Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online)   Volume 33, Issue 1, 2022   DOI: http://doi.org/10.23851/mjs.v33i1.1076

Research Article **Open Access**

Covid-19 Prediction Model using Data Mining Algorithms

Zahraa Naser Shah weli

Department of Computer Science, Al-Nahrain University, Baghdad, IRAQ.

Contact: zah2005muh@gmail.com

ArticleInfo

Received
29/07/2021

Accepted
14/10/2021

Published
10/03/2022

ABSTRACT

From 2019 until today, the whole world is panicking due to the pandemic of the Corona virus or the so-called COVID-19, so it was inevitable to search for a way to predict the disease before it occurs. Disease forecasting requires huge databases, human hands, and high-speed technologies. Due to the rapid spread of this pandemic, scientists have looked at using data mining methods to predict diseases. Predicting and early diagnosis of disease through mining algorithms reduces human errors, saves money and scientists make the most accurate decision. It then shortens the long time needed to detect COVID-19 using available algorithms and tools that rely on data such as lung images or disease symptoms such as temperature, among others. In this research, two data mining algorithms were used to predict COVID-19 that support vector machine (SVM) and naïve bays (NB) using a dataset from the Korea Centers for Disease Control (KCDC). Feature selection was performed using Correlation based Feature Selection (CFS) after previously processed data. Performance measures for the research proved that SVM is best classifier of NB with accuracy, sensitivity and specificity of SVM being 96.72%, 94.08%, and 97.96% respectively. The receiver operating characteristic (ROC) curve also demonstrated better SVM performance than NB for predicting COVID-19.

KEYWORDS: COVID-19; Predicting; KCDC dataset; Correlation based Feature Selection (CFS); ROC curve; SVM; NB.

الخلاصة

منذ عام 2019 وحتى اليوم، يشعر العالم كله بالذعر بسبب جائحة فيروس كورونا أو ما يسمى بـ COVID-19، لذلك كان لا مفر من البحث عن طريقة للتنبؤ بالمرض قبل حدوثه. يتطلب التنبؤ بالأمراض قواعد بيانات ضخمة، وأيدي بشرية، وتقنيات عالية السرعة. لكن الانتشار السريع للمرض ، الجأ العلماء إلى التفكير باستخدام أساليب التنقيب عن البيانات للتنبؤ بالمرض. إن التنبؤ والتشخيص المبكر للمرض من خلال خوارزميات التنقيب يقلل من الأخطاء البشرية ويوفر المال ويتخذ العلماء القرار الأكثر دقة. ثم يقصر الوقت الطويل اللازم لاكتشاف COVID-19 باستخدام الخوارزميات والأدوات المتاحة التي تعتمد على البيانات مثل صور الرئة أو أعراض المرض مثل درجة الحرارة وغيرها. في هذا البحث، تم استخدام خوارزميتين لاستخراج البيانات للتنبؤ بـ COVID-19 التي تدعم آلة ناقلات (SVM) وخلجان ساذجة (NB) باستخدام مجموعة بيانات من المراكز الكورية للسيطرة على الأمراض (KCDC). تم إجراء اختيار الميزة باستخدام تحديد الميزة المستند إلى الارتباط (CFS) بعد معالجة البيانات مسبقًا. أثبتت مقاييس الأداء للبحث أن SVM هو أفضل تصنيف من الـ NB مع دقة وحساسية وخصوصية لـ SVM بلغت 96.72٪ و 94.08٪ و 97.96٪ على التوالي. أظهر منحنى خاصية تشغيل المستقبل (ROC) أيضًا ان أداء SVM أفضل من NB للتنبؤ بـ COVID-19.

INTRODUCTION

COVID-19 is a global pandemic that emerged at the end of 2019 in the Chinese city of Wuhan as alleged and spread to the whole world and cause of many deaths [1]. The rapid spread of this virus caused a state of astonishment for scientists, and they were unable to find ways to stop it or even treat it.

Scientists have not yet found a powerful drug for COVID-19, but the vast rapid spread of this epidemic has provided huge data, some statistical and some clinical. This huge data is employed for early pandemic prediction [2]. Predicting the disease based on its symptoms and causes is the best way to reduce the number of deaths. Prediction is one of the main ways to treat diseases by providing human and clinical equipment [3]. To keep pace with the spread of the disease, scientists had to use tools and algorithms with high speed and accuracy, such as data mining algorithms.

