INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 11, Issue 12, (2021) E-ISSN: 2222-6990

Prediction of Early Symptoms of COVID-19 Infected Patients Using Supervised Machine Learning Models

Zaidah Ibrahim, Norizan Mat Diah, Nor Azreen Rizal and Muhammad Naim Yuri

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor

Email: zaidah@tmsk.uitm.edu.my, norizan@tmsk.uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v11-i12/11991 DOI:10.6007/IJARBSS/v11-i12/11991

Published Date: 26 December 2021

Abstract

The Coronavirus disease 19 (COVID-19) is an ongoing global pandemic where it is easily transmittable and life threatening the world. The number of infected and non-survived patients is increasing in almost all the affected countries. Currently, there is no clinically approved vaccine available yet. Early prediction is necessary to assist the healthcare systems to strategize and reduce the spread of this virus. This is a very critical decision that is considered as a potential threat to others. Supervised Machine Learning (SML) models have demonstrated promising performance in various prediction applications that can improve decision making. Thus, this research investigates the capabilities of SML models to predict whether a patient is infected with COVID-19 or not based on certain symptoms. A comparative analysis of the impact of seven standard SML prediction models has been conducted. They are Adaboost, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Neural Network, Random Forest, and Support Vector Machine. A publicly available dataset from kaggle.com has been utilized for this research that consists of twenty symptoms collected from eight different countries. The outcome from Random Forest revealed that the five most important symptoms are tiredness, fever, dry cough, nasal congestion and those whose age is more than 60. These symptoms are consistent for all eight countries. Besides that, experimental results of the SML models also indicate that Neural Network achieves the best predictive results followed by Adaboost.

Keywords: Adaboost, COVID-19 Infected , K-Nearest Neighbor, Logistic Regression, Machine Learning, Naive Bayes, Neural Network, Random Forest, Support Vector Machine.

Introduction

Coronavirus disease 19 (COVID-19), an infectious pandemic, has shocked the world through its global wide-spread. It has affected the world economy, medical and public health infrastructure. As of 20th August 2020, more than 23,000,000 confirmed cases with more than 800,000 death cases have been reported involving 216 countries, including Malaysia (World Health Organization, 2020). Currently, since there is no approved vaccine for this virus yet, prevention is vital. Various clinical data about the patients have been collected that includes demographic, places that have been visited, and symptoms (Rajkumar, 2019). The availability of intelligent tools for the collection, storage and analysis of these data like Supervised Machine Learning (SML) models, prediction of the infected patients is possible. The early prediction results of these models are necessary to assist the healthcare systems to strategize and reduce the spread of this virus.

SML models, with vast amount of data, provides an effective way to automate analysis and diagnosis in healthcare (Wang and Summers, 2012). A high accuracy in such tasks consequently improves the efficiency of the healthcare decision making. Various SML models have been applied for predictive analysis in medical. A review on various SML models for prediction and classification that includes Support Vector Machine (SVM) and Random Forest (RF) with accuracy more than 95% have been reported where data applied for SVM is lung images while blood test is used for RF (Hanumanthu, 2020). Breast cancer risk prediction has been investigated using SVM, Decision Tree (DT), Naïve Bayes (NB) and K-Nearest Neighbour (K-NN) with SVM producing the highest accuracy (Asri *et. al.*, 2016). (Finkelstein and Jeong, 2017) compare SVM and NB for early prediction of asthma exacerbations with 80% accuracy achieved by SVM while 70% by NB. SVM and NB have also been compared for diabetes prediction and in this case, NB has higher accuracy compared to SVM (Sisodia and Sisodia, 2018). Decision Tree (DT), RF, Neural Network (NN) and Logistic Regression (LR) have been used to predict the risk of coronary heart disease (Beunza *et. al.*, 2019) and the best accuracy of 84% is produced by LR. For some other researches, for instance, prediction of potential druggable proteins indicate that NN performs better than SVM and RF with 89.98% accuracy (Jamali *et.al.*, 2016).

