© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

International Journal of Integrated Engineering

COVID-19: Symptoms Clustering and Severity Classification Using Machine Learning Approach

Nurul Fathia Mohamand Noor[1], Herold Sylvestro Sipail[1], Norulhusna Ahmad[1], Bayram Annanurov[2], Norliza Mohd Noor[1]*

[1]Razak Faculty of Technology and Informatics,
University Teknologi Malaysia, Kuala Lumpur, 54100, MALAYSIA

[2]Center for Biomedical Informatics.
Wake Forest School of Medicine, Winston-Salem, NC, USA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2023.15.03.001
Received 30 October 2022; Accepted 29 December 2022; Available online 31 July 2023

Abstract: COVID-19 is an extremely contagious illness that causes illnesses varying from either the common cold to more chronic illnesses or even death. The constant mutation of a new variant of COVID-19 makes it important to identify the symptom of COVID-19 in order to contain the infection. The use of clustering and classification in machine learning is in mainstream use in different aspects of research, especially in recent years to generate useful knowledge on COVID-19 outbreak. Many researchers have shared their COVID-19 data on public database and a lot of studies have been carried out. However, the merit of the dataset is unknown and analysis need to be carried by the researchers to check on its reliability. The dataset that is used in this work was sourced from the Kaggle website. The data was obtained through a survey collected from participants of various gender and age who had been to at least ten countries. There are four levels of severity based on the COVID-19 symptom, which was developed in accordance to World Health Organization (WHO) and the Indian Ministry of Health and Family Welfare recommendations. This paper presented an inquiry on the dataset utilising supervised and unsupervised machine learning approaches in order to better comprehend the dataset. In this study, the analysis of the severity group based on the COVID-19 symptoms using supervised learning techniques employed a total of seven classifiers, namely the K-NN, Linear SVM, Naive Bayes, Decision Tree (J48), Ada Boost, Bagging, and Stacking. For the unsupervised learning techniques, the clustering algorithm utilized in this work are Simple K-Means and Expectation-Maximization. From the result obtained from both supervised and unsupervised learning techniques, we observed that the result analysis yielded relatively poor classification and clustering results. The findings for the dataset analysed in this study do not appear to be providing the correct result for the symptoms categorized against the severity level which raises concerns about the validity and reliability of the dataset.

Keywords: COVID-19 symptom, machine learning, classification

1. Introduction

The COVID-19 outbreak has swept over the world since its onset in November 2019, infecting more than a hundred million individuals and killing more than 4 million people as of July 8, 2021 [1]. COVID-19 vaccination program may have been deployed globally since December 2020, with efficacy varying from 50.38 per cent to 95 per cent. While immunisation is the most effective strategy to contain the virus, there is concern that COVID-19 mutations can render the vaccination ineffective [2]. Immunisation will be a lengthy process as various COVID-19 variations evolve. Therefore, rapid COVID-19 detection alternatives are essential to minimise the virus from spreading.

*Corresponding author: norliza@utm.my

At the end of 2019, a pneumonia outbreak with an unknown aetiology was identified at the Wuhan Market, located in the province of Hubei, China. [3]. In January 2020, the unknown virus that caused it was identified by the World Health Organization (WHO) as COVID-19. This novel coronavirus is called severe acute respiratory syndrome coronavirus, SARS-CoV-2 [4]. The WHO declared a pandemic in February 2020 due to the epidemic and confirmed cases worldwide. [5]. More than a year later, in July 2021, the disease affected at least 195 countries worldwide [6], with a higher number of cases in high population density [7]. The virus mainly transmits through aerosolised droplets from the lungs when a person who has been infected coughs or sneezes [8]. According to WHO clinical recommendations, the presentation of symptoms varies, and most infected people experience fever, congestion, tiredness, loss of appetite, difficulty breathing, and myalgia. There were also reports of non-specific symptoms, including headache, throat irritation, nasal congestion, and the loss of smell or taste before experiencing respiratory difficulty [9].

Machine learning is a branch of artificial intelligence that focuses on creating systems that can learn from examples and evolve independently without complicated programming. The three fundamental machine learning approaches are supervised learning, unsupervised learning, and reinforcement learning. Predictive modelling and regression are two of the most utilised methods in supervised learning. Because of the learning technique, the system may learn from a dataset with pre-specified labels. On the other hand, unsupervised learning is used to discover patterns and extract features from enormous amounts of unlabelled data. High-dimensional and massive datasets are subjected to clustering and dimensionality reduction, the two most prevalent methods of learning used in this type of learning. According to a recent study, an artificial intelligence (AI) and machine learning (ML) strategy is suggested to make a significant contribution to mitigating the severity of viral outbreaks [10]. Maliki *et al.* in [11] predicted the COVID-19 patients' mortality rate based on the impact of weather variables such as temperature and humidity on the transmission of COVID-19. Various machine learning techniques are presented by extracting the relationship between the number of confirmed cases and the weather variables in certain regions. However, the authors are not focused in terms of prediction model accuracy. A review based on data mining and machine learning algorithm for the Coronavirus family is presented in [12]. The reviews clearly show a need to further analyse datasets for COVID-19 to minimise the spread of this virus. In order to predict a possible COVID-19 outbreak, authors in [13] devised a predictive approach that included Support Vector Machines (SVM) and achieved 98.88% accuracy. The authors claim that they used data from WHO; however, details of the data are not revealed. The research fields that are constantly done are to forecast infection and mortality rates and establish a system to classify symptoms according to their medical assessment [14, 15]. These investigations are essential, and the outcomes will significantly help healthcare workers be prepared and focus on implementing all the necessary protocols to prevent the virus from spreading even more widely.

