A Machine Learning Approach for Diagnosing and Identifying Symptoms of COVID-19

Ahmed Banimustafa, Olla Bulkrock and Jamal Al Qundus

December 7, 2022

A Machine Learning Approach for Diagnosing and Identifying Symptoms of COVID-19

Ahmed BaniMustafa, IEEE Senior Member
Data Science and Artificial Intelligence Department
Isra University, Amman, Jordan
a.banimustafa@iu.edu.jo

Olla Bulkrock
Data Science Department,
Princess Sumaya University for Teachnology
Amman, Jordan
Oll20228067@std.psut.edu.jo

Jamal Al Qundus
Artificial Intelligence Department
Middle East University
Amman, Jordan
Jalqundus@meu.edu.jo

***Abstract*— Coronavirus 2019 (COVID-19) is a pandemic that hit the world and was responsible for the death of millions and the life disruption of billions of people. One of the most critical challenges faced during the earlier breakthrough of the diseases was identifying symptoms confused with colds, flu, and other common infections. Nevertheless, despite all the effort and research conducted for this purpose, this challenge continues as more strains, variants, and mutations appear. This work presents a solution for this problem based on machine learning classification and variable importance algorithms. A public dataset of 274,957 cases has been classified into typical and COVID-19 cases based on the reported symptoms and other variables. The dataset was used for classifying the reported cases using K-nearest neighbor (KNN), Naïve Bayes, and Decision Trees (DT) algorithms and identifying the significant symptoms that were decisive in classifying the patients using Gini, Information Gain, and Information Gain Ratio algorithms. Naïve Bayes and Decision Trees performed best with a Classification Accuracy (CA) score of 95.2% and 96.3%, respectively. The Naïve Bayes classifier scored an Area Under the Curve (AUC) of 88.75%. In addition, the applied variable importance algorithms identified headache, fever, and sore throat as the most important symptoms.**

Keywords—Machine Learning, Data Mining, Health Informatics, Medical Diagnosis, COVID-19 / SARS-Cov-2.

I. Introduction

Coronavirus 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first case of COVID-19 was reported in Wuhan, China, in December 2019 [1]. Since then, the virus has continued to spread worldwide, threatening the lives of millions all around the globe. The World Health Organization (Who) announced the disease as a global pandemic at the beginning of 2020. Until 2022, 293 million cases have been reported worldwide, with a toll of 5.45 million deaths [2]. The disease continues to spread worldwide, and its virus mutates through several dominant variants, such as Alpha, Delta, and Omicron strains, making the virus even more difficult to contaminate.

Machine learning provides innovative solutions for classifying samples and identifying variable scoring importance that contributes to the classifications [3], which can be utilized for diagnosing COVID-19 variants and identifying its most discriminant symptoms.

This work proposes a machine learning approach for predicting COVID-19 infections and identifying its related signs and most significant symptoms, which are usually shared with other similar diseases, and which can also be used as significant indicators for predicting COVID-19 infections such as cough, fever, sore throat, shortness of breath, and headache. The results of this research can be used to narrow down the diagnosis symptoms and differentiate COVID-19 infection from those related to other common diseases such as cold and seasonal flu. They can also help provide better screening for the disease before conducting Polymerase Chain Reaction (PCR) tests and X-Radiation (X-Ray) based diagnostics. For this purpose, we applied three machine learning algorithms to classify the cases into positive and negative binary classes: Naïve Bayes, Decision Trees (DT), and K-Nearest Neighbour (KNN).

The second section of this paper investigates the related work, while the following section describes the input dataset and its attributes. The fourth section describes the applied methodology, including the machine learning and variable importance ranking algorithms and the evaluation metrics used. The fifth section summarises the reported results and their discussion, while the sixth section draws a conclusion and comments on future work.