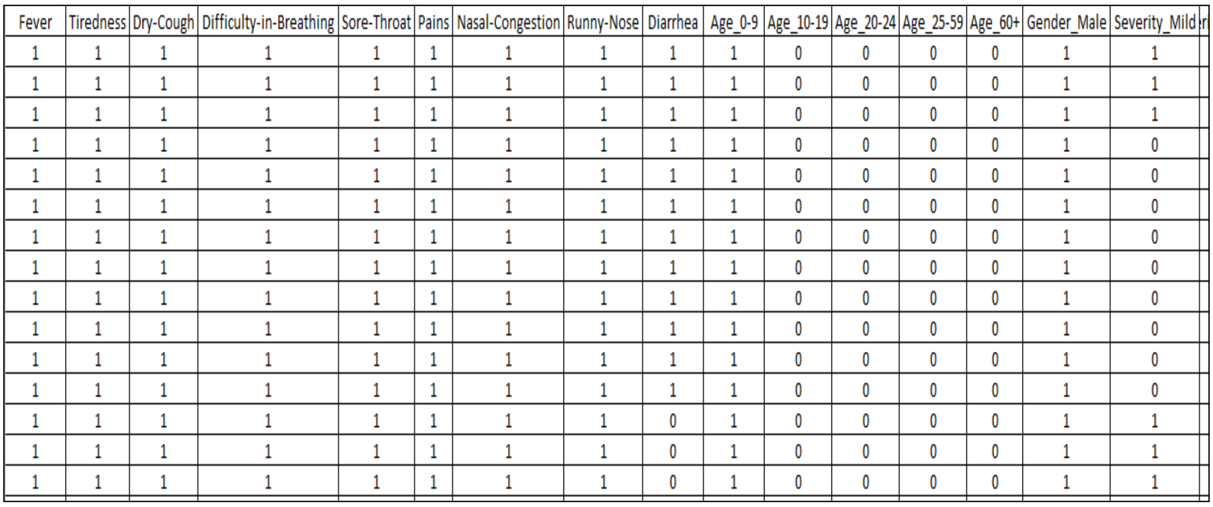
Quite a few other publications have reported good prediction results by SML models related to COVID-19. LR has shown to perform at the top of the list compared to RF, Adaboost (AB) and SVM for detecting COVID-19 (Khanday *et. al.*, 2020). Forecasting the number of new infected cases, the number of deaths and the number of recoveries from COVID-19 has been conducted using LR and SVM and the results show that LR forecasts better than SVM (Rustam *et. al.*, 2017). On the other hand, Adaboost performs better than SVM, RF and DT in identifying early stage symptoms of COVID-19 based on different age categories (Ahamad *et. al.*, 2020).

Besides applying SML models in medical areas, they have also been utilized in other areas such as banking systems (Nor *et. al.,* 2019), human resource (Ab Mutalib *et. al.,* 2017) and facilities (Shariff *et. al.*, 2018).

Based on the previous researches that have been accomplished, it seems that there are seven popular SML models being applied for predicting medical type of problem. Thus, the main objective of this research is to examine the performance of predicting early symptoms for COVID-19 using these seven SML models that are AB, K-NN, LR, NB, NN, RF, and SVM.

Data and Method

In this work, a publicly available dataset has been downloaded from Novel Corona Virus 2019 dataset that consists of twenty symptoms from eight countries that are China, France, Germany, Iran, Italy, Korea, Spain and UAE (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset). The symptoms are fever, tiredness, dry cough, difficulty in breathing, sore throat, pains, nasal congestion, runny nose, diarrhea, age between 0-9, age between 10-19, age between 20-24, age between 25-59, age more than 60, gender, severity of the illness either mild, moderate or severe, and either the patient has any contact with an infected person or not. The output is whether the patient is infected or not. Due to hardware limitations, we only extract 5000 data from each country where 2500 data that are from patients that have been infected with COVID-19 while another 2500 data are from patients that are not infected with COVID-19. Figure 1 shows some sample data used in this research.



Fever	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat	Pains	Nasal-Congestion	Runny-Nose	Diarrhea	Age_0-9	Age_10-19	Age_20-24	Age_25-59	Age_60+	Gender_Male	Severity_Mild
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1

Figure 1 Some sample data used in this research (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset)

The results of the experiments were based on the number of true positive (TP), true negative (TN), false negative (FN) and false positive (FP) where TP is the number of patients that are correctly predicted as infected. FP is the number of non-infected patient predicted as infected. TN is the number of non-infected people predicted and not infected. FN is the number of infected people incorrectly predicted as not infected. Then, the prediction performance of the SML models were evaluated based on the following criteria:

$$\text{Sensitivity} = TP / (TP + FN)$$
$$\text{Specificity} = TN / (TN + FP)$$
$$\text{Accuracy} == (TP + TN) / (TP + FP + TN + FN)$$

Sensitivity measures the proportion of patients that are infected and are correctly predicted to be infected while specificity measures the proportion of patients that are not infected and are correctly predicted as not infected. Accuracy computes the total number of correct predictions divided by the total number of predictions.

Adaboost (Adaptive Boosting), introduced by Freud and Shapire (1997), builds a stronger predictive model from the mistakes made by several weaker models. It starts by generating a predictive model from the training data. Then, a second model is generated from the previous model by reducing the errors made by the previous model. This process continues until the training data is predicted accurately.