Since the pattern in the datasets is unclear, the focus of this work is to investigate the reliability of the dataset by employing supervised and unsupervised machine learning techniques to classify the symptoms experienced by COVID-19 patients according to the severity level recorded in the dataset. The methods considered are K-Nearest Neighbour (K-NN), Linear SVM, Naive Bayes, Decision Tree (J48), Ada Boost, Bagging, and Stacking for supervised machine learning approaches and Simple K-Means and Expectation-Maximization (EM) for unsupervised machine learning techniques. We analysed it using WEKA 3: Machine Learning Software in Java [16] to determine the confusion matrix for each classifier and compared their accuracy. The paper is organised as follows: Section 2 highlights the machine learning algorithm applied in this paper, followed by the material and methods in Section 3, the result and discussion in Section 4 and lastly, the conclusion.

2. Machine Learning Techniques

Supervised and unsupervised learning are two techniques used in machine learning and Artificial Intelligence. Supervised learning techniques can be distinguished by using labelled datasets to train algorithms that accurately classify data or predict outcomes [17]. Unsupervised learning works on unlabelled data to find structure in its input and detect hidden patterns in the dataset [18].

The paper presents work on supervised and unsupervised machine learning techniques to determine the accuracy of the dataset. The supervised learning used in this work includes K-NN, Linear SVM, Naive Bayes, J48, Ada Boost, Bagging, and Stacking. Supervised learning techniques utilise a labelled dataset to train algorithms for effectively classifying data or predicting. The unsupervised learning technique used on the dataset is Simple K-means and EM for patterns discovery and gain perception in an unclear dataset.

2.1 Supervised Learning

The K-NN classifier was based on the concept of length and vicinity of connecting dots in a graph. The Euclidean distance is a well-known distance matrix used to compute the distance between two points [19]. For example, the Euclidean distance d between node q and p for a total of n is given as

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

Since the K-NN algorithm is non-parametric, the algorithm makes no assumptions about the underlying data. It is also known as a lazy learner algorithm since it does not instantly learn from the training set but instead saves it and uses it to classify.

Linear SVM find the best decision boundary or hyperplane to separate the points and classified into different classes using only single or multiple straight lines based on the data. The support vector represents the points, while the gaps between this point and the hyperplane represent the margin. Linear SVM aims to optimise the margin [20].

Based on Bayes' theorem, Naïve Bayes classifiers are among the simplest probabilistic classifiers known and have strong feature independence assumptions. The classifier used this method based on the premise that the impact of each attribute value on a given class is independent of the impact of the other attributes on the class [21].

J48 method can be used in various circumstances by employing the classification and regression strategy, a layered approach in which one choice leads to another. Pruning can be a precise tool in some situations [22]. It addresses issues such as numeric attributes, missing values, pruning, predicting error rates, decision tree induction complexity, and creating rules from trees.

AdaBoost is a type of ensemble learning created to improve binary classifiers' accuracy. AdaBoost employs an iterative strategy to improve poor classifiers by learning from their mistakes. This method is used to generate and repair faults in various sequential models. It will then give faulty predictions more weight so that the next model may deliver more accurate predictions [23].

Bagging is among the ensemble models that the machine learning algorithm employs to learn new knowledge. This strategy will aggregate all the different outcomes from different models to provide a generic result. Because the outputs from each model may be the same, it is not necessarily the case that merging numerous models would result in a definitive conclusion. One of the solutions that have been proposed is the use of bootstrapping techniques as a result [24].

Stacking is a strategy to improve an ensemble learning model produced by pooling predictions from multiple nodes and combining them into a single model. This final model is applied to the test dataset to generate predictions, which are then checked against the data [25]. The core idea is to use a training dataset to train machine learning algorithms and then use these models to produce a new dataset. The combiner machine learning algorithm then uses this new dataset as input.

2.2 Unsupervised Learning

Simple K-Means is an easy-to-use algorithm with quick convergence; hence, it is frequently employed in clustering. However, the K-value must be provided beforehand because it will impact the convergence outcome [26]. In this study, we are using K-value is determined to be 4, which corresponds to the four severity clusters.

The EM method estimates the maximum likelihood of parameters and ensures the likelihood function's convergence. It is an iterative method that clusters the symptoms based on the probability of the symptoms belonging to a cluster. This algorithm is effective in dealing with unknown connections [27].

Simple K-Means is usually only relevant to numerical data. The distance function utilised was Euclidean, and the initialisation methods, such as Random and Farthest First, were compared. It was decided that four clusters would correspond to the number of nominals in the Severity Level class, which was where the clusters were to be evaluated. The algorithm categorises the data in terms of its likelihood of belonging to a specific cluster in EM.

3. Material and Methods

The focus of this work is to investigate the reliability of the dataset by employing supervised and unsupervised machine learning techniques to classify the symptoms experienced by COVID-19 patients according to the severity level recorded in the dataset. The methods considered are K-Nearest Neighbour (K-NN), Linear SVM, Naive Bayes, Decision Tree (J48), Ada Boost, Bagging, and Stacking for supervised machine learning approaches and Simple K-Means and Expectation-Maximization (EM) for unsupervised machine learning techniques.

In this paper, we used SPSS Statistical Software and Microsoft Excel to generate the summary statistics for the data set. WEKA data mining software is chosen for the machine learning algorithm due to its simplicity and free licensing. We summarised the data by selecting only the Severity Level and converted the categorical data with the symptoms as the nominal data. A trial-and-error approach using The simplified data sets employs K-Means and EM to cluster symptoms based on severity. The data is converted into nominal string values before the unsupervised learning algorithm in WEKA.

WEKA can conduct clustering on nominal data, whereas Simple K-Means is usually only relevant to numerical data. The distance function is Euclidean, and the comparison is made using both Random and Farthest First initialisation approaches. The number of clusters was increased to four to correspond to the number of nominals in the Severity Level class, which was used to evaluate the clusters. In EM, the algorithm classifies data according to their likelihood of belonging to a particular cluster.

All the analyses were carried out using a laptop with an Intel Core i5 processor and 6 Gigabytes of RAM, which caused some limitations in analysing the data.

3.1 Dataset

The dataset used in this work was sourced from the Kaggle dataset retrieved from [28]. The information was gathered through a survey of people of diverse genders and ages who had visited at least ten countries. There are four levels of severity based on the COVID-19 symptom, which was developed following the World Health Organization (WHO) and the Indian Ministry of Health and Family Welfare recommendations. These recommendations are then used as the guidelines for the survey.

