International Journal of Advanced Technology and Engineering Exploration, Vol 10(103)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101228

Research Article

Performance evaluation of classifiers for the COVID-19 symptom-based dataset using different feature selection methods

Fauzan Iliya Khalid[1*], Mokhairi Makhtar[2], Rosaida Rosly[3] and Aceng Sambas[2]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[1]
School of Computer Science, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia[2]
Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Terengganu, Malaysia[3]

Received: 12-March-2023; Revised: 19-June-2023; Accepted: 22-June-2023

Abstract
Classification algorithms are commonly employed in healthcare systems to aid decision support processes, such as treatment regimens, diagnosis, and illness prediction. The recent emergence of dominant variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), widely known as the coronavirus disease (COVID-19), has emphasized the significance of early detection for ensuring appropriate treatment and protecting unaffected populations. This study assesses the performance of various classification models on a COVID-19 dataset, utilizing two distinct feature selection methods: the wrapper method (WrapperSubsetEval) and the correlation-based feature subset evaluation (CfsSubsetEval). The effectiveness of these methods is evaluated based on the number of features selected for the reduced subset, execution time, and classifier accuracy. The experimentation is conducted using WEKA tools, and five different classifiers are selected for computation and comparison of accuracy: J48 decision tree (DT), support vector machine (SVM), naïve Bayes (NB), sequential minimal optimization (SMO), and k-nearest neighbor (KNN). The performance of each model is assessed using a 10-fold cross-validation technique, and the accuracy of the models is measured. The evaluation results, including comparisons before and after the implementation of the classification process and feature selection methods, indicate that KNN employing WrapperSubsetEval+KNN outperforms other algorithms, achieving the highest accuracy of 98.81%. In summary, the utilization of feature selection methods can be considered an effective approach for COVID-19 prediction.

Keywords
Classification, Machine learning, Feature selection, COVID-19.

1.Introduction
Recently, a global pandemic with high fatality rate has been affecting many lives and it is caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) known as coronavirus disease COVID-19. Numerous studies have been published in the last three years identifying the symptoms in COVID-19 patients in various countries around the world. The typical signs of an infection may include experiencing a high temperature, difficulties with breathing, a dry cough, a sore throat, feeling breathless, fatigue, loss of smell and taste, and other respiratory issues [1, 2].

As a result, having an independent predictive model that detects for COVID-19 in an individual is important. Creating a COVID-19 prediction model requires the utilization of fundamental resources such as machine learning software, datasets, and classification methods. The researchers in [3] utilized a supervised machine learning method to diagnose COVID-19 using the epidemiology dataset. It is mentioned that traditional machine learning such as naïve bayes (NB), logistic regression (LR), support vector machine (SVM), decision tree (DT) and artificial neural network models were developed to predict the COVID-19 infection.

Classification is a process in data mining that involves using machine learning algorithms to learn how to identify and group instances based on the input data provided. To create a model, a training dataset with various sets of inputs and outputs based on which the

*Author for correspondence

model learns is required for classification and prediction. The model created is then calculated to determine the method that can most effectively match the provided input data to a particular class label using the training dataset where the prediction result is later produced [4]. The findings of applying several classification algorithms to disease datasets for the diagnosis of chronic disease are highly encouraging, especially on COVID-19 disease. Therefore, an innovative classification technique that can speed up and make the diagnosis of chronic diseases simpler is urgently needed [5].

While machine learning techniques have proven to be effective, the quality of features derived from various algorithms significantly influences their performance. However, the creation of these features can be time-consuming and impractical. This is the primary flaw in the machine learning approach; despite its success, these techniques exhibit a significant performance reduction [6]. Clinical spectrum of SARS-CoV-2 infection in patients might range from no symptoms to critical illness. Thus, when performing COVID-19 diagnostics, a large amount of time is needed to carry out the analysis as it is very difficult to identify the presence of COVID-19. Reduced diagnostic time with increased accuracy is the main purpose of machine learning research for COVID-19 diagnosis because it allows for rapid treatment for the patients [7]. To tackle this issue, it is crucial to employ a feature selection technique that can extract the important and relevant features from the input data. The feature selection technique is one of the effective data preprocessing methods for lowering the complexity of the data. Thus, finding the most essential disease-related risk variables is crucial for medical diagnosis. This is important and necessary because feature recognition aids in the reduction of redundant and irrelevant attributes present in the dataset, which leads to quicker and better outcomes [5].