K-Nearest Neighbor (KNN) is a simple non-parametric model that assumes that similar data exist in close proximity (Cover and Hart, 1967). It is a non-parametric model because it does not learn from the training dataset immediately but just store the data during the training phase, and at the time of prediction, it performs an action on the data. It assigns to an unpredicted data the prediction of the nearest of a set of previously predicted data.

Logistic Regression (LR) measures the relationship between the dependent variables (patient will be infected or not) and the independent variables (the twenty symptoms), by estimating probabilities using its underlying logistic function. Then, these probabilities are transformed into binary values for the predicted output using sigmoid function. Naïve Bayes (NB) is a SML model based on Baye's theorem where it computes the probability of an event based on the following steps:

i. Compute the prior probability for the given class labels or output;
ii. Find the likelihood probability with each input or symptoms for each class or output;
iii. Place these values in Bayes formula and compute the posterior probability;
iv. Examine which class has a higher probability, given the symptom belongs to the higher probability class.

Artificial Neural Network or sometimes called as Neural Network (NN) was inspired by the learning process of the biological human brain and the first NN was created by psychologist Frank Rosenblatt called perceptron (Kay, 2001). There are various types of NN and one of it is multi-layer perceptron. It consists of input, hidden and output layers and each layer has neurons. These neurons are interconnected within each layer and weights are assigned to each of these connections. Activation function is applied to determine the output and errors may be discovered at the output layers. Weights are adjusted and the process of applying the activation function is repeated until the convergence criteria are met. Figure 2 illustrates the structure of the 3-layer NN for this research where there are 20 neurons in the input layer and 1 neuron in the output layer. Backpropagation learning algorithm is applied for the training process.

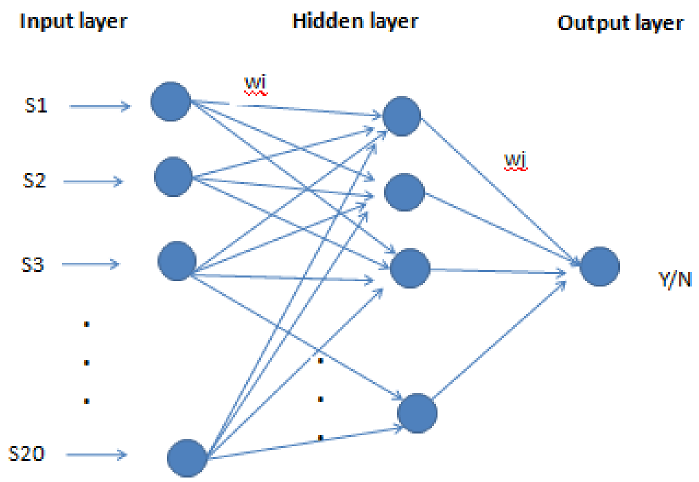


Figure 2 The structure of 3-layer NN for prediction.

Random Forest (RF) has been introduced by (Breiman, 2001). It consists of a collection of tree-structured classifiers {h(X, $Ø_n$), N= 1,2,3,…L} where X represents the input data while identical family and dependent distributed random vectors are denoted by {$Ø_n$}. In each decision split, the features are selected randomly and it reduces the correlation between trees which improves the power of prediction. In other words, the concept of RF is that the aggregate results of multiple predictors or decision trees provide better prediction compared to the best individual predictor. Figure 3 demonstrates how RF reaches its predicted output.

Support Vector Machine (SVM) was first developed by Vladimir Vapnik and his colleagues at AT&T Bell Laboratories for binary classification type of problem (Vapnik and Vapnik, 1998). For our dataset, SVM is very suitable since the predicted output is a binary decision which is either the patient will survive or not. Eq.1 shows how prediction decision is made using SVM. The prediction of a dataset is written as in Eq. (1).

$$f(x) = \sum_{i=1}^{l} \alpha_i \, y_i \, k\langle x_i . x\rangle + b \qquad (1)$$

The *y* values represent the infected (1) and not infected (0) of the prediction for training and testing. The *a* variable is a langrage multipliers obtained in the minimalization process of data *(x,y)*. The *l* variable, also known as support vectors is the decision borderline of bi-class of 1 and 0 values. This decision boundary defined as the hyper-plane is positioned in such a way that it is as far apart as possible from the nearest data points in each class. Such closest points are considered as support vectors. Kernel performs in non-linear mapping is defined using *k* variable. Figure 4 illustrates the distribution of data and the hyper-plane where *k* is the linear kernel function.

