based on influenza's database of the Human Surveillance Records accessed from the Influenza Research Database website [14]. This database contains 15651 records of 22 different health institutions of different countries. For the Influenza research community, the source has been very helpful in terms of projects involving sequence and structure analysis, comparative genomics, and virus phenotype studies among others. However, among all the projects, only 38 projects/papers are based on machine learning prediction, and out of these projects, there is no paper/project which has an objective similar to the present investigation. **Figure**

2 shows the distribution of records/samples of patients in different countries in the data used in the present study.

2.2 Data Used

Influenza Human Surveillance Data acquired for the years 2006–2017, was downloaded from the Influenza Research Database website [14]. Initially the data set consisted of 58 attributes with 15651 records. **Figure 3** shows the Year wise and Gender wise distribution of the records in the data.

Fig.2. Country wise distribution of Records in Dataset

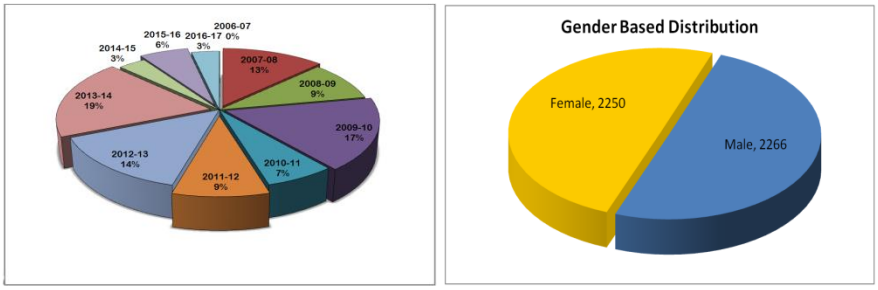

Fig.3. (Left) Year-Wise Distribution. **(Right)** Gender-Wise Distribution of 'Positive' Class

Certain columns/attributes related to the drug resistance, proteins sequence, strain name, contact email address, pre –visit medications, post-visit medications, test sample type, disease outcome, miscellaneous comments, ARDS and hospitalization duration were removed in the beginning itself as these attributes were not relevant for our work. For our work we are primarily interested in the symptoms, vaccination status, and medical conditions of the patient for the purpose of early prediction of the Flu in

the patient. The total relevant records for the positive flu cases in the data are almost equally distributed gender-wise, i.e., Female (48.74%) and Male (51.26%) approximately. However, in the data the distribution based on Vaccination Status, the records with 'Not Vaccinated' status (87.28%) dominates heavily over 'Vaccinated' (12.72%) records.

2.3 Data Pre-Processing

In order to extract the relevant information for the analysis, several pre-processing techniques have been implemented on the dataset. Initially, "Flu test status" is considered the target variable in which the positive label (1) denotes "Positive to influenza infection" while negative label (2) denotes "Negative to influenza infection". 'Symptoms' attribute consists of 24 different symptoms with their state. For each record, 24 or less symptoms are provided. To identify each symptom individually for each record, we divide the 'Symptoms' attribute in 24 'Symptom names' attributes that hold the binary values 0 (for 'No' or 'empty' values) and 1(for 'Yes'). **Table 1** shows the distribution of the 'Symptoms'.

Initially, there are 71 unique combinations of medical diseases for records in data. To make it relevant for the data, four most repeating factors were identified and based on these; a power set is defined over these 4 factors, dividing the factors in 16 unique combinations. **Table 2** of Medical Condition defines the value substitution in the data.

Most of the missing values are handled by substituting a value of 0 for them. Here it is assumed that the 'NA' values means that either that parameter is not known while collecting the data or since that value had no positive response for the particular patient record, i.e. left empty by the data manager of the source. We cannot remove or substitute these values with any other positive value. Hence, we directly replaced these values with 0 in our data. In order to handle 63 records with unknown values of gender, record is removed from the data because values assigned to 'Gender' are not binary values. The gender, 'Female' is represented by '1' and 'Male' is represented by '2'.At this stage 14,592 records are present in the preprocessed data, with 5255 records stating Flu Test Status as 'Positive' and 9333 records stating Flu Test Status as 'Negative'.

Finally, the improper ratio (5255:9333) of the Flu Test Status 'Positive' records to 'Negative' values needs to be balanced, A careful observation of the 5040 records yields the information that among the data there are patient's with Flu Test Status 'Negative' with all the 'Symptom' attributes as '0' and Vaccinated Status as '1'. Clearly, these entries are not relevant for our analysis, thus removing these makes the Positive to Negative ratio as 5255: 4293, which is almost a balanced one.