Data mining is a technique of extracting new, hidden and effectual patterns or knowledge from data, which also used for the diagnosis and prognosis for multiple diseases. Which implied many types of algorithms such as neural network (NN), support vector machine (SVM), naïve bays

(NB) and others [4]. To reach high prediction accuracy when using data mining algorithms, it is necessary to preprocess and configure the data and get rid of unnecessary attributes that reduce the accuracy of the results, so a Correlation based Feature Selection algorithm was used in this research. The performance measures employed are sensitivity, specificity and ROC curve due to their importance in medical research as well as accuracy. This suggested work was split into four stages: the data collection, the pre-processing, feature selection and the prediction stage. The prediction stage was divided into two levels: prediction (application) level and evaluation of results level.

In Section 2, former researches are presented. Section 3 and Section 4, show the algorithms and equipment's used in search. The experimental outcomes of proposed model are clarified in section 5. At last, the conclusion is demonstrated in Section 6.

RELATED WORK

Rachid Zagrouba, Muhammad Adnan Khan [5] proposed a predictive model for COVID-19 outbreaking using support vector machine. The dataset that was collected from WHO and divided into two sections, 80% are employed for training and 20% are employed for validation. To assess the suggested model, multiple measures are utilized. The accuracy for predictive model shows 98.88% and 96.79% for training and validation successively.

Shawni Duttaa, Samir Kumar Bandyopadhyay [6] used deep learning neural network such as Long short-term memory (LSTM) and Gated Recurrent Unit (GRU) for classify COVID -19 cases. Kaggle dataset from 20th January 2020 to 12th march 2020 are intended for model. The accuracy are acquired from the suggested model is 40.5%, 62%, 67.8 and 87% for released, deceased, negative and confirmed cases where another essential parameter RMSE exhibits 13.72%, 4.16%, 49.4% and 30.15 for released, deceased, negative and confirmed respectively.

Ekta Gambhir, Ritika Jain and Alankrit Gupta [7] Proposed model analyzed COVID-19 spread in the world and attempt to outbreak it in India, using support vector machine algorithm and regression algorithm. The dataset used available in 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Regression algorithm shows accuracy reach 93%.

L. J. Muhammad, Md. Milon Islam, Sani Sharif Usman and Safial Islam Ayon [8] Use various data mining algorithms such as NB, SVM, decision tree, logistic regression, K-nearest neighbor, random forest. The model predicts the age of patients with COVID-19 who are more dangerous and the number of days required to recover from virus. Decision tree perform best results in comparison with other algorithm where the accuracy for it is 99.85. Dataset (COVID-19 patients of South Korea) used in this work.

Performance Measure Metrics

Accuracy (Acc.): It is the ratio of accurately classified cases to total cases [9]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Sensitivity (Sn.): It is the quotient of accurately categorized positives to the total number of positives [10]:

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \qquad (2)$$

Specificity (Sp.): It is the quotient of accurately categorized negatives to the total number of negatives [11]:

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad (3)$$

where:
TP= true positive cases. TN= true negative cases.
FN= false negative cases. FP= false positive cases.

Receiver Operator Characteristic (ROC)

Receiver Operator Characteristic (ROC) is a graphical chart with two axes, the y-axis manifest sensitivity and the x-axis display (1 – specificity). It is supposed to select the best classifier. The curve draws high values of the sensitivity and low values of (1– specificity). Better decision of classifier states when the point moving toward the upper left corner in the ROC. ROC curve is vital in medical applications where negative states are as important as positive states [12].

MATERIALS AND METHODS

Proposed methodology is as follow: the dataset was preprocessed and selected the best features for prediction using CFS feature selection algorithm, then divided dataset into three partitions, two random partitions for learning set and third one for testing set. In the learning set, 10 fold cross validations suggested for training and validation. Two random partitions used by prediction algorithms (SVM, NB) in application layer. Evaluation layer evaluates the work using accuracy, sensitivity, specificity and ROC curve depending on third partition. Figure 1 depicts proposed model architecture.

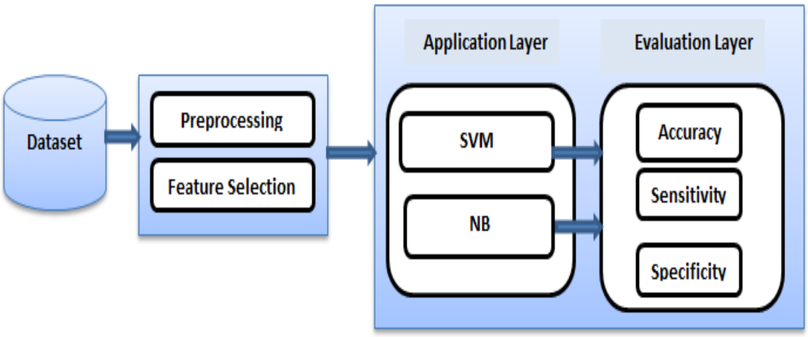

Figure 1. Proposed Model Architecture.