The ongoing COVID-19 pandemic has highlighted the need for accurate and efficient tools for identifying and tracking cases. In this context, supervised machine learning techniques have been proposed as a way to forecast COVID-19 cases based on symptoms. However, the literature on this topic is still evolving, and more research is needed to have a more robust prediction model.

It is also appearing that there are not many published studies that use COVID-19 symptom-based dataset as the input parameters. Moreover, the existing literature has mostly focused on comparing the performance of a single classifier or a limited number of classifiers using a single feature selection method on a COVID-19 dataset. There are relatively few studies that examine the performance of different classifiers on a COVID-19 dataset using various feature selection methods especially in WrapperSubsetEval feature selection method. Considering this, the research questions for this study are as follows:

Research Questions:

1. How do different feature selection methods impact the accuracy of classifiers when predicting COVID-19 cases?
2. What is the performance of different classifiers when applied to a COVID-19 symptom-based dataset using various feature selection methods?
3. Which classifier performs best when applied to a COVID-19 symptom-based dataset using different feature selection methods?

This research paper aims to evaluate the effectiveness of various feature selection methods, such as the wrapper method (WrapperSubsetEval) and correlation-based feature subset evaluation (CfsSubsetEval), on a COVID-19 dataset's classification models. The evaluation results show that K-Nearest-Neighbor, when using WrapperSubsetEval+KNN, outperforms other algorithms with an accuracy of 98.81%. This indicates that the use of feature selection methods can be considered an effective approach for predicting COVID-19. To summarize, this research emphasizes the significance of utilizing feature selection techniques to enhance the effectiveness of classification models when dealing with COVID-19 datasets.

The organization of this paper is presented in the following manner: in section 2, relevant literature and recent studies on feature selection methods are presented. Section 3 outlines the dataset used in the experiment, as well as the feature selection technique and diagram used to visualize the entire process. Section 4 details the results obtained from the experiment, while in section 5, the study's limitations and discussion are presented. Section 6 describes the conclusion drawn from the study and outlines suggestions for future research endeavours.

2.Literature review

Feature selection is a critical step in building accurate and efficient machine learning models. Several techniques for selecting features have been created in recent times to tackle high-dimensional datasets, such as those in medical applications. This literature review

aims to provide an overview of recent studies on feature selection methods, with a focus on their effectiveness in improving classification performance.

2.1Wrapper feature selection methods

One of the feature selection methods is the Wrapper Subset Evaluator algorithm but we found that less researches focusing on this algorithm. Thus, this paper proposes a potential to increase the performance of classifiers by incorporating the feature selection method. Wrapper Subset Evaluator algorithm utilizes a classifier to assess the performance of different subsets of features and select the subset that yields the highest classification accuracy [8]. This approach is computationally expensive, but it can provide better results compared to other feature selection methods [9]. Some other popular wrapper methods include recursive feature elimination (RFE)[10,11], genetic algorithms (GA)[12], and simulated annealing (SA)[13]. These methods have shown good performance in some applications, but their high computational cost is a limitation.

2.2Filter feature selection methods

Filter methods, on the other hand is one of the popular feature selection methods which include chi-squared[14,15], information gain[12,16], and correlation-based feature selection (CFS)[17–19]. The CFS algorithm, which has been shown to effectively reduce feature dimensionality while maintaining classification accuracy in medical datasets. For instance, in a recent study by [20], CFS was used to select features for the diagnosis of COVID-19 using chest, computed tomography (CT) scans. The findings revealed that the CFS-selected features obtained greater classification accuracy than utilising all of the features in the dataset. Filter methods are computationally efficient, but they may miss relevant features that have low individual relevance[21].

2.3Embedded and Hybrid Feature Selection Methods

There are other feature selection methods also including hybrid and embedded methods. Incorporating feature selection into the learning process of a model is a characteristic of embedded methods [22]. These methods optimize the model parameters and the feature subset simultaneously. Embedded methods have shown good performance in some applications but they may overfit the data in some cases. Hybrid methods [23, 24] combine different approaches such as wrapper and filter methods to improve the feature selection process. Wrapper-Filter approach was used, which uses a filter method to enable fast selection and then applies a wrapper method to provide accurate selection [25]. However, in [24] it was noted that the hybrid model, while offering improved feature selection, may be less accurate than other methods. This is because the filter and wrapper steps are performed separately, which can lead to a loss of information and reduced accuracy.