II. Related Studies

Several studies have been conducted to identify the symptoms of COVID-19 disease. A study reported in [3] analyzed the symptoms of 169 positive cases, which were surveyed and collected from patients using a mobile phone application, which involved recording the temperature of patients, taking pictures of the ear surface, and videos of the patients' faces. The findings were used to predict COVID-19 infections and then compared to the results of PCR tests for the same patients. In another study reported by [4], the researchers applied machine learning to identify the disease based on cough sounds after creating a multimedia database for the sound of the cough of the infected patients. However, this study was theoretical, as its results have never been applied in a real diagnosis system. An expert system was also proposed in [5] to help doctors diagnose the COVID-19 disease based on the symptoms identified based on a human expert's knowledge and an embedded adaptive learning algorithm. A study conducted in Bangladesh in 2014 reported using evidential reasoning and belief rule-based systems for differentiating several types of flu diseases. The study reported that the system was more successful than the traditional doctor's diagnosis [6]. A knowledge-based system was reported in [7] to classify 32 lung diseases that share common symptoms. The system successfully classified the diseases with 70% accuracy, which is relatively low compared to the capabilities of machine learning if it were used in the study.

The analysis of the related studies shows that most of the reported studies failed to address the problem of discriminating COVID-19, flu, respiratory, and other

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

associated diseases based on the patient-reported or expert-provided symptoms. However, despite using machine learning [4] to identify the diseases based on a cough sound database, the study failed to achieve reliable results and was conducted only as a theoretical study. On the other hand, the study reported in [7], based on knowledge-based systems, was applied to other lung-related diseases and can also be applied to COVID-19. However, the lack of accuracy of this study can be enhanced if the symptoms were identified based on machine learning analysis of real-world reported cases. However, the other study was reported in [6], which aimed at differentiating several Influenzas that can also be applied to COVID-19 and other respiratory-related diseases. However, the study's dependence on human experts to identify the symptoms makes it less dependable, particularly for diagnosing COVID-19. The findings of the related work analysis show that all the reported studies failed to address the problem of the differentiation of the symptoms of COVID-19 from those shared with other colds, flu, and other respiratory-related diseases. One of the issues in these studies is that they depended on human experts. A more successful approach can utilize machine learning to identify symptoms that distinguish COVID-19 and its variants from those related to the common cold, flu, and other respiratory-related diseases. The dataset must include a substantial number of positive and negative cases; the symptoms must be examined by physicians and then confirmed by laboratory tests. This gap is aimed to be addressed by this study.

III. MATERIALS/DATASET

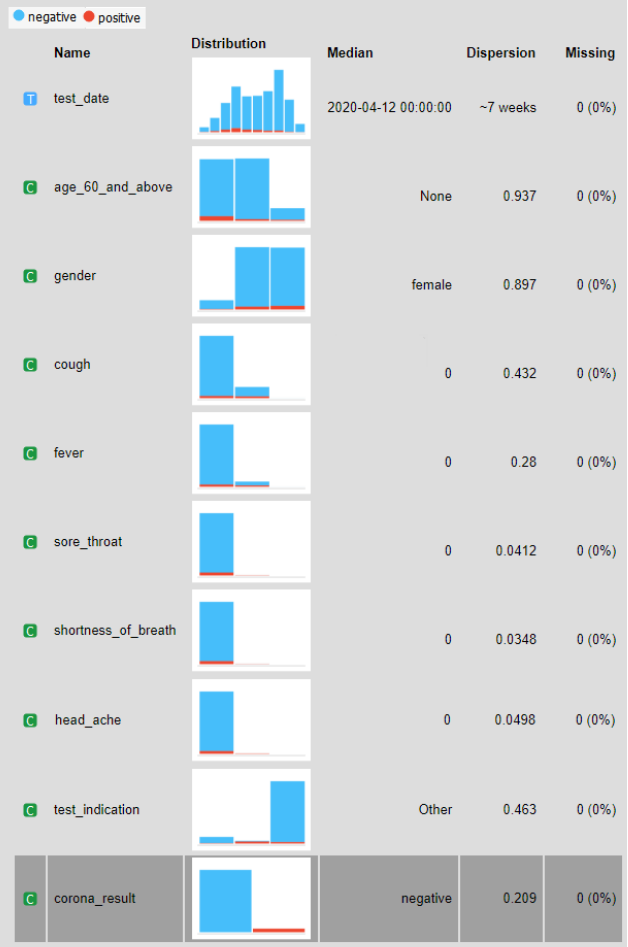
The dataset used in this research consists of 274,957 cases, which covers symptoms that include: cough, fever, sore throat, shortness of breath, headache, and other important factors such as age, gender, test date, test indication, and infected people contacting information [8]. Table I. summarises the variables of the dataset used in this study, while Figure 1. shows the dataset distribution and the basic statistics of the dataset features. The statistics cover missing values, data dispersion, median, and values distribution.