Dataset

The dataset was implied from Coronavirus dataset of Korea Centers for Disease Control (KCDC) which available on Kaggle Website [13]. The dataset has 305689 instances with 8 features. Table 1 illustrates the attributes of dataset.

Table 1. Description of Dataset.

No.	Feature	Type	Description
1	case id	Numeric	the ID of the infection case
2	province	String	Special City/ Metropolitan City/ Province
3	city	String	City / Country / District
4	group	Boolean	TRUE: group infection / FALSE: not group
5	infection case	String	the infection case (the name of group or other cases)
6	confirmed	Numeric	the accumulated number of the confirmed
7	latitude	Numeric	the latitude of the group
8	longitude	Numeric	the longitude of the group

The dataset was prepared, and cleaned from noise data, duplicate data and transform Boolean feature into numerical form. The non-numeric features converted into numeric form using python code, and then normalize all features to harmonize it.

Feature selection (FS)

Feature selection in machine learning points to selecting the best features in our data to support for our model. There are various types of FS, in this work the correlation-based feature selection (CFS) was employed.

CFS is one of filter feature selection algorithms that employ heuristic evaluation function as a feature subset. The evaluation function is biased against features that are closely related to the class and uncorrelated with each other. Features with low correlation with class should be ignored. Excessive features will be excluded and ignored because they correlate with remaining features. Feature selection relied on how predictable the class is that other features can't [14]. Feature subset evaluation function of CFS is:

$$Ms = \frac{K\overline{rcf}}{\sqrt{K + K(K-1)\overline{rff}}} \tag{4}$$

MS: is the heuristic "merit" of a feature subset S containing k features.

$\overline{rcf}$: is the mean feature-class correlation ($f \in S$).

$\overline{rff}$: is the average feature- feature intercorrelation. This equation shows a set of features which predictive the class [15].

After investing the CFS, the dataset contains 5 features with the same number of instances and the other 3 features (case-id, city, and group) are omitted.

Classification algorithms

The purpose of the suggested work is to distinguish the features that lead to the prediction of COVID-19 from others. In the following, illustrate each algorithm used briefly.

– *Naïve Bayes (NB)*

NB is a probabilistic algorithm employed for clustering and classification and uses Bayes theorem for classification and prediction tasks. It forecasts the class values considering feature sets. NB algorithm trained effectually in a supervised learning setting counting on the probability model [16].

Bayes' theorem:

$$p\left(\frac{x}{y}\right) = \frac{p(x,y)}{p(x)} = p(\frac{x}{y})p(y)/p(x) \qquad (5)$$

NB uses Bayes' classifier to reduce misclassification through the equation below:

$$p(wj,x) = \frac{p(wj,x)}{p(x)} = p\left(\frac{x}{wj}\right)p(wj)/\sum_{j=1}^{c} p(\frac{x}{wj})\,p(wj) \qquad (6)$$

For many applications, maximum likelihood is the approach used by NB to estimate parameter. NB classifier requires small instances of training data to estimate the means and variances of the variables. Just only the variances of the variables for each class need to be specified and not the entire covariance matrix, because only independent variables are intended [17].

– *Support Vector Machine (SVM)*

SVM is a supervised learning algorithms used for classification and regression. The SVM is demonstrating a system that will foresee target values; this is the basic goal of it. Another goal is to separate different classes by a decision boundary through maximize the margin between different classes [18].Decision boundary relied on parameters like kernel, gamma, C, degree etc.

SVM has special advantages when solving problems that have small, missing and highly correlated inputs [19]. Figure 2 summarizes liner SVM work.

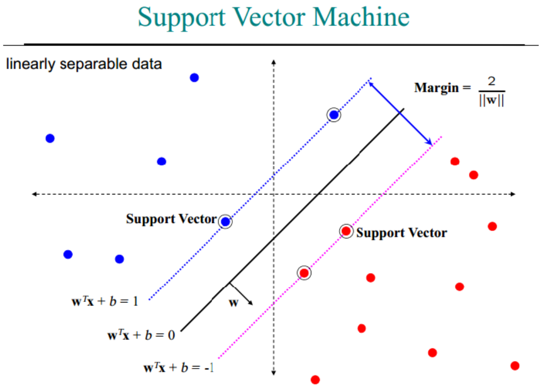

Figure 2. Linear SVM [20].