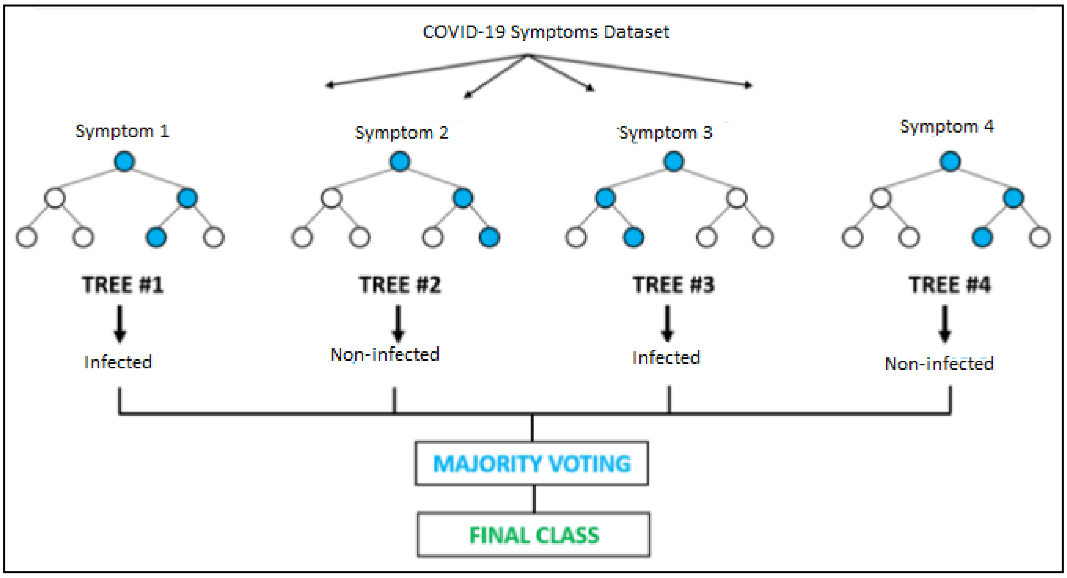


Figure 3 Sample process of RF.

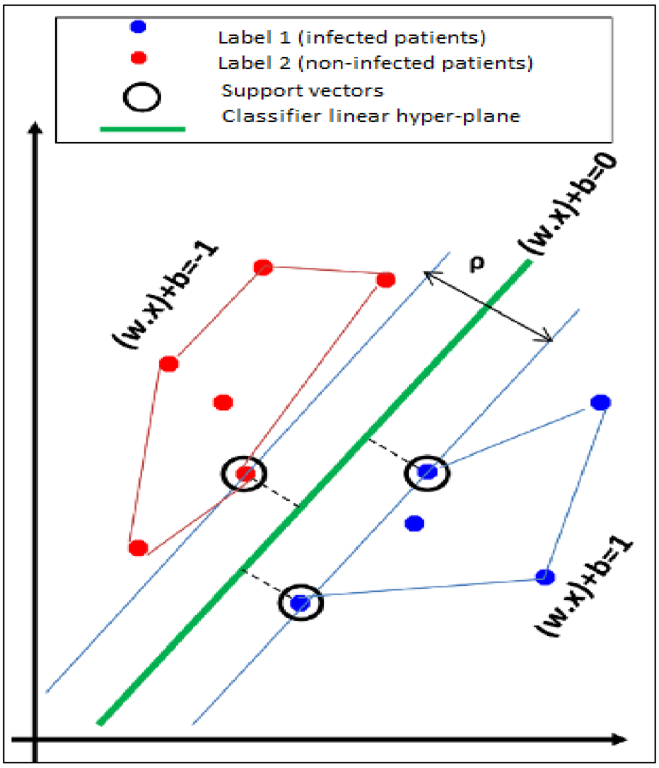

Label 1 (infected patients)
Label 2 (non-infected patients)
Support vectors
Classifier linear hyper-plane
(w.x)+b=-1
(w.x)+b=0
ρ
(w.x)+b=1

Figure 4 Example of support vectors and the hyper-plane in SVM.

Results and Discussion

Scikit-Learn tool has been used for the application of seven SML prediction models for COVID-19 infections in this research. From the twenty symptoms that have been recorded, RF has revealed three common early symptoms for all eight countries that are tiredness, fever and dry cough. Besides that, nasal congestion and age over 60 interchangeably become the fourth and fifth symptoms that can lead to this virus. Table 1 shows the feature importance score for all the symptoms in all the eight countries produced by RF. This outcome is similar with the results produced in (Khanday *et. al.*, 2020) that use other datasets. These experimental results confirm that anybody who has at least these five symptoms should go for further check-up or isolation to avoid the spread of this virus.

