2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-3368-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOSEC51865.2021.9591801

Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) – A Comparative Analysis

[1]L. William Mary, [2]S.Albert Antony Raj

[1]Research Scholar, Department of Computer Applications, SRM Institute of Science and Technology, Kattankulathur, Chennai, India. lwilliamary@gmail.com
[2]Associate Professor & Head, Department of Computer Applications, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

***Abstract*--The new public health crisis threatens the entire world. The infectious virus SARS-CoV-2 (COVID-19) is a new virus that has spread rapidly across the world. Coronavirus spreads faster than the early virus, SARS-CoV and MERS-CoV, the first major beta-coronavirus in the human respiratory system. There are approximately 9,033,508 reported cases worldwide. CT scans have been used to diagnose suspicious negative and positive cases. To detect infected cases and mortality rates, many epidemiological models are being used. Therefore, Machine Learning techniques play a vital role in effective prediction. This technology has been used to extract information on large data sets and predictive performance analysis. As a result, a variety of Machine Learning prediction techniques were employed. This paper aims at determining which Classification method performs a high accuracy rate for the collected data samples of COVID-19 positive cases. The Support Vector Machine (SVM) gives 85% of accuracy, K-Nearest Neighbor gives 80% accuracy, and Naive Bayes (NB) text classifier method gives 65% accuracy.**

***Keywords*--COVID-19 Prediction, Feature Selection Techniques, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes.**

I. INTRODUCTION

The novel coronavirus (COVID-19 or 2019-nCoV) was discovered in Wuhan, Hubei Province, China. Bats were responsible for the initial spread, which was later spread to humans by the raccoon dog and palm civet [1], [2], [3]. The common COVID-19 symptoms are fever, dry cough, and tiredness. The major symptoms are difficulty in breathing, chest pain, and loss of speech. The Percentage of COVID-19 with fewer symptoms is shown in Table 1below[4].

TABLE 1 THE PERCENTAGE OF FEW SYMPTOMS

Signs and Symptoms	Total Percentage (%)
Dry Cough	60.4
Breathing problems or shortness of breath	41.1
Fever	55.5
Muscle ache	44.6
Headache	42.6
Sore throat	31.2
Disturbance of smell and taste	64.4
Fatigue	68.3

The author proposed a machine learning system that uses CT scan slices to detect cases of lung infection using a series of methods (such as multiple thresholds, threshold filters, feature extraction, selection, fusion, and classification). Chaotic-Bat-Algorithm and Kapur's Entropy thresholding use the CTS enhancement tool. The classifier methods Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine with linear kernel were used to achieve the highest accuracy (5]. The author has discussed the standard Naive Bayes model for uncertain data classification. The kernel density estimation method is extended to deal with unknown data. The value of each data item is represented by a probability distribution function [6].

For text classification, the Naive Bayes classifier method is commonly used. The feature selection system is used for text classification, and an auxiliary feature method is used for text subspace classifications [7]. The model was developed using a correlation coefficient analysis of dependent and independent features. 80% of the training set and 20% of testing have been considered[8]. The most accurate model is the regression and classifier model [9]. For classification, the author contributed the K-Nearest Neighbor method and the Hybrid Feature

Selection Method (HFSM), which combines the filter and wrapper approaches[10].

II. RESEARCH METHODOLOGY

This section explained the dataset used for these predictions, considered positive cases in the comparison.

Dataset collections

This section describes the datasets used in this study. To predict a SARS-CoV-2 obtained the data sets from the common survey. The dataset has 350 instances with multiple attributes, such as infected and non-infected males and females from various countries and cities. Experimentally validated to construct the training and testing datasets for this study and listed the various forms of data visualization. Figure 1 displays the collected data in a sequence.