3.1.1 The Dataset Preparation

This dataset comprises a total of 316,800 data combinations with 17 attributes of demographic information that determine the classes: country, contact, age group, gender, severity level, and the symptoms themselves. Fever, weariness, dry cough, difficulty breathing, sore throat, aches, nasal congestion, runny nose, diarrhoea, and no symptoms are among the COVID-19 symptoms listed in the dataset. Gender, age, severity, and country groups have been incorporated and substituted into the raw dataset to accommodate the restrictions of the classifier used to create the dataset. Furthermore, the numerical value for the severity group and country have been converted to a nominal value. Table 1 shows the dataset's information categorised into gender, age group, close contact, severity level, and country.

Table 1 - Dataset information

	Replaced with	Description (if any)
Gender		
Female	1	
Male	2	
Transgender	3	
Age Group		
0-9	1	Child
10-19	2	Teens
20-24	3	Adolescent
25-59	4	Adult
>60	5	Senior citizen
Close Contact		
No	1	Close contact with positive patient
Yes	2	No close contact with positive patient
Do not Know	3	Do not know
Severity Level		
None	1	
Mild	2	
Moderate	3	
Severe	4	
Country		
China	1	
France	2	
Germany	3	
Iran	4	
Italy	5	
Republic of Korean	6	
Spain	7	
UAE	8	
Other-EUR	9	
Other	10	

3.1.2 Summary Statistic

Using Microsoft Excel's summary descriptive tool, we obtained the summary statistic displayed in Table 2 for four distinct degrees of COVID-19 severity. The results show a balance of data distribution among all severity categories in which the authenticity of the data can be questionable.

Table 2 - Summary statistic by severity

COVID 19 Severity Summary Statistic				
	Mild	**Moderate**	**Severe**	**None**
Mean	4.1591	4.1591	4.1591	4.1591
Standard Error	0.0057	0.0057	0.0057	0.0057
Standard Deviation	1.5982	1.5982	1.5982	1.5982
Sample Variance	2.5543	2.5543	2.5543	2.5543
Range	7	7	7	7
Minimum	2	2	2	2
Maximum	9	9	9	9
Sum	329400	329400	329400	329400
Count	79200	79200	79200	79200

Table 3 and Table 4 show the result of the dataset in nominal form as generated using SPSS software. Table 3 shows the COVID-19 symptoms in binary form, and Table 4 presents demographic information in a categorical form, summarised in the frequency column. The frequency column reveals that each variable has a frequency divisible by 100.

Table 3 - Summary statistic by COVID-19 symptoms

Variable	**Value**	**Frequency**	**Per cent**	**Valid Percent**	**Cumulative Percent**
Fever	no	217800	68.8	68.8	68.8
	yes	99000	31.3	31.3	100
Tiredness	no	158400	50	50	50
	yes	158400	50	50	100
Dry Cough	no	138600	43.8	43.8	43.8
	yes	178200	56.3	56.3	100
Difficulty in Breathing	no	158400	50	50	50
	yes	158400	50	50	100
Sore Throat	no	217800	68.8	68.8	68.8
	yes	99000	31.3	31.3	100
No Symptom	no	297000	93.8	93.8	93.8
	yes	19800	6.3	6.3	100
Pains	no	297000	93.8	93.8	93.8
	yes	19800	6.3	6.3	100
Nasal Congestion	no	144000	45.5	45.5	45.5
	yes	172800	54.5	54.5	100
Runny Nose	no	144000	45.5	45.5	45.5
	yes	172800	54.5	54.5	100
Diarrhoea	no	201600	63.6	63.6	63.6
	yes	115200	36.4	36.4	100
Runny Nose	no	144000	45.5	45.5	45.5
	yes	172800	54.5	54.5	100
None Experiencing	no	288000	90.9	90.9	90.9
	yes	28800	9.1	9.1	100

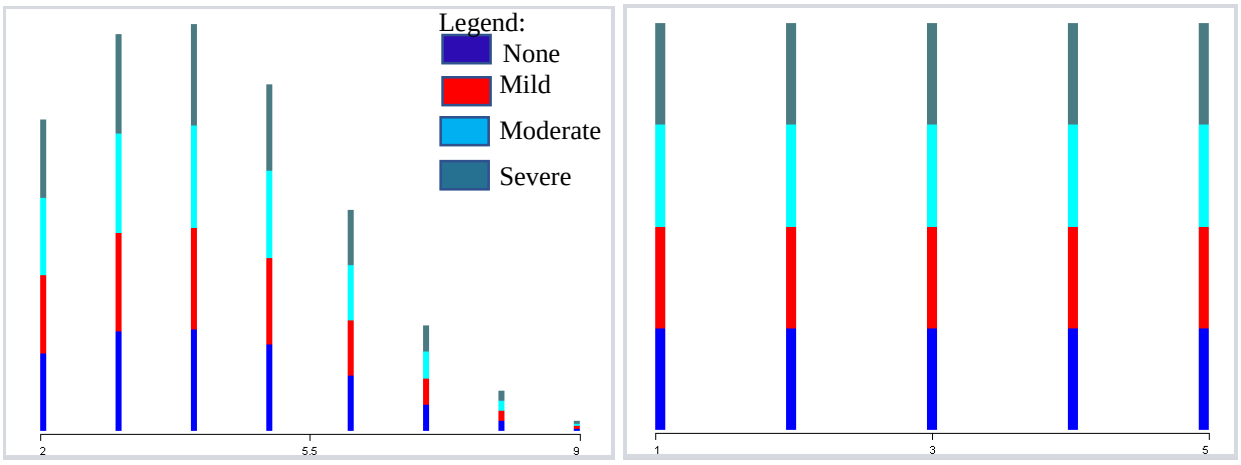
Table 4 - Summary statistic of demographic information