The experiments were conducted separately for each country to examine their similarities and differences in the prediction performance. The data is being divided into 70:30 ratio where 70% of the data is for training while the other 30% is for testing the SML models. 5-fold cross-validation was conducted for all seven SML models for each country separately. Then, an average is computed for each evaluation criteria. Table 2 illustrates the individual prediction performance (accuracy, sensitivity and specificity) for each country while Table 3 lists the comparative analysis of the average for each evaluation criteria for all seven SML models. By referring to Table 3, we can see that on the average, NN has proven to be the best SML model for early symptom prediction for COVID-19 since it produces 0.99 for all three evaluation criteria. The next second best model is AB since on average it produces about 0.97. Even though KNN achieves 1.0 performance for average specificity, the range of performance for average accuracy and sensitivity is more compared to NN. The same situation happens for NB and SVM where both of these models arrives at 1.0 for average sensitivity, the range of performance for the other two evaluation criteria is a bit high. RF seems to produce the lowest performance compared to the other models.

Conclusion

Currently, the spread of the infectious COVID-19 pandemic is a dangerous threat to global health. One option to control this spread is by predicting early symptoms of this virus since a vaccine is not available yet. SML models have demonstrated promising results to address this problem. Experimental analysis indicates that the significant symptoms are tiredness, fever, dry cough, nasal congestion and age that is more than 60. The predicting performance of the seven SML models are based on accuracy, specificity and sensitivity, and the results show that Neural Network and Adaboost seem to be the best two predictive models to predict the early symptoms for COVID-19 infectious patients. However, the size of the COVID-19 dataset was not extensive enough to provide sufficient results. Future work includes the use of larger datasets and the integration of x-ray images of the lungs for the prediction. It is hoped to acquire Malaysian data to perform similar predictive analysis.

Acknowledgements

The authors gratefully acknowledge the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, for sponsoring this research.

Table 1 Lists of Importance Scores for all Symptoms Related to COVID-19.