2.4 Training and Testing Sample Datasets

For the analysis purposes, standard R Tools are used. Training samples and testing samples were acquired in 70:30 ratios with 6685 training samples and 2863 testing samples. Distribution of the dependent variable 'Flu Test Status' in the training sample divided as 'Positive' (55.03%) and 'Negative' (44.97%).

Formula defined for our analysis:

Flu Test Status ~ Gender + Vaccination Status + (All the 24 Symptoms) + Medical Conditions[1]

Table 1. Symptoms Distribution

Symptom Name	Values
Running Nose	0,1
Cough	0,1
Myalgia	0,1
Headache	0,1
Throat	0,1
Fever	0,1
Fatigue	0,1
Temperature	0,1
Diarrhea	0,1
Nausea	0,1
Sudden Onset	0,1
Short Breath	0,1
Malaise	0,1
Wheezing	0,1
Chills	0,1
Vomiting	0,1
Loss of appetite	0,1
Arthalgia	0,1
Ear Ache	0,1
Aches	0,1
Sinus Congestion	0,1
Rash	0,1
Chest Pain	0,1
High_temp_home	0,1

Table 2. Medical Conditions Distribution

Factors Combination	Values
Other	0
Immuno	1
Asthma	2
Diarrhea	3
Chronic	4
Immuno, Asthma	5
Immuno, Diarrhea	6
Immuno, Chronic	7
Asthma, Diarrhea	8
Asthma, Chronic	9
Diarrhea, Chronic	10
Immuno, Asthma, Diarrhea	11
Immuno, Asthma, Chronic	12
Immuno, Diarrhea, Chronic	13
Asthma, Diarrhea, Chronic	14
Immuno, Asthma, Diarrhea, Chronic	15

2.5 Classification Algorithm

Support Vector Machines (SVM)

Support Vector Machine is a machine learning algorithm that is mainly used for binary classification. In SVM, outliers are weighed down, if it is not possible to separate the classes using a linear classifier in the space; thereby projecting the data points into another dimension (usually higher). There are several tuning parameters in SVM,

[1] For the Symptoms attribute, refer to **Table 1**.

depending on the type of kernel used. These parameters include regularization parameter(C), kernel, degree, and gamma. Authors are interested in the 'linear' kernel. Mathematically,

$$d_H(\phi(x_0)) = \frac{|w^T(\phi(x_0)) + b|}{||w||_2}$$

where 'w' is the vector normal to the hyperplane, b is the intercept or bias term for the hyperplane, dH is the distance of the point from the line separating the hyperplane.The above equation gives the prediction value for the positive class (if dH > 0) and negative class (if dH < 0). In the train() method, method = 'svmLinear' sets the kernel to 'Linear' value and the default value of C =1 for classification.

k -Nearest Neighbour (k-NN)

k-NN or k-Nearest Neighbor algorithm is a non-parametric method that is used for estimation, prediction, and classification. k-NN classifies a new unclassified record by comparing it to the most similar set of records in the training set. To achieve this, the training data set is stored. For estimation and prediction, a locally weighted averaging method is used. 'k' in the k-NN defines the number of nearest neighbors included in the majority of the classification process [25]. A common distance function used is Euclidean distance:

$$d_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

For estimation and prediction, locally weighted averaging method is used. In locally weighted averaging, the estimated target value $\hat{y}$ is:

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

By default, k equals sqrt (N), where N is the total number of samples. Since the createDataPartition() method of the 'Caret' package in R is used for creating partitioning, it uses the bootstrapped sampling, i.e. the values selected for k equal to 5, 7, and 9.

Neural Networks

As pointed out by Ripley [26], neural networks consist of units, which are further arranged in layers. These layers consist of Input Layer, Hidden Layers, and Output Layer. All these layers are linked using a particular weight (w) which multiplies the signal traveling along with them by that factor. Also, each unit sums its input and adds a bias (constant) as a total input (x), and applies a function φ to input(x) to return as output(y).

$$y_k = \phi_o\left(\alpha_k + \sum_{i \to k} w_{ik} x_i + \sum_{j \to k} w_{jk} \phi_h\left(\alpha_j + \sum_{i \to j} w_{ij} x_i\right)\right)$$

The default number of iterations 'maxit' in nnet is 100. The 'nnet' package of the R provides the default value for decay as 0 and it allows the skip-layer units by initializing the size to 0 but it is a required argument. Also, the train() of 'Caret' package in R also provides an implementation of 'nnet' method. In this, size and decay tuning parameters are optional and set as size as 1, 3, and 5, and decay as 0.0001, 0.1, and 0.