IV. METHODOLOGY

This section provides details regarding the methodology and the algorithms applied for predicting the diagnosis of COVID-19 cases and the variable importance algorithms that have been applied for identifying the significance of COVID-19 symptoms.

TABLE I. DATASET VARIABLES DESCRIPTION

Variable	Description	
	Data Type	***Values***
test_result	Binary	{Positive, Negative}
Test_date	DateTime	Date of the test
test_indication	Categorical	{Contact with confirmed, Abroad, Other}
Gender	Categorical	{Male, Female}
age_60_and_above	Binary	{Yes, No, Other}
Cough	Binary	{0,1}
Fever	Binary	{0,1}
Sore_throat	Binary	{0,1}
Shortness_of_breath	Binary	{0,1}
Head_ache	Binary	{0,1}



● negative ● positive

	Name	Distribution	Median	Dispersion	Missing
T	test_date		2020-04-12 00:00:00	~7 weeks	0 (0%)
C	age_60_and_above		None	0.937	0 (0%)
C	gender		female	0.897	0 (0%)
C	cough		0	0.432	0 (0%)
C	fever		0	0.28	0 (0%)
C	sore_throat		0	0.0412	0 (0%)
C	shortness_of_breath		0	0.0348	0 (0%)
C	head_ache		0	0.0498	0 (0%)
C	test_indication		Other	0.463	0 (0%)
C	corona_result		negative	0.209	0 (0%)

Figure 1. Data distribution and basic statistics.

A. *Classification*

Three machine learning algorithms are applied in this work: Naïve Bayes, K-Nearest Neighbor, and Decision Trees.

1) *Naïve Bays*

A supervised learning algorithm that can be used for classification. This algorithm works by mapping a set of observations S to a set of labels or classes C. It calculates the probability of the sample belonging to the label as well as the frequencies of its belonging to the assigned class by finding the maximum likelihood using equation 1 [9].

$$P(C\backslash S)= P(S\backslash C)\ P(C)/P(S) \qquad (1)$$

Where P(C) is the probability of the prior probability of the C class, P(S) is the probability of the samples, and P(S\C) is the likelihood probability of the sample S given the class C. In contrast, P(C\S) is the posterior probability of class C; given the observation S, The Naïve Bayes algorithm considers the prior knowledge of the observation classification by multiplying the probability of this prior knowledge by the possible labeling or classification likelihood. The new observations are then labeled by assigning their classes, considering the need for increasing the calculated posterior value.

The model's training is conducted by an estimated value for the probability of the sample assigned to a class according to the frequency of the sample belonging to the class during training. Naïve Bayes was reported successful in several machine learning applications [10].

2) *K-Nearest Neighbor (KNN)*

K-Nearest Neighbor (KNN) is an instance-based learning technique for classifying data samples by measuring their proximity to neighboring data points belonging to a set of pre-labeled classes. This technique was introduced by Evelyn Fix and JL Hodges, Jr. in their unpublished technical report while working at the United States Air Force (USAF) School of Aviation [11, 12].

The KNN algorithm measures the Euclidean distance between predicted and training values belonging to a predefined class in a two-dimensional space using equation 2.

$$\Delta(x_i, x_j) = \sqrt{\sum_{i=1}^{n}(|\ x_{in} - x_{jn}\ |)^2} \qquad (2)$$

The predicted classes for a point are determined based on a plurality vote regarding its distance from neighboring data points that belong to the adjacent classes. Figure 1. illustrates how the value x is classified based on its Euclidean distance to the neighboring classes [13]. The application of the KNN algorithm was reported successful in several machine learning applications [14].

3) *Decision Trees Decision Trees (DT)*

Decision Trees (DT) are a popular machine learning algorithm with several implementations, including Decision Trees Induction (ID3), C4.5, C5.0, and J48. Decision Trees gain popularity to the simplicity of model interpretation and to the transparency of the model, which enables tracing that results in the form of logical rules that can be visualized as a tree-like diagram. Decision Trees were successfully used in several applications to classify samples into a set of given classes or rank the variables' importance toward classification. [15, 16]. The gain algorithm can also find the feature with the maximum information about the sample labeling or classification [17].

B. *Variables Importance Ranking*

The variable's importance ranking in this research includes information gain and GINI algorithms. In this subsection, we provide an overview of two techniques.