In summary, wrapper, filter, embedded and hybrid methods are widely used in feature selection. Each method has its own strengths and weaknesses, and the choice of the most appropriate method depends on the application and the available resources. The performance of each method may vary depending on the dataset and the algorithm used, so it is important to compare different methods and select the one that performs best for a particular task. In this experiment, WrapperSubsetEval, which is a wrapper method, and CfsSubsetEval, which is a filter method, are chosen to perform feature selection.

2.4Related work

Recent studies have explored the use of classification models and feature selection methods for predicting COVID-19. For instances, a study published in 2023 used metaheuristic optimization and artificial intelligence for optimal feature selection in COVID-19 detection with CT images, achieving a classification accuracy of 87.2% [26]. However, this study did not explore the use of different classifiers and their impact on accuracy. A different study [27] assessed 14 various feature selection approaches based on biochemical parameters for COVID-19 diagnosis. These approaches encompassed filter methods, spiral methods, and embedded methods. According to the results, the feature selection methods successfully reduced the initial set of 16 features to 5, resulting in the best performance for the KNN algorithm with an fsvFS score of 86.4%. The paper suggests that the use of artificial intelligence, and specifically machine learning, can be beneficial in the diagnosis process of COVID-19. Feature selection can lead to more meaningful results with less data by choosing the most meaningful parameters among those that affect the result. However, the effectiveness of the proposed methods would depend on the quality and quantity of the data used.

In another study [28], feature selection techniques were compared for predicting SARS-CoV-2 pneumonia severity prognosis using a COVID-19 dataset. The researchers used a range of feature selection techniques, including filters, wrappers, and embedded methods. The authors also considered

computation time as a factor, given that feature selection is often part of a larger machine learning pipeline. However, the dataset used had a marked class imbalance in terms of the number of patients by pneumonia severity class. This could potentially bias the results of the feature selection techniques. *Table 1* provides a summary of key characteristics from several studies, including the employed feature selection methods, classification models, results, advantages, and limitations, in the context of COVID-19 prediction and diagnosis.

Table 1 Review analysis based on feature selection and classification models applied to COVID-19 dataset

Ref.	Dataset	Feature selection	Classification models	Results	Advantages	Limitations
[29]	Chest CT images COVID-19	RFE as a wrapper feature selection and extra tree classifier as embedded feature selection	NB and restricted Boltzmann machine (RBM)	The RBM model achieved the highest accuracy of 99.924%	High accuracy with RBM technique, effective feature selection methods and extraction of optimal number of features from diverse data.	The study is limited by the feature selection and classification methods used, where other methods might yield different results.
[30]	COVID-19 symptom-based dataset	Chi-squared statistics and mutual-information statistics	SVM, DT, and neural network (NN) classifiers	The NN model achieved the highest accuracy of 97.08%	By utilizing feature selection, the accuracy results between the classifiers were not statistically significant, indicating comparable classification abilities. This implies that the statistical feature selection employed in the study makes machine learning easier by allowing the classifiers to effectively classify input data using the most relevant features.	The method requires a significant number of computational resources, which may not be feasible for some applications.
[31]	COVID-19 hospital-based registry data	Wrapper based feature selection	K-nearest neighbor (KNN), DT, multi-layer perceptron (MLP), and SVM	The DT model achieved the highest accuracy of 93.8%	The proposed model is capable of handling both categorical and continuous data, designed to handle missing data and capable of handling high-dimensional data, which is often a challenge in machine learning.	The limitations in the study includes low data quantity, non-optimal data quality, limited generalizability due to a single-center dataset with a small sample size, the use of only four machine learning algorithms for prediction analyses based on specific clinical features, and the absence of important para-clinical variables in the dataset.
[32]	COVID-19 clinical dataset	CfsSubsetEval, Chi-square and Information Gain	Bagging, C4.5 decision tree in WEKA (J48), LR, random forest (RF),	The bagging model achieved the highest	This research utilizes a recently published dataset from the Harvard Dataverse, which has not been employed in	The study also highlights the limitations of some of the machine learning models