RESULTS AND DISCUSSION

Assumed methodology is as follow: dividing dataset into three partitions, two random partitions for learning set and third one for testing set. In the learning set, 10 fold cross validations invested for training and validation. Then evaluated the work through using performance measures such as accuracy, sensitivity, specificity and ROC curve. The proposed work was established on a system having 4 GB RAM and 1.8 GHz Intel Cori5- 3317 processor, 64-bit operating system and Python 3.8.5 programming language.

Table 2 explains the results for both classification algorithms, then figure 3 represent Roc curve for both algorithms.

Table 2. Results for Algorithms Used.

Classifier Model	accuracy	sensitivity	specificity
NB	93.81%	92.63%	96.54%
SVM	96.72%	94.08%	97.96%

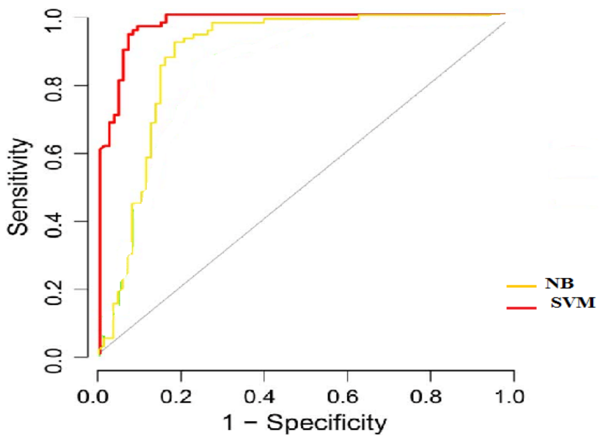

Figure 3. ROC curve for algorithms used.

It seems that SVM gave the best classification results from NB, which is also illustrated by the ROC curve.

In table 3 a comparison was made between this work and one of the earlier research work No. [21] To explain the contribution:

Table 3. proposed work compares with previous work

	Dataset	algorithms	results

Proposed work	KCDC (8 attributes, 305689 instances)	SVM	Sn=94.72
			Acc=96.72

		NB	Sn=92.63
			Acc=93.81

Research [21]	VAERS (14 attributes, 6745 instances)	NB	Sn=69.7
			Acc=69.74

		KNN (K-Nearest Neighbo)	Sn=81.3
			Acc=81.6

		DT
		(Random Tree)
			Sn=100
			Acc=99.98

		RT
		(Random Tree)
			Sn=96.2
			Acc=96.17

CONCLUSION

Corona virus is not like other diseases, it has spread broadly very fast and caused the death of many people around the world, and it is in a continuous development and spread. Governments put all money and efforts to reduce COVID- 19 and find a cure, but to no avail, so early predicting it has become the best way to control it. In this work two prediction algorithms were employed, SVM and NB. Where the accuracy for SVM is 96.72% while 93.81% for NB. The dataset was preprocessed and select the best features by Correlation based Feature Selection. To train algorithms, 10 fold cross validation was used.

REFERENCES

[1] Omran S. H.; Abbas K. A.; Alaa H. A.; Nihad Q. M. "Epidemiological Features of COVID-19 Epidemic in Basrah Province-Southern Iraq-First Report ", The Medical Journal of Basrah University, 38, 6-17, 2020.
http://doi.org/10.33762/mjbu.2020.126943.1008

[2] Mateus M. ; Jonatha S. P.; Ivalbert S. P.; João G.; Marcos E. B.; Anderson A., "Convolutional Support Vector Models: Prediction of Coronavirus Disease Using Chest X-rays", Information, 11, 2020.
http://doi.org/10.3390/info11120548

[3] N.; Adel G.; Mohammad S., "COVID-19 Prediction Classifier Model Using Hybrid Algorithms in Data Mining", Int J Pediatr, 9, 12723-12737, 2021.
http://doi.org/10.22038/ijp.2020.54272.4290

[4] Zahraa N. S., "Data Mining in Cancer Diagnosis and Prediction:Review about Latest Ten Years", Current Journal of Applied Science and Technology, 39, 11-32, 2020.
http://doi.org/10.9734/CJAST/2020/v39i630555