1) *Information Gain*

The Information Gain algorithm is based on the information theory and entropy concepts used in Decision Trees and other machine learning algorithms. An information gain algorithm can also find the feature with the maximum information about the sample labeling or classification [17].

2) *Gain Ratio*

The gain ratio is a ratio of information gain to intrinsic information. This algorithm was proposed by Ross Quinlan, to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute [11, 18].

3) *GINI*

The GINI index calculates the degree or probability of a variable misclassification as it is randomly selected, similar to finding the GINI coefficients. GINI can be applied to categorical data, and its outcome is either a successful or failed classification [18]. The GINI index is also used for the variable's importance ranking based on the variable's importance for building a successful classifier [14].

C. *Evaluation Criteria*

The evaluation strategy applied in this research involves using the classification accuracy (CA), Area Under the Curve (AUC) [19], and Receiver Operator Characteristics plot (ROC) [20, 21] in addition to the use of a Confusion Matrix. The evaluation was performed based on five-fold cross-validation [22]. The confusion matrix visualizes the classification model performance in a tabular form. The rows represent the samples in a predicted class predicted by the classification model, while the columns represent the samples in the actual class [23]. The Confusion Matrix helps evaluate the model robustness by showing the number of samples in true-positive, true-negative, false-positive, and false-negative classification. The equation for calculating the classification accuracy (CA) is shown in Equation 3.

$$\text{Classification Accuracy} = \text{TP+TN/ N.} \qquad (3)$$

Where TP represents the number of samples classified as belonging to the assigned class, TN represents the number of samples classified as not belonging to the assigned class. N is the total number of samples.

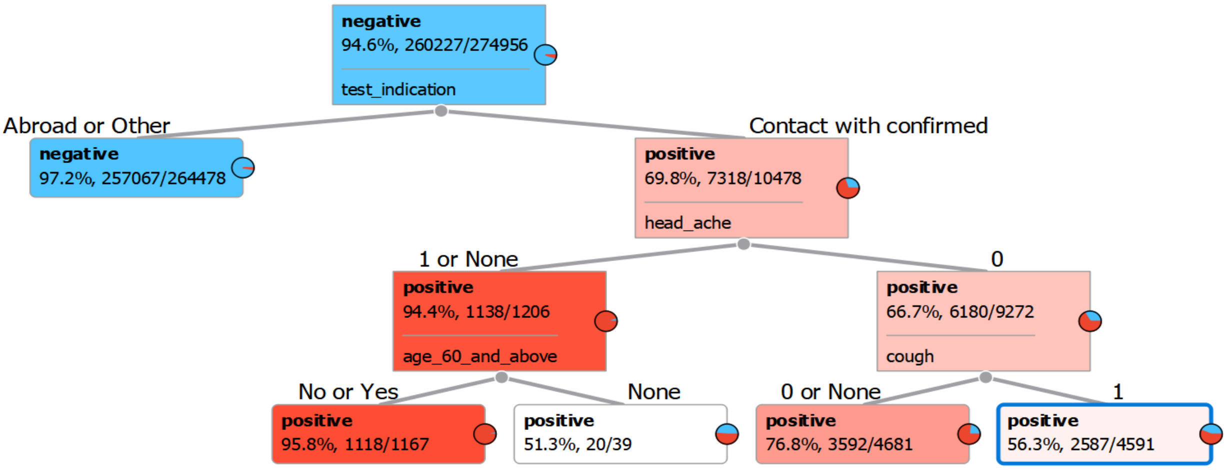


Figure 2. Visualization of the Decision Trees classifier

ROC plots are also used in the evaluation. The ROC curve maps the false positive rate to the true positive rate of the class prediction. This curve is widely accepted as an excellent and accurate metric of machine learning model performance, in addition to the Area Under the Curve Measure (AUC) [20].

V. Results and Discussion

In this section, we present the results obtained in this research using machine learning classification and variables importance ranking algorithms in addition to their evaluation using Confusion Matrix, Classification Accuracy, and ROC plots.

A. Classification Results

The classification results of applying the three machine learning algorithms demonstrated that all the applied techniques could predict COVID-19 infections with a considerable classification accuracy that exceeded a threshold of 92% based on the given symptoms and other related predictors such as age, gender, test date test indication and infected people contacting information. Table II. Summarize the classification performance of the three applied classifiers based on AUC and CA metrics.