Variable	Value	Frequency	Per cent	Valid Percent	Cumulative Percent
Age Group	Adolescent	63360	20	20	20
	Adult	63360	20	20	40
	Child	63360	20	20	60
	Senior citizen	63360	20	20	80
	Teens	63360	20	20	100
Gender	Female	105600	33.3	33.3	33.3
	Male	105600	33.3	33.3	66.7
	Transgender	105600	33.3	33.3	100
Severity Level	Mild	79200	25	25	25
	Moderate	79200	25	25	50
	None	79200	25	25	75
	Severe	79200	25	25	100
Contact	Do not Know	105600	33.3	33.3	33.3
	No	105600	33.3	33.3	66.7
	Yes	105600	33.3	33.3	100

WEKA was used to create a visual representation of the dataset. Table 5 shows the summary statistic of the presented dataset. The result shows that the mean and standard deviation equal 4. 159 and 1.598, respectively. Fig. 1 and Fig. 2 depict the experiment's outcome. Because there are no missing data in this dataset, it is possible to use it without data interpolation, which is necessary when the missing data can lead to biased estimation and erroneous conclusions.

Table 5 - WEKA Tool's summary statistic

Statistic	Value
Minimum	2
Maximum	9
Mean	4.159
Standard Deviation	1.598



(a) Symptoms Sum (b) Age group

Fig. 1 - WEKA Visualization (a) symptoms sum; (b) age group where the x-axis is the bin, and the y-axis is the frequency

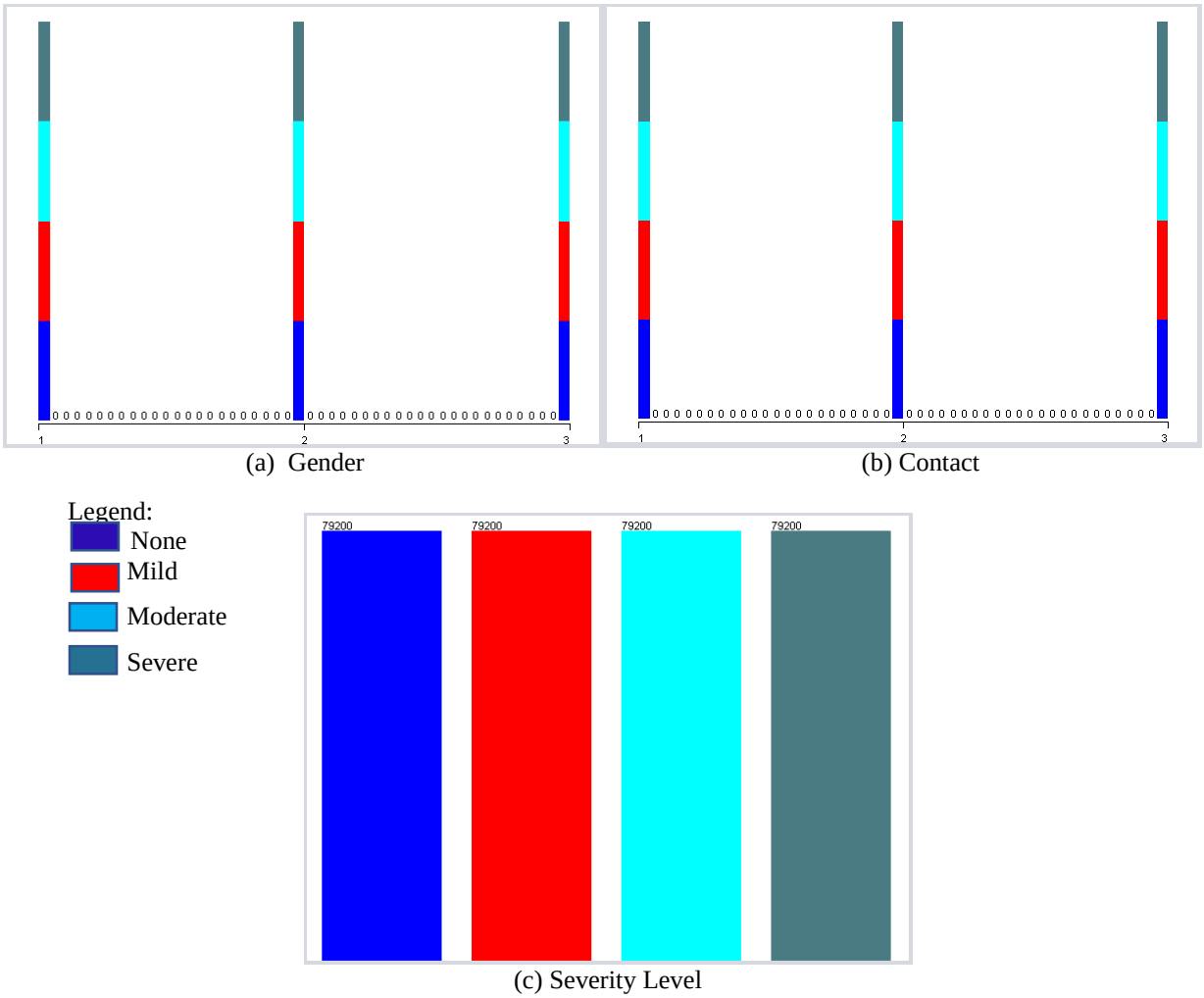


Fig. 2 - WEKA Visualization on the dataset based on (a) gender; (b) contact; (c) severity level where the x-axis is the bin, and the y-axis is the frequency

Based on the early observation result from Microsoft Excel, SPSS and Weka Summary Statistic and visualisation, the data for severity level, symptoms, and other variables are equally distributed. We examined this dataset again using two different types of machine learning: unsupervised and supervised learning methods.

3.2 Performance Evaluation

For the sake of this study, all classifiers will be fitted with $K = 3, 5, 10$ - fold validations. The number of k determines the size differences between the training set and the resampling subsets. If k is reduced to a smaller value, the size difference will expand in magnitude, and vice versa. In predicting a model, two error types can be defined: False Positives (Type 1) and False negatives (Type 2). Type 1 error occurred when an event was predicted, but there was no event. Meanwhile, the Type 2 error occurs when no event is predicted, but one occurs.