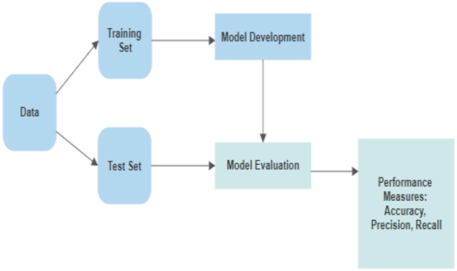


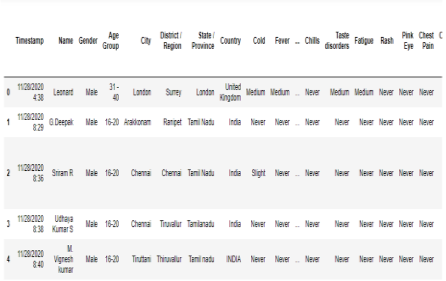
Fig. 1. Model creation and evaluation in machine learning

Data Preprocessing

Without preprocessing, raw data cannot be sent for model development. Data preprocessing is important before sending through a data model. Preprocessing involves the following steps,

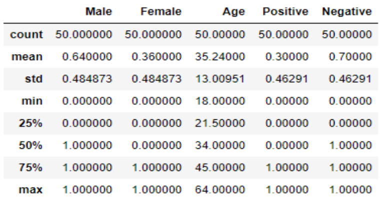
Import libraries: Pandas, Numpy are the major libraries are used to import data. Data manipulation and data analysis have done using pandas. Numpy is used for computation. Matplotlib and Seaborn used for visualization.

Read data: Collected raw data are available in the the.CSV file, which can read using pandas pd_read_csv.



	Timestamp	Name	Gender	Age Group	City	District / Region	State / Province	Country	Cold	Fever	...	Chills	Taste disorders	Fatigue	Rash	Pink Eye	Chest Pain	C
0	11/28/2020 4:38	Leonard	Male	31 - 40	London	Surrey	London	United Kingdom	Medium	Medium	...	Never	Medium	Medium	Never	Never	Never	
1	11/28/2020 8:29	G.Deepak	Male	16-20	Arakkonam	Ranipet	Tamil Nadu	India	Never	Never	...	Never	Never	Never	Never	Never	Never	
2	11/28/2020 8:36	Sriram R	Male	16-20	Chennai	Chennai	Tamil Nadu	India	Slight	Never	...	Never	Never	Never	Never	Never	Never	
3	11/28/2020 8:38	Udhaya Kumar S	Male	16-20	Chennai	Tiruvallur	Tamilanadu	India	Never	Never	...	Never	Never	Never	Never	Never	Never	
4	11/28/2020 8:40	M. Vignesh kumar	Male	16-20	Tiruttani	Thiruvallur	Tamil nadu	INDIA	Never	Never	...	Never	Never	Never	Never	Never	Never	

Fig.2. COVID-19 Data samples



	Male	Female	Age	Positive	Negative
count	50.000000	50.000000	50.00000	50.00000	50.00000
mean	0.640000	0.360000	35.24000	0.30000	0.70000
std	0.484873	0.484873	13.00951	0.46291	0.46291
min	0.000000	0.000000	18.00000	0.00000	0.00000
25%	0.000000	0.000000	21.50000	0.00000	0.00000
50%	1.000000	0.000000	34.00000	0.00000	1.00000
75%	1.000000	1.000000	45.00000	1.00000	1.00000
max	1.000000	1.000000	64.00000	1.00000	1.00000

Fig.3. Mean and Standard Deviation Variant

Checking for missing values: Dataset labelled as '0' and '1' to find the missing values.

Checking for categorical and variable data: Identified the target variable in the classification. Seaborn display () feature distribution techniques have been applied.

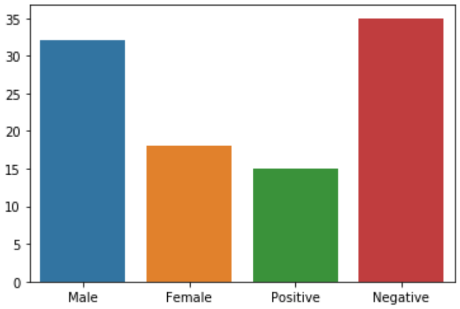


Fig.4. Distribution of Gender, Positive and Negative Cases

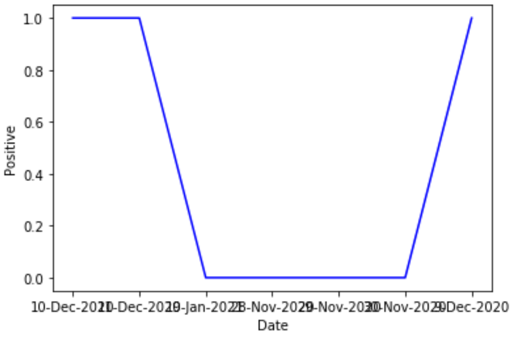


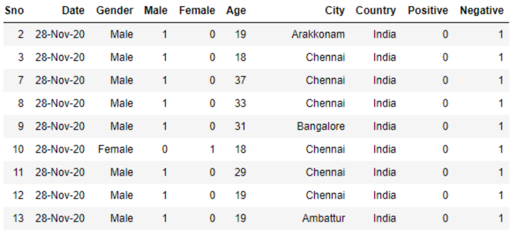
Fig.5. Distribution of Date Wise Positive Cases

Standardize the data: The dataset contains numerical input variables for analysis.