Random Forests (RF)

According to Leo Breiman. [27], a random forest is defined as "A classifier consisting of a collection of tree-structured classifiers {h(x, k), k = 1,..} where the {k } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x." For classification, the default value of x is [√p] and minimum node size is 1, where p is number of variables in x. The advantage of using Random Forest is that it is claimed to "cannot over fit the data" in most cases. Increasing the tree samples will not over-fit the random forest sequence [28]. The train function in R allows implementing Random Forest with the tuning parameter 'mtry' = sqrt (p). Hence, the values of 'mtry' are 2, 14, and 27.

2.6 Accuracy Assessment and Comparisons

Parameters used for evaluation are as provided in the work of Zhu et. al.,[29]. *Accuracy:* Accuracy is defined as the number of correct cases predicted to the total number of cases present in the dataset. It measures the degree of accuracy of a case on a condition.

$$\text{Accuracy} = (TN + TP) / (TN+TP+FN+FP)^2$$

Sensitivity: Sensitivity is defined as the proportion of the true positives cases that are discovered accurately by the diagnosis (diagnostic test). It is the percentage of people having infection from the dataset. Sensitivity is also called as- True Positive Rate (TPR).

$$\text{Sensitivity} = TP/(TP + FN)$$

Specificity: Specificity is defined as the true negatives cases that are discovered accurately from the diagnosis. It is the percentage of people having no infection from the dataset. Specificity is also called as- True Negative Rate (TNR).

$$\text{Specificity} = TN/(TN + FP)$$

F1-Score: F1 score combines precision and sensitivity (recall) in one metric. It calculates the harmonic mean between precision and sensitivity. Range of F1 score is [0, 1].

$$\text{F1-Score} = 2 * (\text{Sensitivity} * \text{Precision}) / (\text{Sensitivity} + \text{Precision})$$

Here, Precision = TP/ (TP+FP).

ROC curve: ROC curve or Receiver Operating Characteristic curve plots Sensitivity versus Specificity at different classification thresholds where, FPR = FP/ (FP+TN).

AUC: AUC or Area under the ROC curve provides a performance measurement for the classification problems. It measures how well predictions are ranked and the quality of the technique's predictions. It calculates the entire 2-Dimension area under the ROC curve. The higher the value of AUC, the better the techniques are at predicting cases with infection or no infection.

---

[2] TP (True Positive): number of True Positive values
TN (True Negative): number of True Negative values
FP (False Positive): number of False Positive values
FN (False Negative): number of False Negative values

3 Results

Initially, the performance measures selected for the models are Accuracy, Sensitivity and Specificity. **Table 3** provides the results for these models. The results are shown in ranges as the performance measure varies for different samples for training and testing. The average mean of the accuracy with the worst to best training samples acquired is 5.185%, i.e., based on the selected samples for training and testing, accuracy may vary by 5% (approximately).

Table 3. Value Band of Performance Measures for Different Machine Learning Techniques

Methods	Accuracy	Sensitivity	Specificity
SVM with Linear Kernel	76.08 - 82.36	68.44 - 72.97	85.76 - 93.86
k- Nearest Neighbour	79.75 - 83.58	66.44 - 80.85	87.26 - 96.6
Neural Network	79.85 - 83.83	67.00 - 75.82	93.63 - 96.12
Random Forest	**80.24 - 86.9**	**67.13 - 82.11**	**92.77 - 96.84**

From **Figure 4**, it can be discerned that the method of the random forests performs better analysis than other three techniques. On an average Random Forest gives 86.9% accuracy which is 3.64% more accurate results than others based on the parameter of accuracy. Although the difference in sensitivities of k-NN and Random Forest is less than 2%, but vis-a-vis the other two models it is higher by more than 5%. Similarly, the specificities of these two models, k-NN and Random Forest is less than the specificities of the other two models.

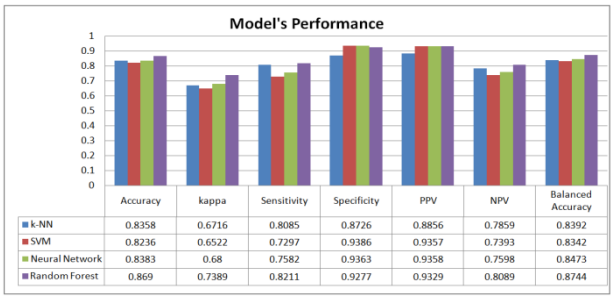

Fig.4. Performance Measures for different Machine Learning Techniques

Again in **Figure 5** and **Figure 6,** clearly represents a better performance for the Random Forest over the other techniques. On the other hand, the AUC graph shows a better result for the Neural Networks compared to the k-NN technique but the performance is opposite in the F1- Score Graph. This prediction occurs due to the higher variance in the values with respect of sensitivity and sensitivity of the model.