The accuracy, A_c of a machine learning model can be determined as [29]

$$A_c = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. True Positives is determined as a positive outcome prediction with a positive outcome result. In contrast, True Negatives produced negative outcome results as predicted. A better degree of accuracy can be reached when the dataset is good.

The True Positives Rate TP_R, also known as Recall or sensitivity, is the proportion of adequately classified positive data items in comparison to all positive data points, which is given as

$$TP_R = \frac{TP}{TP + FN} \tag{3}$$

The False Positives Rate, FP_R, is the fraction of negative data points that are incorrectly interpreted as positive, in comparison to all negative data points and can be determined as

$$FP_R, = \frac{FP}{FP + TN}. \tag{4}$$

Another vital performance evaluation considers in this paper is the precision, τ^2. It is defined as the proportion of correct positive findings to those predicted by the classifier.

$$\tau^2 = \frac{TP}{TP + FP} \tag{5}$$

The balance between precision and Recall can be measured through F-Measure or F1-Score. It indicates the precision and robustness of the classifier and is measured in the range of [0 1]. The higher the F1 Score, the better our model's performance. Mathematically, it can be expressed as :

$$F1_m = 2\frac{\tau^2 \times TP_R}{\tau^2 + TP_R} \tag{6}$$

Matthews Correlation Coefficient (MCC) is another way in binary classification and can be expressed mathematically as

$$\varphi_{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

The MCC value ranges from -1 to 1, with 0 indicating that the classifier is no better than a random coin flip. When MCC is equal to one ($\varphi_{MCC} = 1$), the classifier is perfect and positively correlated. Conversely, the classification is misclassified when $\varphi_{MCC} = -1$.

Fig. 3 shows the relationship between TP_R and FP_R are depicted as a line on a graph. It is represented as the Receiver Operator Characteristic (ROC) curve and the Area Under Curve (AUC) represents the balance between the two. It is preferable if TP_R is higher and FP_R is lower for each threshold than the other way around. Therefore, classifiers with graph curves that are prone on the top left side outperform those with curves that are prone on the bottom right side of the graph curve. On the other hand, the Precision-Recall Curve (PRC) Area is the area below the PRC when the precision and recall measure is plotted at the different thresholds.

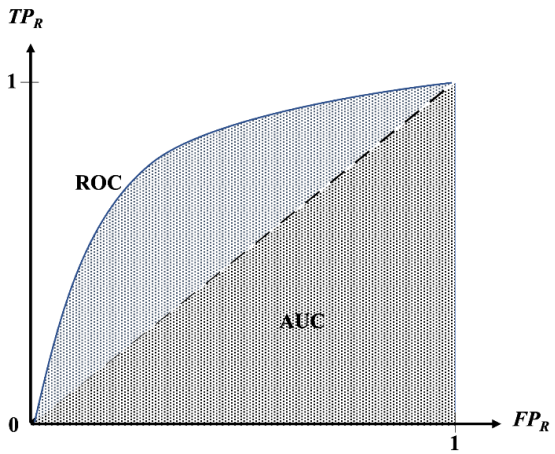


Fig. 3 - ROC curve and AUC

4. Result and Discussion

This section explains the classifier performance results and the accuracy of supervised machine learning algorithms used to categorise the severity group of COVID-19 symptoms. The severity group was classified using supervised learning techniques such as K-NN, Naive Bayes, Decision Tree (J48), Ada Boost, Bagging, and Stacking.

Tables 6, 7, and 8 show the results of each classifier's performance when K-fold validation is performed with $K = 3, 5,$ and 10, respectively. All attributes have a similar amount of data, so no attributes are excluded from consideration in this project. The highest precision classifier is AdaBoost with 0.248 when $K = 5$. It is discovered that the majority of classifier performance measures, including TP Rate, FP Rate, Precision, and Recall, are less than 0.5, indicating that the proposed classifier performs poorly.

Table 6 - Classifier results when $K = 3$

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
K Nearest Neighbour (with Euclidian Distance)	0.009	0.33	0.008	0.009	0.007	-0.323	0.102	0.181
Naïve Bayes	0.244	0.252	0.244	0.244	0.244	-0.008	0.494	0.246
Decision Tree (J48)	0.206	0.265	0.206	0.206	0.206	-0.059	0.446	0.219
Ada Boost (with classifier Decision Stump)	0.248	0.251	0.247	0.248	0.228	-0.003	0.497	0.249
Bagging (with REPTree Classifier)	0.048	0.317	0.048	0.048	0.048	-0.269	0.218	0.157

Table 7 - Classifier results when $K = 5$

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
K Nearest Neighbour (with Euclidian Distance)	0.002	0.333	0.001	0.002	0.001	-0.334	0.043	0.186
Naïve Bayes	0.243	0.252	0.243	0.243	0.243	-0.009	0.491	0.245
Decision Tree (J48)	0.225	0.258	0.225	0.225	0.225	-0.033	0.471	0.233
Ada Boost (with classifier Decision Stump)	0.248	0.251	0.248	0.248	0.241	-0.003	0.497	0.248
Bagging (with REPTree Classifier)	0.029	0.324	0.029	0.029	0.029	-0.295	0.159	0.149

Table 8 - Classifier results when $K = 10$

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
K Nearest Neighbour (with Euclidian Distance)	0	0.333	0	0	0	-0.34	0.012	0.203
Naïve Bayes	0.24	0.253	0.24	0.24	0.24	-0.013	0.488	0.243
Decision Tree (J48)	0.239	0.254	0.239	0.239	0.238	-0.014	0.487	0.242
Ada Boost (with classifier Decision Stump)	0.246	0.251	0.246	0.246	0.236	-0.005	0.495	0.247
Bagging (with REPTree Classifier)	0.019	0.327	0.019	0.019	0.019	-0.308	0.114	0.144

Fig. 4 shows the ROC area for classifying COVID-19 severity from the dataset for various K-fold values. It shows that when $K = 10$, K-NN, Naïve Bayes and Bagging classifier performance is reduced compared to the J48 classifier and no changes for the Ada Boost classifier. The KNN and Bagging classifiers show the lowest capability in distinguishing between the severity class, which means that the model reciprocates the classes. The remaining classifiers have ROC areas close to 0.5, indicating that the model cannot discriminate between positive and negative classes.