Data splitting: To obtain the expected accuracy in the model, three classification algorithms, SVM, KNN and NB, are used.

Machine Learning Classification Algorithms

During this study, we have investigated three different prediction algorithms for supervised classification.



Sno	Date	Gender	Male	Female	Age	City	Country	Positive	Negative
2	28-Nov-20	Male	1	0	19	Arakkonam	India	0	1
3	28-Nov-20	Male	1	0	18	Chennai	India	0	1
7	28-Nov-20	Male	1	0	37	Chennai	India	0	1
8	28-Nov-20	Male	1	0	33	Chennai	India	0	1
9	28-Nov-20	Male	1	0	31	Bangalore	India	0	1
10	28-Nov-20	Female	0	1	18	Chennai	India	0	1
11	28-Nov-20	Male	1	0	29	Chennai	India	0	1
12	28-Nov-20	Male	1	0	19	Chennai	India	0	1
13	28-Nov-20	Male	1	0	19	Ambattur	India	0	1

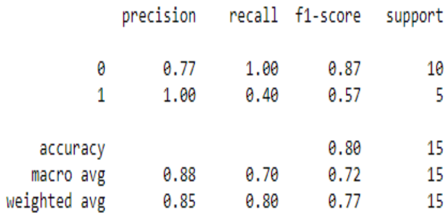
Fig.6. Features selected for the classification

Support Vector Machine Algorithm (SVM)

Vapnik was the one who came up with the concept of SVM [11]. It is a supervised learning method for classifying and predicting data. In the classification task, training and testing are involved in the SVM with the data. Each instance contains the target values in training [12]. SVM is divided into three parts. i) Model creation for classification or regression problems. ii) Application of optimization techniques to new models and paradigms. iii) Formulation and solution of optimization problems [13].

Support Vector, Hyperplane, and Margin

The closest vector point of the hyperplane is the support vector points. Applying a support vector machine is to identify the best line in two dimensions or the best hyperplane over one dimension. The hyperplane has been found through the maximum margin set. Margin is the distance of the vectors from the hyperplane. The support vector is made up of the points of the opposite class that are nearest to each other [14].



```
              precision    recall  f1-score   support

           0       0.77      1.00      0.87        10
           1       1.00      0.40      0.57         5

    accuracy                           0.80        15
   macro avg       0.88      0.70      0.72        15
weighted avg       0.85      0.80      0.77        15
```

Fig.7. SVM Metrics for the positive case prediction

The Precision, Recall, F1-Score, and Support values have been calculated as shown in Figure 7.

KNN (Kneareast Neighbor)

KNN is a case-based learning method that is used to solve the classification and regression problem. K-NN relies on labelled data to provide the correct output for unlabeled data.KNN is one the effective classification method which convincing the Reuters corpus in text categorization to bring out the high classification accuracy.

KNN Algorithm Procedure

1. Load the COVID-19 dataset.
2. Initialize K to select a positive neighbour for each positive case in the data set.
3. Calculate the positive value from the data set, and then add the index value.
4. Sort the collected data from the smallest to the largest value in ascending order.
5. Get the top K entries from the sorted set of the COVID-19 data set.
6. Obtain the labels of a K positive case that has been chosen.
7. If there is a regression, return the label's K-means. If the label is classified, the label's K mode is restored.

Naive Baye's Algorithm

A subset of Bayesian theory is naive Bayes. For text classification, a Naive Bayes classifier is used. Classifies the transferred data in a readable format. Naïve Baye's classification algorithm discriminates against the dataset based on particular features or attributes [15].

The algorithm can be applied to a wide range of real-world scenarios, including [16-23]:

Text Classification: A probabilistic learning model is used to classify text. It's the most common method for categorizing text documents and categorical variables [24-25].