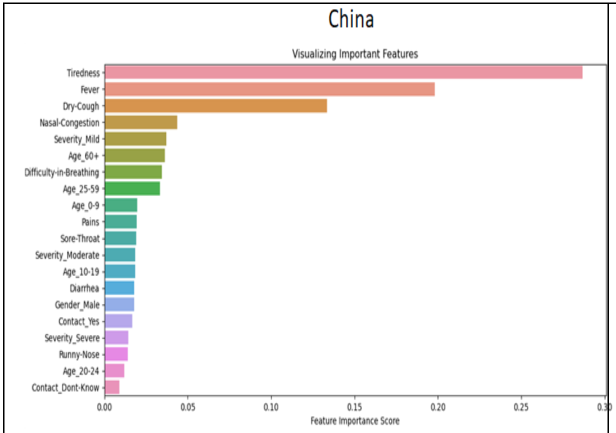
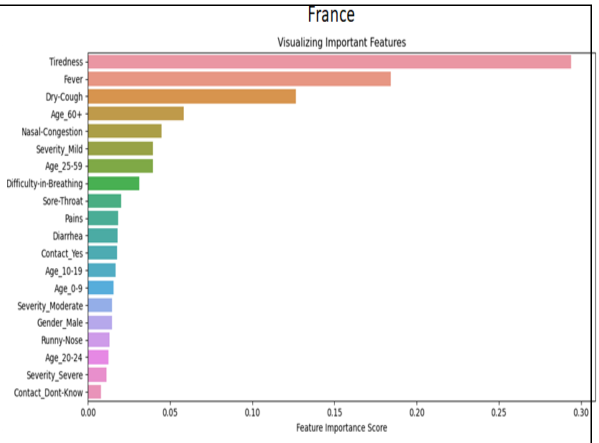
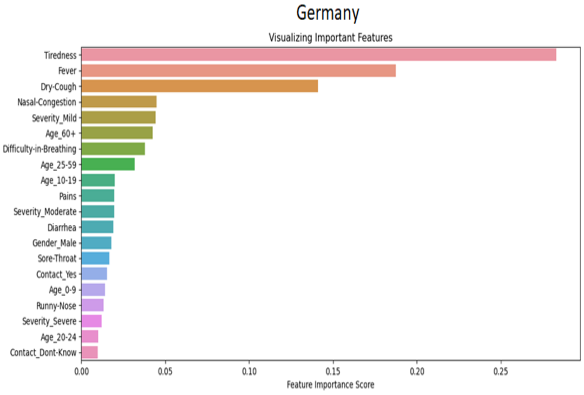
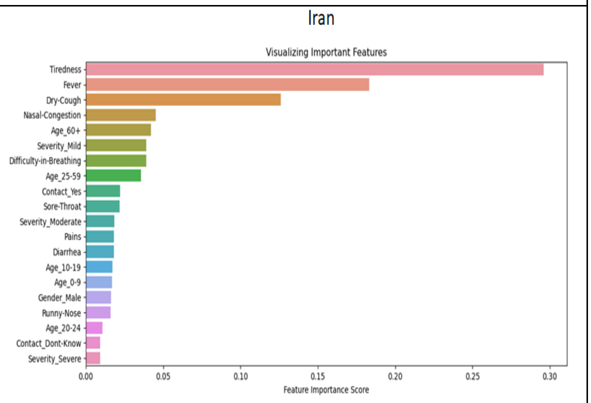
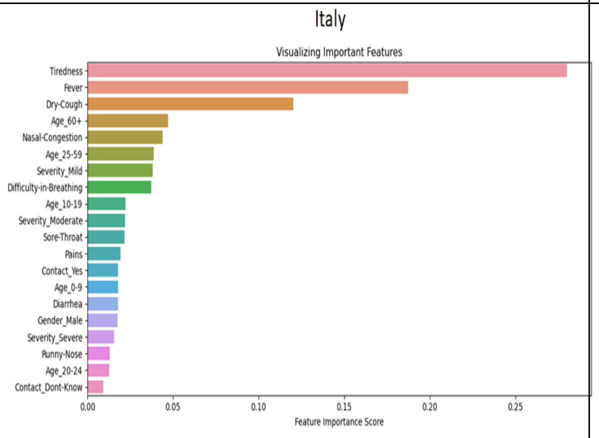
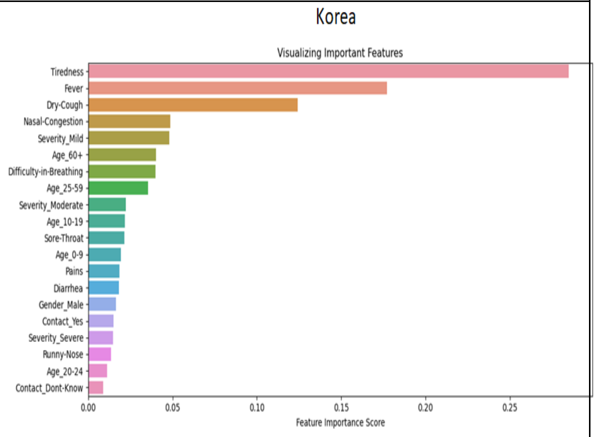
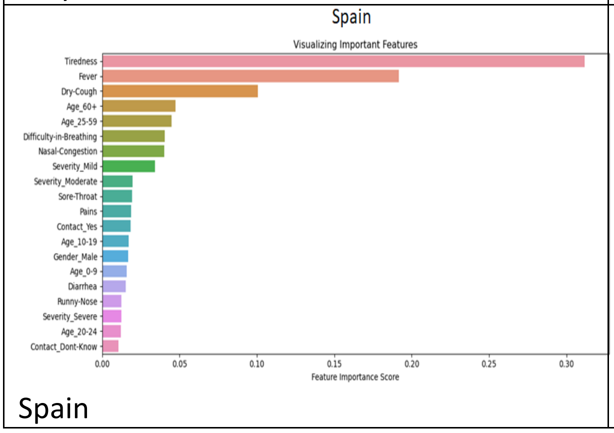
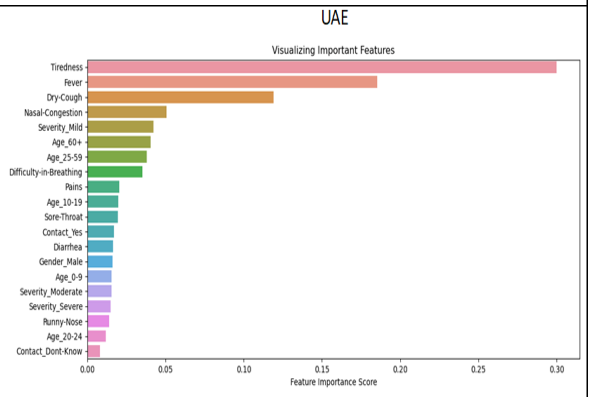
 China	 France
 Germany	 Iran
 Italy	 Korea
 Spain	 UAE

Table 2 Prediction Performance for Seven SML Models By Country.