Ref.	Dataset	Feature selection	Classification models	Results	Advantages	Limitations
			SVM, NB, and threshold selector	accuracy of 83.55%	any prior studies. Multiple experiments were conducted in this study, employing various feature selection techniques, with the aim of improving classifier performance and identifying the specific features accountable for these performance improvements.	used. For example, it mentions that the Bagging classifier made inaccurate predictions regarding the mortality outcome of 38 individuals. This indicates that while the models used in the study have high accuracy, they are not perfect and can still make incorrect predictions.
[33]	COVID-19 clinical dataset and Proteomics dataset	CfsSubsetEval, and InfoGainAttributeEval	J48, RF, IterativeClassifierOptimizer, AdaBoostM1, LogitBoost, BayesNet and sequential minimal optimization (SMO)	The random forest model achieved the highest accuracy of 89.47%	The study succeeded in identifying key clinical parameters and proteins that could be utilized to predict the prognosis of COVID-19 patients. These identified features can be evaluated as biomarkers that can help identify the patients who require immediate medical attention.	The imbalance in the number of survivors and deaths in both clinical and proteomics data could potentially affect the models' performance.
[34]	COVID-19 survey dataset	InformationGain and CfsSubsetEval	RF, NB, SVM, and LR	The LR model achieved the highest accuracy of 83.25%	The study employed validated measures to assess the outcomes, thereby increasing the reliability and validity of the results. The method used in this study has sped up the training process.	The study relied on self-reported measures, which could be subject to reporting bias.
[35]	COVID-19 patients from clinical text	A new hybrid feature selection using Improved Binary Flamingo Search Algorithm (IBFSA)	RF, MLP, Nearest Mean and SVM	The IBFSA with the SVM classifier model achieved the highest accuracy of 97.1119%	The IBFSA algorithm achieves better results than any other competing approaches in terms of feature selection accuracy.	The performance of IBFSA might be dependent on the quality and nature of the data.

3.Materials and methods

The dataset and methods used to forecast COVID-19 were discussed in this section. It describes COVID-19 experimental dataset, data preprocessing, feature selection, classification algorithms and prediction models include in this work. *Figure 1* illustrates the researchers' proposed system's workflow, which shows the preprocessing conducted on the dataset and the extracted several sets of relevant features by using the feature selection technique. The diagram outlines the steps involved in the experiment, which include loading the COVID-19 dataset, applying the feature selection method, training classification models using all available attributes and retraining on only the selected features, evaluating the models' performance, and determining the accuracy of the models for each feature selection method.

The diagram also includes an algorithm that outlines the steps involved in the experiment. The first step

involves loading the dataset. The second step is to preprocess the data. The third step is to apply feature selection methods to select a subset of relevant features from the dataset and select a classifier as a hyperparameter. The fourth step involves training the classification models by using a 10-fold cross-validation method on the selected features and evaluating their performance. The final step is to determine the accuracy result of the models for each feature selection method.

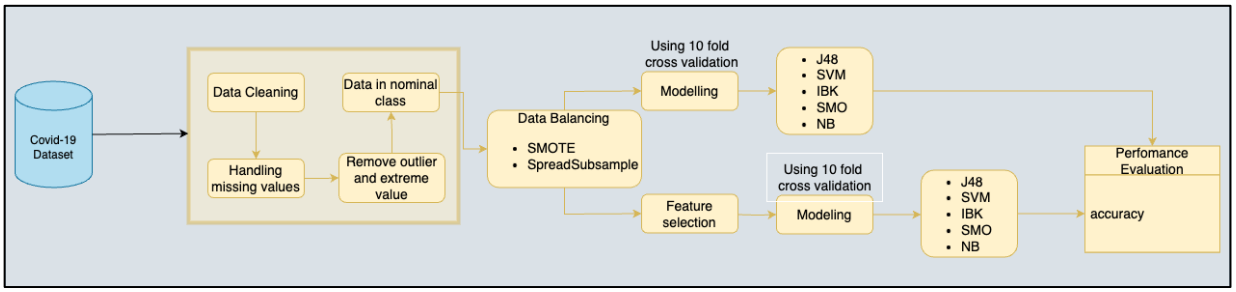

Figure 1 The block diagram of the proposed system

3.1Data collection

The dataset used is the COVID-19 dataset, which was collected from the Kaggle Website under the title "Symptoms and COVID Presence (May 2020 data)". The dataset was created by Hemanth Harikrishnan, based on WHO guidelines, in March 2020 in India. The COVID-19 dataset includes all potential symptoms, which should facilitate the prediction of whether COVID is likely to be present. There were 20 features of possible symptoms and one class attribute to determine the existence of COVID-19. The class label consists of Yes (COVID-19 presence) and No (No COVID-19 presence) as shown in *Figure 2*. This dataset comprises 5434 instances, some of which have missing values. As illustrated in *Table 2*, the list of attributes and the descriptions of the attributes. Each attribute is categorized either 'Yes' or 'No', where 'Yes' signifies the presence of the symptoms and 'No' indicates their absence.