TABLE II. Classiers Performance

Model	**Performance**	
	AUC	***CA***
Naïve Bayes	88.7%	95%
Decision Tree	74.0%	96.3%
KNN	50.6%	92.1%

The visualization in Figure 1. illustrates the Decision Trees created by the Decision Trees classifier for a maximum depth of four levels. The score on each node shows the classification level of confidence and the number of classified cases, while the labels on the Trees branches (edges) show the value for the node split Decision factor.

The Decision Trees show that infected people contracting COVID-19 play the most significant role in diagnosing positive cases. At the same time, headache and cough are the most important symptoms for diagnosing the disease. Table III. Demonstrates the confusion matrix for the Decision Trees classifier. The confusion matrix for the Decision Trees classifier shows that it was more successful in classifying negative cases than its capability to classify positive cases.

This result is quite important, knowing that the most common PCR tests suffer from poor false negative results, which might reach 40%, making the Decision Trees classifier a viable candidate for confirming or complementing the results of PCR tests. On the other hand, the confusion matrix of the KNN classifier is illustrated in Table III. shows that KNN performed less than Decision Trees when it comes to negative cases, while it achieved similar performance to Decision Trees when classifying positive cases. These results confirm the Decision tree's superiority over KNN algorithms, as indicated by the classification accuracy metric, as it achieved 96% compared to the 92% accuracy achieved by the KNN classifier. Table IV. shows the confusion matrix of the KNN classifier. However, the confusion matrix of Naïve Bayes is demonstrated in Table V., which shows better classification results for negative cases than KNN, but worse than Decision Trees. On the other hand, the Naïve Bayes classifier performs better than both KNN and Decision Trees regarding positive cases, which might explain the superiority of the Naïve Bayes classifier in the AUC metric compared to both Decision Trees and KNN classifiers. This observation confirms the concern over the use of classification accuracy as the core evaluation metric for judging the performance of machine learning classifiers and the most recent recommendations of data mining and machine learning practitioners to depend more on ROC and AUC as more accurate techniques when judging and comparing classifiers as reported in the recent literature.

TABLE III. The Confusion matrix for the Decision Trees classifier

		Predicted		
		Negative	**Positive**	**Sum**
Actual	Negative	**258416**	1811	**260227**
	Positive	8373	**6356**	**14729**
	Sum	**266789**	**8167**	**274956**

TABLE IV. The Confusion matrix for the KNN classifier

		Predicted		
		Negative	**Positive**	**Sum**
Actual	**Negative**	**246306**	13921	**260227**
	Positive	7814	**6915**	**14729**
	Sum	**254120**	**20836**	**274956**

TABLE V. The Confusion matrix for the Naive Bayes classifier

		Predicted		
		Negative	**Positive**	**Sum**
Actual	**Negative**	**253601**	6626	260227
	Positive	6643	**8086**	14729
	Sum	**260244**	**1412**	**274956**

The ROC curves of the three applied classification algorithms: KNN, Decision Trees, and Naïve Bayes, illustrated in Figure 3. confirm our previous findings and discussion related to the three classifiers' performance shown in confusion. Hence, based on the ROC plot of the applied classifiers, we can conclude that Naïve Bayes has a better ROC curve and AUC score than the other classifiers. The Decision Trees classifier, despite its excellent classification accuracy, is the worst regarding the ROC curve and AUC metric.

B. Variable Importance Ranking Results

Variable importance ranking was applied based on thetwo applied ranking algorithms, information gain, and GINI, which have already been discussed in the theoretical framework section for all the reported COVID-19 symptoms. The results show that headache and fever are the most important symptoms, while cough and sore throat are relatively less significant but less important. Shortness of breath was also among the five most significant symptoms.