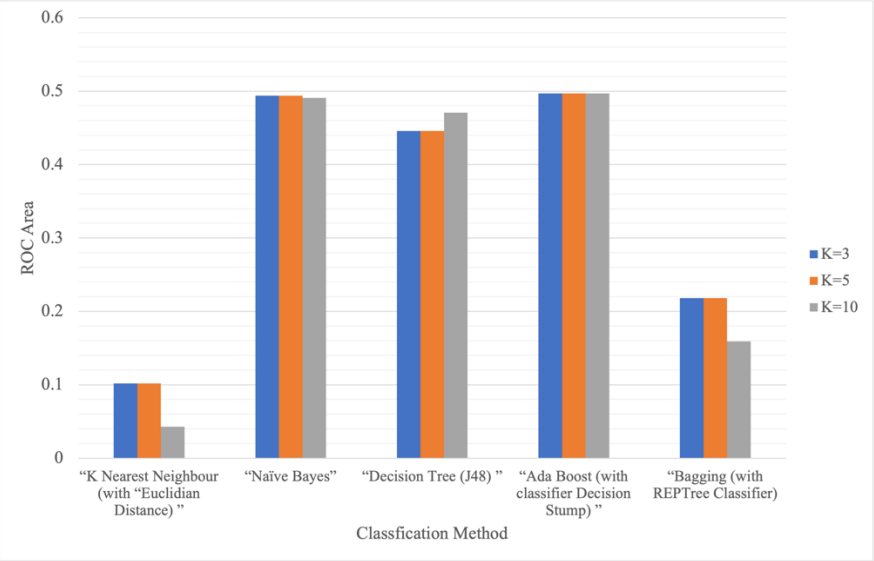


Fig. 4 - The severity group of COVID-19 ROC area for various classification methods with $K = 3, 5, 10$

Table 9 tabulated the results of the classifiers' accuracy in predicting the severity classes for $K = 3, 5, 10$. Except for classifier Linear SVM and Stacking, the classifier's accuracy obtained less than 25%, with the highest accuracy only reaching 24.763% when using classifier "Ensemble Method Ada Boost (with classifier Decision Stump)" when $K = 10$. The KNN is the worst accuracy with only 0.022% when $K = 10$. The accuracy result performance is expected as the TP rate should be nearest to 1.0 to consider it the best value. The performance result shows that the supervised machine learning techniques proposed could not classify the severity class from the dataset.

Table 9 - Classifier accuracy when $K = 3, 5, 10$

Classifier	K=3	K=5	K=10
K Nearest Neighbour (with Euclidian Distance)	0.911%	0.200%	0.022%
Naïve Bayes	24.428%	24.302%	23.996%
Decision Tree (J48)	20.594%	22.538%	23.930%
Ada Boost (with classifier Decision Stump)	24.763%	24.757%	24.605%
Bagging (with REPTree Classifier)	4.801%	2.870%	1.870%

This dataset is then further investigated using the Simple K-Means clustering approach. The Simple K-Means algorithm in WEKA categorises the nominal data using mode instead of means. The result from this algorithm varies depending on the initialisation method used. Table 10 compares the binary result between Random, Farthest First, and K-Means++ for various COVID-19 symptoms, categorised into four clusters: mild, moderate, severe and none. The result shows that the Farthest First initialisation is closer to the standard categorisation for the severity level of COVID-19 disease based on the symptoms.

Table 10 - Simple K-means category result

Initialisation Method	Symptoms	Cluster 0 (Mild)	Cluster 1 (Moderate)	Cluster 2 (Severe)	Cluster 3 (None)
Random	Fever	no	no	no	yes
	Tiredness	yes	no	no	yes
	Dry-Cough	yes	yes	no	yes
	Difficulty-in-Breathing	no	yes	yes	yes
	Sore-Throat	no	no	yes	yes
	None_Sympton	no	no	no	no
	Pains	no	yes	no	yes
	Nasal-Congestion	no	yes	yes	no
	Runny-Nose	yes	yes	no	no

	Diarrhoea	yes	no	no	no
	None_Experiencing	no	no	no	no
Farthest First	Fever	no	yes	yes	no
	Tiredness	no	yes	yes	no
	Dry-Cough	yes	no	yes	no
	Difficulty-in-Breathing	yes	no	yes	no
	Sore-Throat	no	no	yes	no
	None_Sympton	no	no	no	no
	Pains	no	yes	yes	no
	Nasal-Congestion	yes	yes	no	no
	Runny-Nose	yes	yes	no	no
	Diarrhoea	no	yes	no	no
	None_Experiencing	no	no	no	no
K-Means++	Fever	yes	no	no	no
	Tiredness	yes	no	no	no
	Dry-Cough	yes	yes	no	yes
	Difficulty-in-Breathing	no	yes	no	yes
	Sore-Throat	no	no	no	yes
	None_Sympton	no	no	no	no
	Pains	no	no	yes	no
	Nasal-Congestion	no	yes	yes	no
	Runny-Nose	yes	no	yes	yes
	Diarrhoea	no	no	yes	yes
	None_Experiencing	no	no	no	no

There were 237600 (75%) erroneously clustered instances in the runs for the three initialisation techniques with a maximum of 500 iterations and varying symptoms. Table 11 shows the percentage of total data assigned to the clusters from the three initialisation methods. 50% of samples were classified as having a Mild severity level using the Farthest First initialisation. It shows that most COVID-19 patients fall under Cluster 1 and 2 compared to Cluster 3, with an average of 16%.