Table 2 COVID-19 dataset information

Parameter	Type	Definition
Breathing Problem	Nominal	Shortness of breath is frequently linked with heart or lung conditions and can lead to breathing difficulties
Fever	Nominal	An abnormally high body temperature
Dry Cough	Nominal	Cough without expectoration
Sore throat	Nominal	Inflammation of the throat
Running Nose	Nominal	Experiencing runny nose
Asthma	Nominal	A chronic inflammatory disease of the lungs characterized by a narrowing of the airways
Chronic Lung	Nominal	A person has a lung disease
Headache	Nominal	Pain in the head
Heart Disease	Nominal	Aberrations in the heart's structure or function, as well as in the blood vessels that nourish it, can hinder its regular operation
Diabetes	Nominal	A broad expression utilized to denote health ailments distinguished by elevated blood sugar levels and frequent urination
Hyper Tension	Nominal	High blood pressure
Fatigue	Nominal	Physical or mental weariness resulting from effort or activity/ lack of energy
Gastrointestinal	Nominal	Medical conditions that impact any part of the gastrointestinal system, including the oesophagus, rectum, and other organs involved in digestion.
Abroad travel	Nominal	Travelling somewhere outside of the current country
Contact with COVID Patient	Nominal	To potentially contract an infection, an individual must have been in close proximity, within 6 feet, of an infected person for a minimum of 15 minutes, commencing from two days prior to the onset of symptoms, and continuing until the infected person is placed in isolation

Parameter	Type	Definition
Attended Large Gathering	Nominal	Mass gathering
Visited Public Exposed Places	Nominal	Going to public areas where the chances of exposure to COVID-19 disease are higher
Family working in Public Exposed Places	Nominal	Have a family member who lives together, working in public areas
Wearing Masks	Nominal	Wearing face masks such as surgical masks and cloth masks to avoid the contagion from infected people to others
Sanitization from Market	Nominal	The act or process of making something completely clean and free from bacteria
COVID-19	Nominal	COVID-19 presence

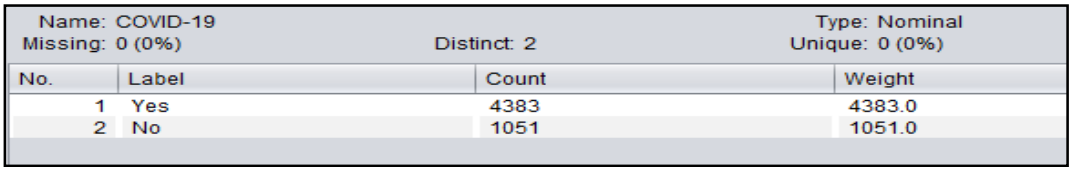

Figure 2 The class label of the COVID-19 dataset

Figure 3 depicted the class distribution for each attribute in the dataset. The *Figure 3* shows the proportion of instances in the dataset that are classified as "Yes" or "No" for each attribute. Each attribute represents a potential symptom of COVID-19, and the "Yes" or "No" classification indicates whether that symptom is present or not. For example, the first attribute shown in the figure is "Breathing Problem", and the figure shows that most instances in the dataset are classified as "Yes" for this attribute, indicating that most individuals in the dataset do have breathing problems.

3.2Data pre-processing

The initial phase of classifying the COVID-19 dataset involves implementing pre-processing techniques. Data pre-processing transforms raw data into an understandable format. The accuracy of data can be enhanced through pre-processing techniques [36]. Its aim is to handle missing values, remove outliers and extreme values, discretize data, and extract features [37]. When dealing with the missing data, the removal of instances with missing values is used to reduce bias. In this case, the missing instance under the attribute 'Hyper Tension' was removed using the 'ReplaceMissingValues' filter in WEKA tools, as shown in *Figure 4*.