[5] Rachid Z.; Muhammad A. K.; Atta-ur-R.; Muhammad A. S.; Muhammad F. M.; Abdur R.; Muhammad F. K., "Modelling and Simulation of COVID-19 Outbreak Prediction Using Supervised Machine Learning", Computers, Materials & Continua, 66, 2397-2407, 2021.
http://doi.org/10.32604/cmc.2021.014042

[6] Shawni D.; Samir K. B., "Machine learning approach for confirmation of COVID-19 cases: positive, negative, death and release", Iberoamerican Journal of Medicine, 3, 172-177, 2020.
http://doi.org/10.5281/zenodo.3822623

[7] Ekta G.; Ritika J.; Alankrit G., "Regression Analysis of COVID-19 using Machine Learning Algorithms", Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020), IEEE Xplore Part Number: CFP20V90-ART, 2021.

[8] Muhammad L. J.; Md. M. I.; Sani S. U.; Safial I. A., "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery", SN Computer Science, 1, 2020.
http://doi.org/10.1007/s42979-020-00216-w

[9] Emmanuel de-G. J. O.; Wen Z.; Nancy Z.; Ning W., "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", Northeast SAS User Group proceedings, Section of Health Care and Life Sciences, Baltimore, Maryland, 14-17, 2010.

[10] Angelina S.; Amponsah K. S.; Abaidoo R., "Sensitivity and Specificity Analysis Relation to Statistical Hypothesis Testing and Its Errors: Application to Cryptosporidium Detection Techniques", Open Journal of Applied Sciences, 6, 209-216, 2016.
http://doi.org/10.4236/ojapps.2016.64022

[11] Kavitha K.; Catherine R. D.; Anuradha D., "Testing the Sensitivity and Specificity of ICU Patients and Diagnose Statistics Hypothetical Errors", International Journal of Innovative and Exploring Engineering (IJITEE), 8, 1042-1045, 2019.
http://doi.org/10.35940/ijitee.J9179.0881019

[12] Jan Y. V.; Ewout W. S.; Hajime U.; Bavo De C.; Laure W.; Gary S. C.; Ben V. C., "ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES", Journal of Clinical Epidemiology, 126, 207-216, 2020.
http://doi.org/10.1016/j.jclinepi.2020.01.028

[13] Coronavirus dataset of Korea Centers for Disease Control & Prevention (KCDC), https: //www.kaggl e.com/kimji hoo/coron avirusdata set/data. Accessed 20 Apr 2020.

[14] Agnieszka W.; Danuta Z., "Integrating Correlation-Based Feature Selection and for Improved Cardiovascular Disease Diagnosis", Complexity, 2018, 2018.
http://doi.org/10.1155/2018/2520706

[15] Rakkrit D.; Terry W., "Correlation-Based and Causal Feature Selection Analysis for Ensemble Classifiers", IAPR Workshop on Artificial Neural Networks in Pattern Recognition, 25-36, 2010.
http://doi.org/10.1007/978-3-642-12159-3_3

[16] Daniel B., "Bayes' Theorem and Naive Bayes Classifier", Encyclopedia of Bioinformatics and Computational Biology, 1, Elsevier, 403-412, 2018.
http://doi.org/10.1016/B978-0-12-809633-8.20473-1

[17] Michael C., "The Naive Bayes Model, Maximum-Likelihood Estimation, and the EM Algorithm", 2012.

[18] Jakub H.; Jaromir V.; Petr S., “Support Vector Machine Methods and Artificial Neural Networks Used for the Development of Bankruptcy Prediction Models and their Comparison”, J. Risk Financial Manag., 13, 2020.

http://doi.org/10.3390/jrfm13030060

[19] Yogita B. B.; Kalyani C.W., “Intrusion Detection System Using Data Mining Technique: Support Vector Machine”, International Journal of Emerging Technology and Advanced Engineering, 3, 581-586, 2013.

[20] Pijush S.; Sanjiban S. R.; Valentina B., “Handbook of Neural Computation”, Academic Press, 2017.

[21] Abdulkareem N., M.; bdulazeez A. M.; zeebaree D. Q.; Hasan D. A., “COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms”, Qubahan Academic Journal, 1, 100-105, 2021.

http://doi.org/10.48161/qaj.v1n2a53

How to Cite

Shah weli, Z. N. **Covid-19 Prediction Model Using Data Mining Algorithms**. *Al-Mustansiriyah Journal of Science*, *33*(1), 45–50, 2022. Doi: http://doi.org/10.23851/mjs.v33i1.1076