Table 12 displays the outcome of the WEKA EM clustering for various symptoms, which is expressed as a percentage. The result shows that Tiredness, Dry-Cough and Difficult in Breathing were classified for Moderate; Nasal-Congestion and Runny-Nose were classified as Severe symptoms. However, all these symptoms are also classified as None, and it is discovered that 75% of occurrences were erroneously classified using Expectation-Maximisation clustering.

Table 11 - Simple K-means clustering percentages

Initialisation Method	**Cluster 0 (Mild)**	**Cluster 1 (Moderate)**	**Cluster 2 (Severe)**	**Cluster 3 (None)**
Random	37%	33%	17%	13%
Farthest First	50%	18%	13%	20%
K-Means++	33%	34%	19%	14%

Table 12 - WEKA expectation-maximization clustering

	Cluster	**0 (Mild)**	**1 (Moderate)**	**2 (Severe)**	**3 (None)**
Symptoms	**% of Total**	**17%**	**26%**	**26%**	**31%**
	Value / Total	**55317.7926**	**80867.5231**	**83413.451**	**97209.23**
Fever	yes	29%	33%	29%	33%
	no	71%	67%	71%	67%
Tiredness	yes	29%	67%	29%	67%
	no	71%	33%	71%	33%
Dry-Cough	yes	0%	100%	0%	100%
	no	100%	0%	100%	0%
	yes	29%	67%	29%	67%

Difficulty-in-Breathing	no	71%	33%	71%	33%
Sore-Throat	yes	29%	33%	29%	33%
	no	71%	67%	71%	67%
None_Sympton	yes	14%	0%	14%	0%
	no	86%	100%	86%	100%
Pains	yes	17%	20%	49%	50%
	no	83%	80%	51%	50%
Nasal-Congestion	yes	0%	0%	91%	100%
	no	100%	100%	9%	0%
Runny-Nose	yes	39%	40%	65%	67%
	no	61%	60%	35%	33%
Diarrhoea	yes	42%	40%	33%	33%
	no	58%	60%	67%	67%
None_Experiencing	yes	23%	20%	0%	0%
	no	77%	80%	100%	100%

The classification result for the data set selected in this study does not provide the correct result for the symptoms categorised against the severity level. First, it is noted that the total number of each value in the samples was all divisible by 100, and even numbers suggest that the data may be synthetic and not sourced from the real world.

Compared to the data source in a study by [30], in which the data was collected by the Center for Disease Control and Prevention of Henan Province (Henan CDC) within a time frame, there was no indicator of an official source for the selected data in this study. The data processing used in the study by Li is also described by frequency and percentage; however, there are no continuous variables for the data in this study.

A study by [31] and the Elbow method [32] were used to determine the good clusters out of the clustering model consisting of 155 countries and territories. Their techniques differed from our clustering, which was based only on the number of nominal values in the categorical class Severity Level. WEKA allows for the model to be evaluated against a categorical class. The same method can better fit the symptoms into clusters from a more significant number of clusters for better optimisation.

The preliminary data processing performed on the data set in this study is also unknown. In the previous study, such as by [33], the exact way the data was recorded is known. However, this may be a limitation of the secondary data source. It may be more beneficial for reliability to link primary and secondary data [34] with open-source data as the secondary data.

On the other hand, the Furthest First initialisation method gives a more distinct difference between the None and Severe severity levels for the type of data set. This initialisation approach alters the K-Means algorithm by selecting the initial centre before computing the distance between it and all other possible centres.[35].

However, the high count (75%) of incorrectly clustered both in Simple K-Means and EM, as shown in Table 12, shows poor performance for the algorithms in classifying the selected data set to severity level, suggesting the data is simulated.

5. Conclusion

In conclusion, the purpose of this study is to apply the clustering and classification method in machine learning to discover the hidden pattern in the dataset, thus generating valuable knowledge on the COVID-19 outbreak. The dataset used in this study comprises a total of 316,800 data combined with 17 attributes of demographic information that determine the classes: country, contact, age group, gender, severity level, and the symptoms themselves. Based on the results obtained, the analysis using supervised and unsupervised machine learning approaches has yielded relatively poor classification and clustering results. The findings for the dataset analysed in this study do not appear to provide the correct result for the symptoms categorised against the severity level, raising concerns about the dataset's validity and reliability. While conducting this research, we could not locate any previous work that had been done on the same data.

Lastly, machine learning is crucial for scientists and healthcare practitioners, especially during the pandemic, to learn the hidden pattern and insightful knowledge on how the outbreak spread and how to contain it. For future work, it is recommended to apply the clustering and classification method using the COVID-19 outbreak dataset obtained from the local hospital or COVID-19 Assessment Centre (CAC), as the symptom may vary due to virus mutations and other external factors such as the economy.

Acknowledgement

The work in this paper is supported by Universiti Teknologi Malaysia and The Ministry of Higher Education Malaysia.

References

[1] Wang, C., Wang, Z., Wang, G., Lau, J., Zhang, K., Li, W. (2021, March 08). COVID-19 in early 2021: Current status and looking forward. Retrieved February 20, 2022, from https://www.nature.com/articles/s41392-021-00527-1

[2] Johns Hopkins Coronavirus Resource Center. (2000). COVID-19 map. Retrieved July 8, 2021, from https://coronavirus.jhu.edu/map.html

[3] Morens, D. M., Breman, J. G., Calisher, C. H., Doherty, P. C., Hahn, B. H., Keusch, G. T., Taubenberger, J. K. (2020). The origin of COVID-19 and why it matters. The American Journal of Tropical Medicine and Hygiene, 103(3), 955-959. doi:10.4269/ajtmh.20-0849

[4] Morens, D. M., Daszak, P., Taubenberger, J. K. (2020). Escaping pandora's box — another novel coronavirus. New England Journal of Medicine, 382(14), 1293-1295. doi:10.1056/nejmp2002106

[5] Mahase, E. (2020). COVID-19: Who declares pandemic because of "alarming levels" of spread, severity, and inaction. BMJ, M1036. doi:10.1136/bmj.m1036