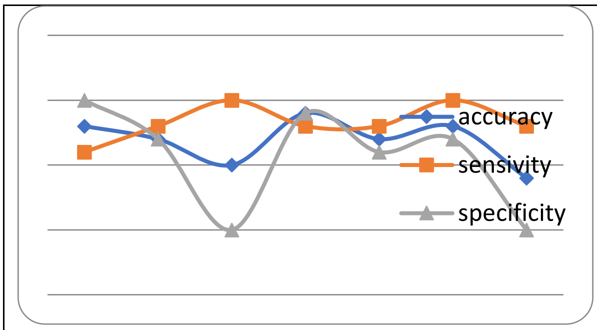
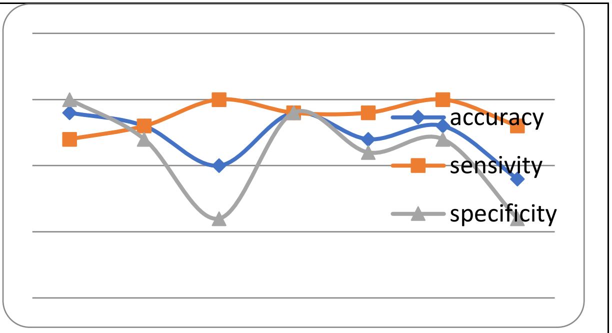
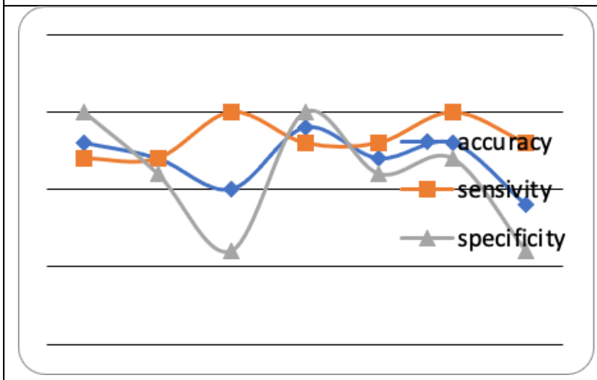
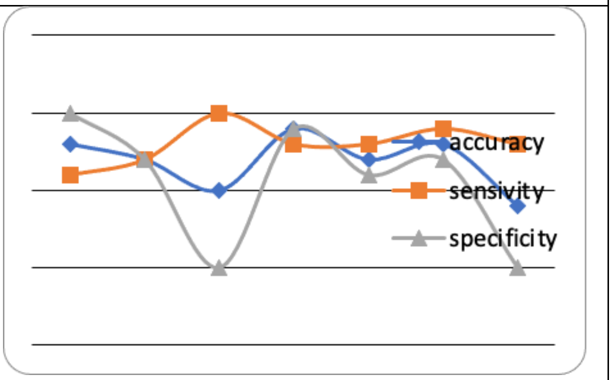
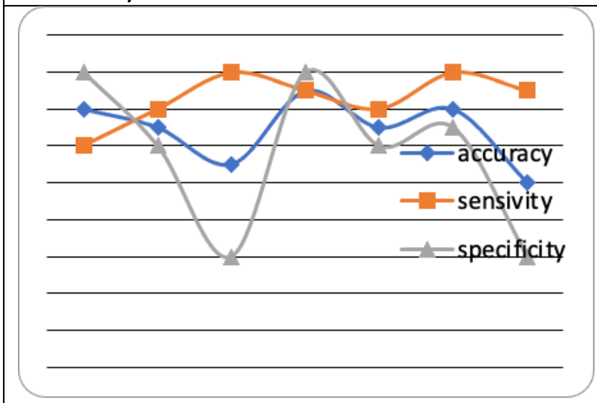
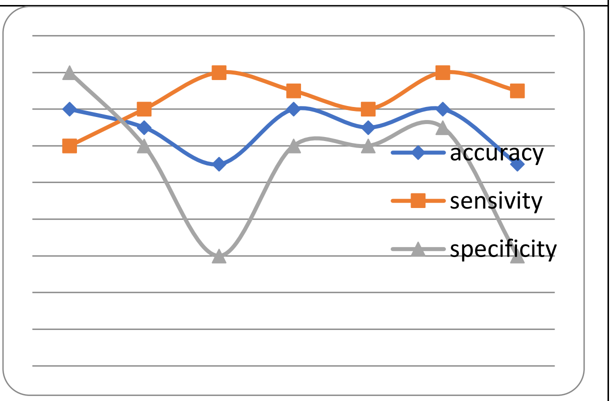
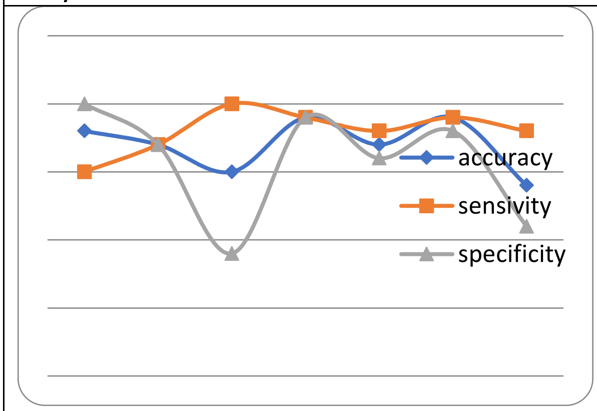
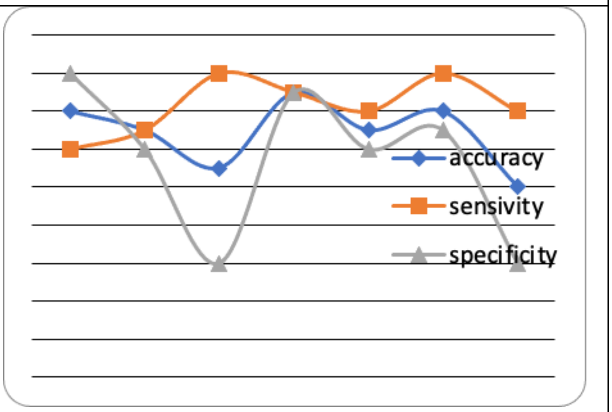
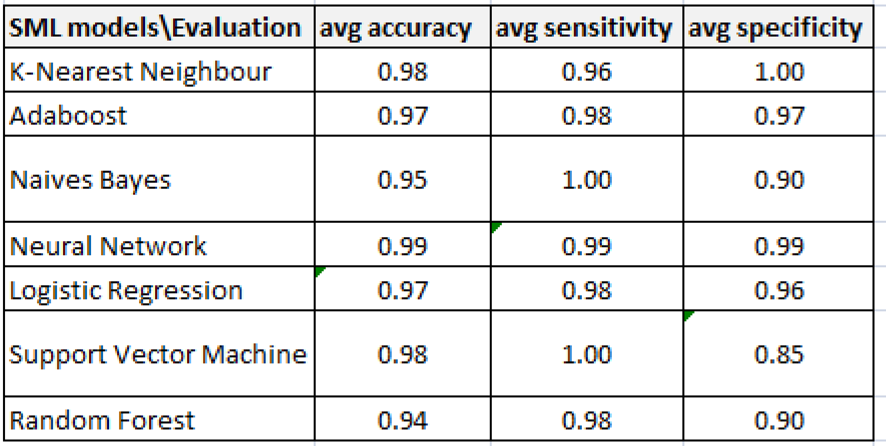
 China	 France
 Germany	 Iran
 Italy	 Korea
 Spain	 UAE