Spam Filtering: It is a widely used method. DSPAM, Spam Bayes, Spam Assassin, Bogofilter, and ASSP email filter are some methods that can be used to differentiate spam filtering [26-29].

Sentiment Analysis: It is a technique for analyzing both positive and negative feedback and reviews.

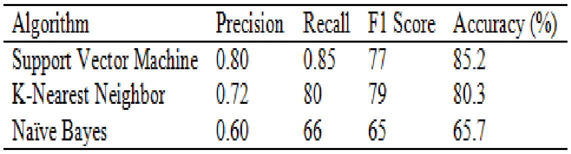
Recommendation System built a hybrid recommendation system for predictive data using a combination of collaborative filtering mechanisms.

III. RESULTS AND ANALYSIS

Comparative Analysis

The COVID-19 infectious disease threatens the global. Experimental results have shown that the various classification techniques provide accurate results. Compared to the K-Nearest Neighbor and Naive Bayes classifiers, the Support Vector Machine (SVM) achieved an accuracy of 85%. In terms of accuracy, precision, recall, and f1 score, the Support Vector Machine is faster and more precise than the other techniques. The table below depicts the results of a comparison of conventional supervised learning classifiers.

TABLE 2. COMPARATIVE ANALYSIS OF TRADITIONAL SUPERVISED LEARNING CLASSIFIER



Algorithm	Precision	Recall	F1 Score	Accuracy (%)
Support Vector Machine	0.80	0.85	77	85.2
K-Nearest Neighbor	0.72	80	79	80.3
Naïve Bayes	0.60	66	65	65.7

Result

The positive rate of coronavirus infection was compared to the predictions of the three models in this analysis. Support Vector Machine (SVM), K Nearest Neighbor (KNN), and the Naive Bayes Algorithm (NB) are three algorithms that were applied to the collected data set to compare the performance. In contrast, it was discovered that SVM is the most effective and has the highest accuracy rate of 85%, which is very useful for COVID-19 medical care with a small data collection.

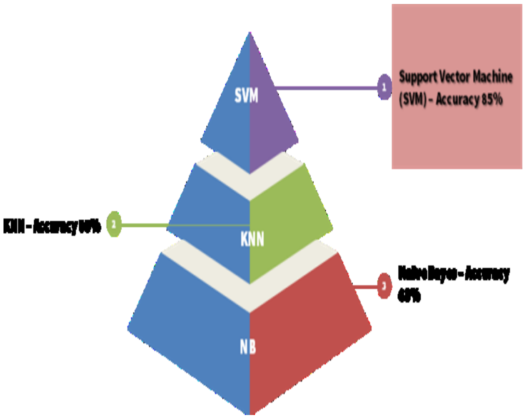

SVM
1
Support Vector Machine (SVM) - Accuracy 85%
KNN - Accuracy 80%
2
KNN
3
Naive Bayes - Accuracy 60%
NB

Fig. 8. SVM provides the highest accuracy

IV. CONCLUSION

In the present study on machine learning predictions, we have achieved various levels of classification accuracy. We use numerical variables and categorical variables for classification. The eventual results are Support Vector Machine algorithm–85% accuracy, K-Nearest Neighbor algorithm–80% accuracy, and Naïve Baye's –66% accuracy. Among these three classification methods, the SVM algorithm provides high accuracy for the provided samples. The investigated classification algorithm would be helpful to predict the Confirmed cases, Cured cases, Death cases, and Active in the future with the immense collections of COVID-19 data set. In the future, we also will investigate lots of other algorithms.

References

[1] Li Y, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection", medRxiv, doi.org/10.1101/2020.02.27.20028 027,2020.

[2] Raphael Dolin MD, "Novel coronavirus from Wuhan China. Mandell, Douglas, and Bennett's principles and practice of infectious diseases", Elsevier, Chapter vol.155,2020.

[3] Shang J, "Structural basis of receptor recognition by SARS-CoV-2. Nature", https://doi.org/10.1038/s41586-020-2179-y,2020.

[4] https://www.medicalnewstoday.com/articles/coronavirus-early-symptoms

[5] Seifedine Kadry, "Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19Class", arXiv.org,arXiv:2004.13122,2020.

[6] I. Rish, "An Empirical Study of the Naïve Bayes Classifier",2014.