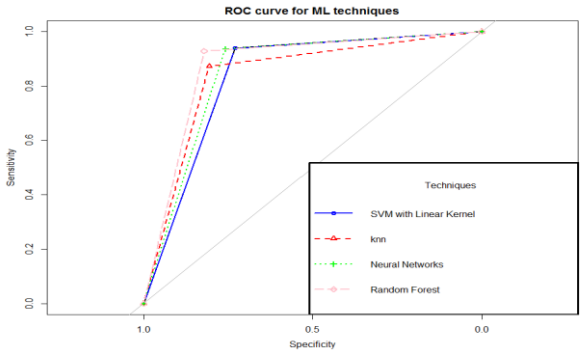

Fig.5. ROC curve of the Machine Learning Techniques

As in [30], it is clearly verified from the AUC graph, that all the selected technique fits on the data model and out of these, random forest shows the best results.

From the above results, currently, random forest model is the best performing model and should be chosen for developing system for the same. But as the data will grow, it might be possible that neural network (nnet) can overrule the random forest accuracy and predict more efficiently. Due to the categorical data and large number of training examples, SVM with linear kernel and k-NN performances are lower than other two techniques.

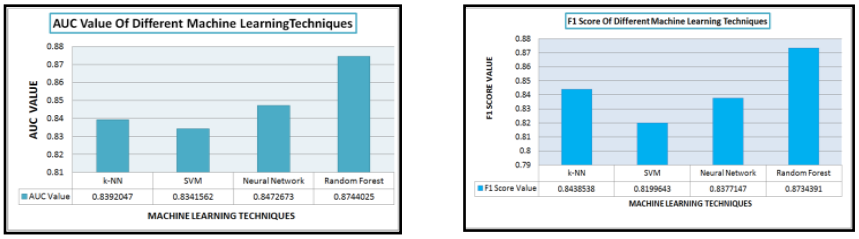

Fig.6. (Left) AUC Graph of Different Machine Learning Techniques.
(Right) F1-Score Graph of Different Machine Learning Techniques.

4 Conclusion

In this paper, four different well established machine learning techniques have been implemented over the analytic data of Human Surveillance Records accessed from the Influenza Research Database. Out of these techniques, Random Forest returns better results with accuracy ranging from 82% - 87% as compared to SVM, k-NN and Neural Networks which show accuracies of 77%- 82%, 79%-83.5% and 79%-84%, respectively. On the basis of ROC Curve, AUC graph and F1 Score graph, Random Forest dominates over other techniques. It is also found that the methods of Neural Networks and k-NN perform almost similar to each other.

From this study, it is clearly identified that any medical tests including RIDTs and DIAs gives accuracy less than 80% of the positive prediction. But this model using any of the four techniques will guarantee to give better results, i.e. > 80%. And also, compared to NAATs, it will be more economical feasible because its features are the common medical tests and information from the patient's medical history.

5 Future Scope

To further improve the accuracy of the models, feature selections, identify top features; sampling using statistical parameters will be done to acquire much better predictions. A mixed strategy of top techniques or a new modeling technique may also develop to improve the efficiency of the machine. Several other parameters will also be included to train the model at the much deeper levels, based on strain type, age category, geographical distributions, etc.

References

1. Peter Spreeuwenberg, MadelonKroneman, John Paget.: Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic. American Journal of Epidemiology 187(12), 2561-2567 (2018). DOI: 10.1093/aje/kwy191
2. W. Paul Glezen.: Emerging Infections: Pandemic Influenza. Epidemiologic Reviews 18(1), 64–76, (1996). DOI: 10.1093/oxfordjournals.epirev.a017917
3. J. S. Malik Peiris, Menno D. de Jong, Yi Guan.: Avian Influenza Virus (H5N1): a Threat to Human Health. Clinical Microbiology Reviews 20 (2), 243-267, (2007). DOI: 10.1128/CMR.00037-06
4. Seth J. Sullivan , et al..: 2009 H1N1 influenza. Mayo Clinic Proceedings 85(1), 64-76, (2010). DOI: 10.4065/mcp.2009.0588
5. Centers for Disease Controls and Prevention (CDC), https://www.cdc.gov/flu/about/burden/2019-2020.html, last accessed 2021/03/05.
6. Squires, R.B., Noronha, J., Hunt, V., García-Sastre, et al.: Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and Other Respiratory Viruses 6(6), 404-416, (2012). DOI: 10.1111/j.1750-2659.2011.00331.x
7. Nada Lavrač,: Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16(1), 3-23, (1999). DOI: 10.1016/S0933-3657(98)00062-1
8. Tan AC, Gilbert D. .: Ensemble machine learning on gene expression data for cancer classification. In: Proceedings of New Zealand Bioinformatics Conference, pp. 13-14. University of Glasgow, Te Papa, Wellington, New Zealand (2003).
9. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F.: Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis, Journal of Personalized Medicine 10(2), 21, (2020). DOI: 10.3390/jpm10020021
10. Rajalakshmi V., Sasikala D., Kala A. , A Predictive Analysis for Heart Disease Using Machine Learning. In: Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 1172. Springer (2021), DOI: 10.1007/978-981-15-5566-4_42
11. Geert Meyfroidt, Fabian Güiza, Jan Ramon, Maurice Bruynooghe, Machine learning techniques to examine large patient databases, Best Practice & Research Clinical Anaesthesiology 23(1), 127-143, (2009). DOI: 10.1016/j.bpa.2008.09.003