Table 3 Average Prediction Results of Seven SML Models.



SML models\Evaluation	avg accuracy	avg sensitivity	avg specificity
K-Nearest Neighbour	0.98	0.96	1.00
Adaboost	0.97	0.98	0.97
Naives Bayes	0.95	1.00	0.90
Neural Network	0.99	0.99	0.99
Logistic Regression	0.97	0.98	0.96
Support Vector Machine	0.98	1.00	0.85
Random Forest	0.94	0.98	0.90

References

Ab Mutalib, S. M., Ramli, N., & Mohamad, D. (2017). Forecasting Unemployment based on Fuzzy Time Series with Different Degree of Confidence. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(1-4), 21-24.

Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., ... & Moni, M. A. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert systems with applications, 160, 113661.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., ... & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). Journal of biomedical informatics, 97, 103257.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cover, T. M. (1967). P, F. Hart, Nearest neighbour pattern classification, I. EEE Trans.

Finkelstein, J., & cheol Jeong, I. (2017). Machine learning approaches to personalize early prediction of asthma exacerbations. Annals of the New York Academy of Sciences, 1387(1), 153.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Swapnarekha, H., Behera, H. S., Nayak, J., & Naik, B. (2020). Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. Chaos, Solitons & Fractals, 138, 109947.

Jamali, A. A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., & Ebrahimie, E. (2016). DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug discovery today, 21(5), 718-724.

Kay, A. (2001). Artificial Neural Network, Computerworld.

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. U. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 12(3), 731-739.

Rajkumar, S. (2019) Novel Corona Virus 2019 Dataset, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=COVID19_line_list_data.csv/

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. IEEE access, 8, 101489-101499.

Shariff, S. S. R., Suhaimi, M. A., Zahari, S. M., & Derasit, Z. (2018). Alternative Methods for Forecasting Variations in Hospital Bed Admission. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 410-416.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

Nor, S. H., Ismail, S., & Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model. Journal of Economics, Finance and Administrative Science, 24(47), 157-170.

Vapnik, V., & Vapnik, V. (1998). Statistical learning theory Wiley. New York, 1(624), 2.

Wang, S., & Summers, R. M. (2012). Machine learning and radiology. Medical image analysis, 16(5), 933-951.

World Health Organization. (2019) Coronavirus Disease (COVID-19) pandemic https://www.who.int/emergencies/diseases/novel-coronavirus-2019/