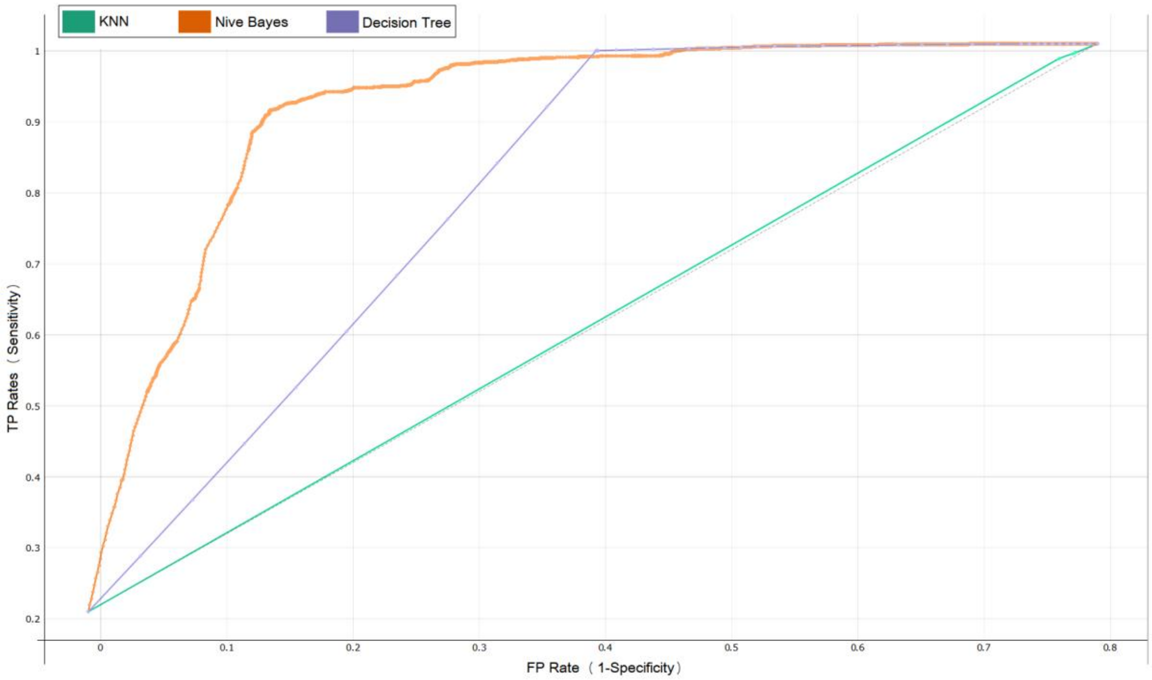
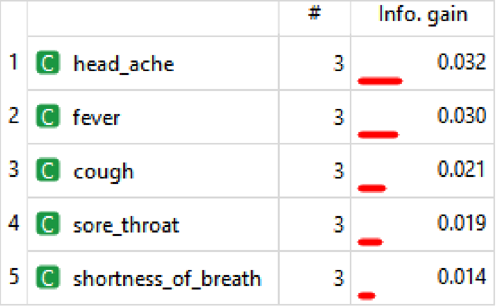


Figure 3. ROC plot of KNN, Naïve Bayes, and Decision Trees classifiers

Variable importance ranking was applied based on the These results are consistent with the recommendations published by the WHO organization and with the guidance of the Jordanian Ministry of Health (MoH). Figure 4. shows the symptoms ranked by the Information Gain algorithm. At the same time, Figure 5. provides a ranking for the importance of COVID-19 symptoms based on the GINI variable importance ranking algorithm, which found that headache was also ranked as the most important predictor for COVID-19 infection. However, sore throat and fever were also found important, while shortness of breath was the least important symptom compared to the former. Nevertheless, GINI ranks cough as the least important symptom, ranked third by the Information Gain algorithm. COVID-19 infection, while shortness of breath was found to be an insignificant indicator of the infection. Other related factors such as age, gender, test date, test indications, and infected people's contact information. Figure 6. The variables' importance ranking for the symptoms was obtained based on the Gain Ratio algorithm.