[6] Dong, E., Du, H., Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in Real time. The Lancet Infectious Diseases, 20(5), 533-534. doi:10.1016/s1473-3099(20)30120-1

[7] Rocklöv, J. and Sjödin, H. (2020). High population densities catalyse the spread of COVID-19. Journal of Travel Medicine, 27(3). doi:10.1093/jtm/taaa038

[8] Almaghaslah, D., Kandasamy, G., Almanasef, M., Vasudevan, R., Chandramohan, S. (2020). Review on the coronavirus disease (COVID-19) pandemic: Its outbreak and current status. International Journal of Clinical Practice, 74(11). doi:10.1111/ijcp.13637

[9] Elliott J. H. and Jeppesen, B. T. (2021). Rapid and living guidance for COVID-19. Annals of Internal Medicine, 174(8), 1171-1172. doi:10.7326/m21-2245

[10] Dargan, S., Kumar, M., Ayyagari, M. R., Kumar, G. (2019). A survey of Deep Learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering, 27(4), 1071-1092. doi:10.1007/s11831-019-09344-w

[11] Malki, Z., Atlam, E., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons and Fractals, 138, 110137. doi: 10.1016/j.chaos.2020.110137

[12] Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-qays, Z., Zaidan, A. A., Zaidan, B. B., Madhloom, H. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. Journal of Medical Systems, 44(7). doi:10.1007/s10916-020-01582-x

[13] Zagrouba, R., Adnan Khan, M., Atta-ur-Rahman, Aamer Saleem, M., Faheem Mushtaq, M., Rehman, A., Farhan Khan, M. (2021). Modelling and simulation of COVID-19 outbreak prediction using supervised machine learning. Computers, Materials & Continua, 66(3), 2397-2407. doi:10.32604/cmc.2021.014042.

[14] Kumar, A., Sharma, A., Arora, A. (2019). Anxious depression prediction in real-time social data. SSRN Electronic Journal. doi:10.2139/ssrn.3383359.

[15] Lalmuanawma, S., Hussain, J., Chhakchhuak, L. (2020). Applications of machine learning and Artificial Intelligence for COVID-19 (SARS-COV-2) pandemic: A Review. Chaos, Solitons & Fractals, 139, 110059. doi: 10.1016/j.chaos.2020.110059

[16] Frank E, Hall MA, and Witten IH. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[17] Chowdhury M. E., Rahman, T., Khandakar, A., Al-Madeed, S., Zughaier, S. M., Doi, S. A., Islam, M. T. (2021). An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. Cognitive Computation. DOI:10.1007/s12559-020-09812-7

[18] Han J, Kamber M, and Pei. J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

[19] Cheng D, Zhang S, Deng Z, Zhu Y., Zong M. (2014). kNN Algorithm with Data-Driven k Value. International Conference on Advanced Data Mining and Applications, pp 499-512. DOI:10.1007/978-3-319-14717-8_39.

[20] Ali, A. H., and Abdullah, M. Z. (2020). An Efficient Model for Data Classification Based on SVM Grid Parameter Optimization and PSO Feature Weight Selection. *International Journal of Integrated Engineering*, *12*(1), 1-12

[21] Nai-Arun N and Sittidech P. (2014). The ensemble learning model for diabetes classification. Advanced Materials Research, 931-932, 1427-1431. doi:10.4028/www.scientific.net/amr.931-932.1427

[22] Saravanan N. and Gayathri, V. (2018). Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48). International Journal of Computer Trends and Technology, 59(2), 73-80. doi:10.14445/22312803/ijctt-v59p112
[23] The Ultimate Guide to AdaBoost algorithm: What is AdaBoost algorithm? G. (2022, January 13). The Ultimate Guide to AdaBoost algorithm: What is AdaBoost algorithm? Retrieved February 10, 2022, from https://www.mygreatlearning.com/blog/adaboost-algorithm/
[24] Rocca, J. (2021, March 21). Ensemble methods: Bagging, boosting and stacking. Retrieved February 15, 2022, https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205
[25] Brownlee, J. (2021, April 26). Stacking Ensemble Machine Learning with python. Retrieved February 15, 2022, from https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/
[26] Yuan C. and Yang H., "Research on K-value selection method of K-means clustering algorithm," J, vol. 2, no. 2, pp. 226-235, 2019.
[27] Suljevic-Pasic A. and Yaman E., "Categorising Stars with Known Properties Using the Expectation-Maximization Clustering Algorithm," Southeast Europe Journal of Soft Computing, vol. 6, no. 2, 2018.
[28] Hungund, B. (2020, March 21). COVID-19 symptoms checker. Retrieved June 5, 2021, from https://www.kaggle.com/iamhungundji/COVID19-symptoms-checker
[29] Kuhn, M. and Johnson, K. (2019). Applied predictive modelling. New York: Springer, Vol. 26, p. 13.
[30] Li J, Chen Z, Nie Y, Ma Y, Guo Q, Dai X. Identification of Symptoms Prognostic of COVID-19 Severity: Multivariate Data Analysis of a Case Series in Henan Province. Journal of Medical Internet Research. 2020;22(6):e19636.
[31] Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Research. 2020;5.
[32] Marutho D, Handaka SH, Wijaya E, Muljono, editors. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. 2018 International Seminar on Application for Technology of Information and Communication; 2018 21-22 Sept. 2018.
[33] Benito-León J, del Castillo MD, Estirado A, Ghosh R, Dubey S, Serrano JI. Using Unsupervised Machine Learning to Identify Age- and Sex-Independent Severity Subgroups Among Patients with COVID-19: Observational Longitudinal Study. J Med Internet Res. 2021;23(5):e25988.
[34] Druschke D, Arnold K, Heinrich L, Reichert J, Rüdiger M, Schmitt J. Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects. Das Gesundheitswesen. 2020;82(S 02):S108-S16.
[35] Fränti P, Sieranoja S. How much can k-means be improved by using better initialisation and repeats? Pattern Recognition. 2019;93:95-112.