[7] W. Zhang., "Procedia Engineering An Improvement to Naive Bayes for Text Classification", vol.15, pp.2160-2164,2011.

[8] Sampathkumar, A., Murugan, S., Rastogi, R., Mishra, M. K., Malathy, S., & Manikandan, R, "Energy Efficient ACPI and JEHDO Mechanism for IoT Device Energy Management in Healthcare", In the *Internet of Things in Smart Technologies for Sustainable Urban Development*, pp.131-140, 2020.

[9] L. J. Muhammad, "Supervised Machine Learning Models for Prediction of COVID- 19 Infection using Epidemiology Dataset. Springer Nature", https://doi.org/10.1007/s42979-020-00394-7,2020.

[10] Warda M. Shaban, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier", Elsevier,2020.

[11] Durgesh K, "Data classification using support vector machine",Journal of theoretical and applied information technology, 2010.

[12] Islam MM, "Prediction of breast cancer using support vector machine and K-nearest neighbours", IEEE region 10 humanitarian technology conference,2017.

[13] S. Kanaga Suba Raja, A. Sathya,S. Karthikeyan, T. Janane, "Multi cloud-based secure privacy preservation of hospital data in cloud computing", International Journal of Cloud Computing ISSN 2043-9989, vol.10, no.2, pp.101-111,2021.

[14] Danny Roobaert, "DirectSVM: A simple support vector machine perceptron", Journal of VLSISignal Processing Systems, 2001.

[15] Muhammad LJ, "Performance evaluation of classification data mining algorithms on coronary artery disease dataset", IEEE 9th international conference on computer and knowledge engineering (ICCKE 2019), Ferdowsi University of Mashhad, 2019.

[16] Pouria Kaviani, "International Journal of Advance Engineering and Research Development",e-ISSN (O): pp.2348-4470, vol.4, no.11, 2017.

[17] Rao, A. N., Vijayapriya, P., Kowsalya, M., & Rajest, S. S. (2020). Computer Tools for Energy Systems. In International Conference on Communication, Computing and Electronics Systems (pp. 475-484). Springer, Singapore.Gupta J., Singla M.K., Nijhawan P., Ganguli S., Rajest S.S. (2020) An IoT-Based Controller Realization for PV System Monitoring and Control. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham

[18] Sharma M., Singla M.K., Nijhawan P., Ganguli S., Rajest S.S. (2020) An Application of IoT to Develop Concept of Smart Remote Monitoring System. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham

[19] Ganguli S., Kaur G., Sarkar P., Rajest S.S. (2020) An Algorithmic Approach to System Identification in the Delta Domain Using FAdFPA Algorithm. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham

[20] Singla M.K., Gupta J., Nijhawan P., Ganguli S., Rajest S.S. (2020) Development of an Efficient, Cheap, and Flexible IoT-Based Wind Turbine Emulator. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham

[21] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A Comparative Analysis of Active Learning for Biomedical Text Mining," Applied System Innovation, vol. 4, no. 1, p. 23, 2021

[22] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLOS ONE, 16(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

[23] Yousaf, A., Umer, M., Sadiq, S., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021b). Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD). IEEE Access, 9, 6286–6295. https://doi.org/10.1109/access.2020.3047831

[24] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models," arXiv preprint arXiv:.15036, 2020.

[25] Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & NAPPI, M. (2021). Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning. Expert Systems with Applications, 115111. https://doi.org/10.1016/j.eswa.2021.115111

[26] Manne, R., & Kantheti, S. C. (2021). Application of Artificial Intelligence in Healthcare: Chances and Challenges. Current Journal of Applied Science and Technology, 40(6), 78-89. https://doi.org/10.9734/cjast/2021/v40i631320

[27] U. Naseem, S. K. Khan, M. Farasat, and F. Ali, "Abusive Language Detection: A Comprehensive Review," Indian Journal of Science Technology, vol. 12, no. 45, pp. 1-13, 2019.

[28] Haldorai and A. Ramu, “An Intelligent-Based Wavelet Classifier for Accurate Prediction of Breast Cancer,” Intelligent Multidimensional Data and Image Processing, pp. 306–319.

[29] S, D., & H, A. (2019). AODV Route Discovery and Route Maintenance in MANETs. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). doi:10.1109/icaccs.2019.8728456