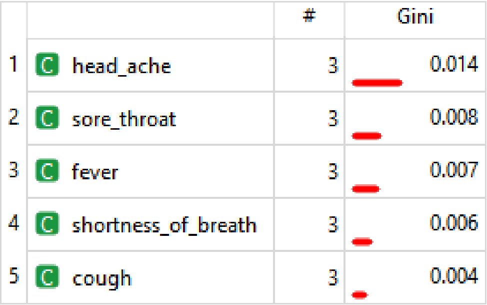
The ranking of the Information Gain Ration algorithm seems to agree with both Information Gain and GINI algorithms in ranking headaches as the most important symptom of COVID-19. However. It agrees more with GINI in ranking the other symptoms with only one exception, as it ranks shortness of breath as a more significant symptom than fever and cough. When comparing and discussing the results of the three algorithms that have been applied to the dataset for ranking the importance of variables for predicting the importance of variables, and for ranking the significance of COVID-19 symptoms for diagnosing the disease, it was found that all three techniques agree that headache is the most important symptoms. At the same time, they disagree in ranking most of the other symptoms. GINI and Information Gain Rate agrees on ranking cough as the least essential symptom, while the Information Gain algorithm ranks it as the third most crucial symptom. On the other hand, both GINI and Information Gain Ratio algorithms rank sore throat as the second most important symptom. However, the rank of the shortness of breath symptom varies significantly from one algorithm to another. While it is ranked the least important by the Information Gain algorithm, it is ranked fourth by GINI and Fifth (before last) by the Information Gain Ratio algorithm.

The performance of all the created classification models is excellent. The KNN model scored 92%, while the Naïve Bayes and Decision Trees models scored 95% and 96.3%, respectively. The AUC score of the naïve Bayes model was 88.7% which is quite acceptable, while DT and KNN scored 74% and 50.4% respectively. The performance of the confusion matrixes of the Naïve Bayes and DT models was quite good; however, the confusion matrix of the KNN model was less acceptable. The ROC curve confirms the naïve Bayes and Decision Trees models' robustness while it uncovers issues with the KNN model. The variable importance ranking results were also helpful in identifying the disease symptoms based on an insightful analysis of the cases, which can help achieve more accurate, reliable, and fast diagnosis, which might help rescue lives and reduce costs. It could also automate the diagnosis process by creating a knowledge-based system (KBS). This would also save costs and reduce the risk of infection.



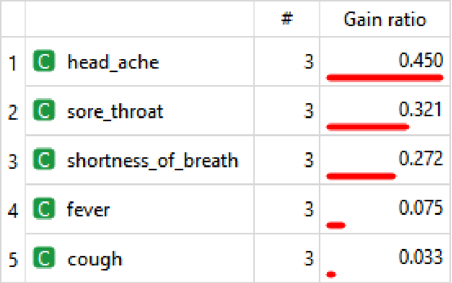
		#	Info. gain
1	C head_ache	3	0.032
2	C fever	3	0.030
3	C cough	3	0.021
4	C sore_throat	3	0.019
5	C shortness_of_breath	3	0.014

Figure 4. Information Gain algorithm symptoms ranking.



		#	Gini
1	C head_ache	3	0.014
2	C sore_throat	3	0.008
3	C fever	3	0.007
4	C shortness_of_breath	3	0.006
5	C cough	3	0.004

Figure 5. GINI algorithm symptoms ranking.



		#	Gain ratio
1	C head_ache	3	0.450
2	C sore_throat	3	0.321
3	C shortness_of_breath	3	0.272
4	C fever	3	0.075
5	C cough	3	0.033

Figure 6. Gain Ratio algorithm symptoms ranking.

VI. Conclusion

Machine learning algorithms successfully predicted COVID-19 infections based on the disease-reported symptoms and three other factors, including age, gender, test date, test indication, and infected people's contact information. The prediction results of COVID-19 infection classification were excellent using all three applied classifiers Decision Trees and Naïve Bayes and KNN, according to classification accuracy as they scored 96%, 94%, and 92%, respectively. However, Naïve Bayes outperformed both Decision Trees and KNN in the AUC metric, which depends on calculating the Area Under the Curve in ROC plots, which confirms the classification performance of the three algorithms in predicting both negative and positive case classes. On the other hand, the variables importance ranking algorithms that were applied to find the most significant symptoms for COVID-19 found that headache and sore throat are the most important symptoms in addition to fever. In contrast, cough and shortness of breath were found to be less significant symptoms for diagnosing COVID-19. These findings are quite important for both public and doctors who depend on the reported symptoms for making the initial diagnosis and for public monitoring and screening. While fever was reported as the most important symptom in most of the Ministry of Health (MoH) and WHO guidelines and publications, the headache was ranked as the most significant symptom, followed by sore throat in this research.

The findings of this research can help create KBS and Expert systems that are used for diagnosing COVID-19 while giving the symptoms a weight factor that depends on their significance and relevance to COVID-19 disease while collecting more data regarding COVID-19 mutations, variants, and strains to identify the difference in there reported symptoms. We also recommend extending this work by applying more classification and variable importance techniques and algorithms to improve these algorithms' performance and accuracy and provide a more rapid and reliable diagnosis for COVID-19 disease cases.