12. Li, H., Sun, F, Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences, Scientific Reports 8, Article number- 10032, (2018). DOI: 10.1038/s41598-018-28308-x
13. Jing Li, Sen Zhang, Bo Li, Yi Hu, et al., Machine Learning Methods for Predicting Human-Adaptive Influenza A Viruses Based on Viral Nucleotide Compositions, Molecular Biology and Evolution (37)4, 1224-1236, (2020). DOI: 10.1093/molbev/msz276
14. Influenza Research Database (Human Surveillance Record), https://www.fludb.org/brc/influenza_humanSurveillanceData_search.spg?method=ShowCleanSearch&decorator=influenza, last accessed 2021/01/ 28.
15. Health Data, https://healthdata.gov/dataset/influenza-surveillance, last accessed 2021/03/05.
16. Data.World, https://data.world/datasets/influenza, last accessed 2021/03/05.
17. WHO (World Health Organization), https://www.who.int/influenza/gisrs_laboratory/flunet/en/, last accessed 2021/03/05.
18. CDC (Centers for Disease Control and Prevention), https://www.cdc.gov/flu/weekly/index.htm, last accessed 2021/03/05.
19. Merckx J, Wali R, Schiller I, Caya C, et. al., Diagnostic Accuracy of Novel and Traditional Rapid Tests for Influenza Infection Compared With Reverse Transcriptase Polymerase Chain Reaction: A Systematic Review and Meta-analysis, Ann Intern Med. 167(6), 394-409, (2017). DOI: 10.7326/M17-0848
20. Pai NP, Vadnais C, Denkinger C, Engel N, Pai M. : Point-of-Care Testing for Infectious Diseases: Diversity, Complexity, and Barriers in Low- And Middle-Income Countries. PLOS Medicine 9(9), (2012). DOI: 10.1371/journal.pmed.1001306
21. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J: Clinical Signs and Symptoms Predicting Influenza Infection. Arch Intern Med. 160(21), 3243–3247, (2000). DOI: 10.1001/archinte.160.21.3243
22. E. Marquez and V. Barrón: Artificial Intelligence system to support the clinical decision for influenza. In: IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), pp. 1-5. Ixtapa, Mexico (2019). DOI: 10.1109/ROPEC48299.2019.9057056
23. Arturo López Pineda, Ye Ye, Shyam Visweswaran, Gregory F. et al.: Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. Journal of Biomedical Informatics 58, 60-69, (2015). DOI: 10.1016/j.jbi.2015.08.019
24. M. A. Khan, W. Ul, M. A., S. H. Almotiri, S. Saqib et al., "Forecast the influenza pandemic using machine learning," Computers, Materials & Continua, vol. 66, no.1, pp. 331–340 (2021). DOI:10.32604/cmc.2020.012148
25. Larose, D.T., Larose, C.D.: Discovering Knowledge in Data: An Introduction to Data Mining. 2nd edn. Wiley, pp. 149–164 (2014). DOI: 10.1002/9781118874059.ch7
26. Ripley, B.: In Pattern Recognition and Neural Networks. Cambridge University Press.pp. 143-180, (1996). DOI: 10.1017/CBO9780511812651.006
27. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001). DOI: 10.1023/A:1010933404324
28. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY (2009). DOI: 10.1007/978-0-387-84858-7_15
29. Zhu, W., Zeng, N., Wang, N. .: Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical SASÒ implementations. In: NESUG Proceedings: Health Care and LifeSciences, Baltimore, Maryland (2010).
30. Jane V. Carter, Jianmin Pan, Shesh N. Rai, Susan Galandiuk.: ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves,Surgery 159(6), 1638-1645, (2016). DOI: 10.1016/j.surg.2015.12.029