VII. References:

[1] J. Page, D. Hinshaw, and B. McKay, "In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market–The man with the first confirmed infection of the new coronavirus told the WHO team that his parents had shopped there," The Wall Street Journal, 2021.

[2] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," The Lancet infectious diseases, vol. 20, no. 5, pp. 533-534, 2020.

[3] D. Çelik Ertuğrul and D. Çelik Ulusoy, "A knowledge‐based self‐pre‐diagnosis system to predict Covid‐19 in smartphone users using personal data and observed symptoms," Expert systems, vol. 39, no. 3, p. e12716, 2022.

[4] A. N. Belkacem, S. Ouhbi, A. Lakas, E. Benkhelifa, and C. Chen, "End-to-end AI-based point-of-care diagnosis system for classifying respiratory illnesses and early detection of COVID-19: a theoretical framework," Frontiers in Medicine, vol. 8, p. 585578, 2021.

[5] M. Alnajjar, "A Proposed Expert System for Cold and Flu Diseases Diagnosis," International Journal of Academic Engineering Research (IJAER), vol. 4, no. 4, pp. 10-16, 2020.

[6] M. S. Hossain, M. S. Khalid, S. Akter, and S. Dey, "A belief rule-based expert system to diagnose influenza," in 2014 9Th international forum on strategic technology (IFOST), Chittagong, Bangladesh, 2014, pp. 113-116: IEEE.

[7] J. Singla, "The diagnosis of some lung diseases in a prolog expert system," International Journal of Computer Applications, vol. 78, no. 15, pp. 37-40, 2013.

[8] J. Hasell et al., "A cross-country database of COVID-19 testing," Journal of Scientific data, vol. 7, no. 1, pp. 1-7, 2020.

[9] K. M. Leung and R. Engineering, "Naive bayesian classifier," Department of Computer Science/Finance and Risk Engineering, vol. 2007, pp. 123-156, 2007.

[10] A. BaniMustafa, "Predicting Software Effort Estimation Using Machine Learning Techniques," in 2018 8th International Conference on Computer Science and Information Technology (CSIT), Amman, 2018, pp. 249-256: IEEE.

[11] B. W. Silverman and M. C. Jones, "E. fix and jl hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951)," International Statistical Review/Revue Internationale de Statistique, pp. 233-238, 1989.

[12] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", 2003, pp. 986-996: Springer.

[13] W. J. Hwang and K. W. Wen, "Fast kNN classification algorithm based on partial distance search," Journal of Electronics letters, vol. 34, no. 21, pp. 2062-2063, 1998.

[14] A. BaniMustafa, "Enhancing learning from imbalanced classes via data preprocessing: A data-driven application in metabolomics data mining," ISeCure, vol. 11, no. 3, pp. 79-89, 2019.

[15] J. R. Quinlan, "Simplifying decision trees," International journal of man-machine studies, vol. 27, no. 3, pp. 221-234, 1987.

[16] A. BaniMustafa and N. Hardy, "Applications of a novel knowledge discovery and data mining process model for metabolomics," arXiv, vol. arXiv:.03755, 2019.

[17] W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik, "Comparison of feature ranking methods based on information entropy," in 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), 2004, vol. 2, pp. 1415-1419: IEEE.

[18] A. Bhattacharya and R. T. Goswami, "Comparative analysis of different feature ranking techniques in data mining-based android malware detection," in Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications, Singapore, 2017, pp. 39-49: Springer.

[19] C. X. Ling, J. Huang, and H. Zhang, "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms," in Advances in Artificial Intelligence, Berlin, Heidelberg, 2003, pp. 329-341: Springer Berlin Heidelberg.

[20] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?," Emergency Medicine Journal, vol. 34, no. 6, pp. 357-359, 2017.

[21] C. E. Metz, "Basic Principles of ROC analysis," Seminars in Nuclear Medicine, vol. 8, no. 4, pp. 283-298, 1978.

[22] A. Banimustafa and N. Hardy, "A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics," IEEE Access, vol. 8, pp. 209964-210005, 2020.

[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Journal of Information Processing Management, vol. 45, no. 4, pp. 427-437, 2009.