Since all data type in this data set was nominal, there was no need for this data set to undergo data discretization. To check the outliers and extreme values in this data set, the filter InterquartileRange was applied as shown in *Figure 5*. This filter will add additional attributes that determine if the values of instances are deemed outliers or extreme values. The outlier was then removed but only with the class label equal to "Yes" (if any), and the same was done with the extreme value.

The class balance ratio for COVID-19 dataset with 4383 of Positive class and 1051 of Negative class is approximately 4:1, which is significantly imbalanced and it is possible that the model's performance could be affected. This means that the model is more likely to predict the positive class because it had seen more examples of it during training. This could lead to poor performance on the Negative class. The ways to improve the balance of the dataset are to undersample the majority class and oversample the minority class by using the SpreadSubsample filter and SMOTE (Synthetic Minority Over-sampling Technique) filter as shown in *Figure 6* and *Figure 7*. This filter could help further balance the dataset and ensure that the model is able to learn from both classes equally well, leading to better overall performance.

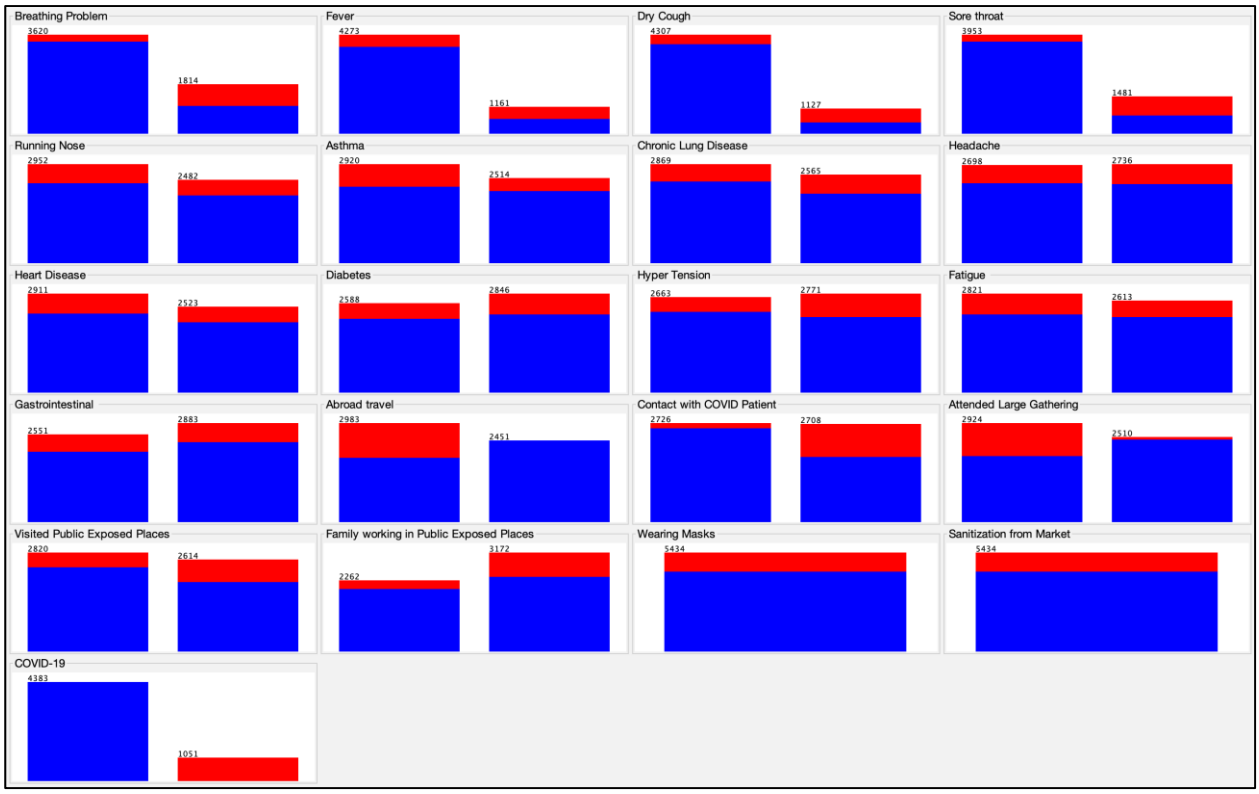

Figure 3 Class distribution of the dataset

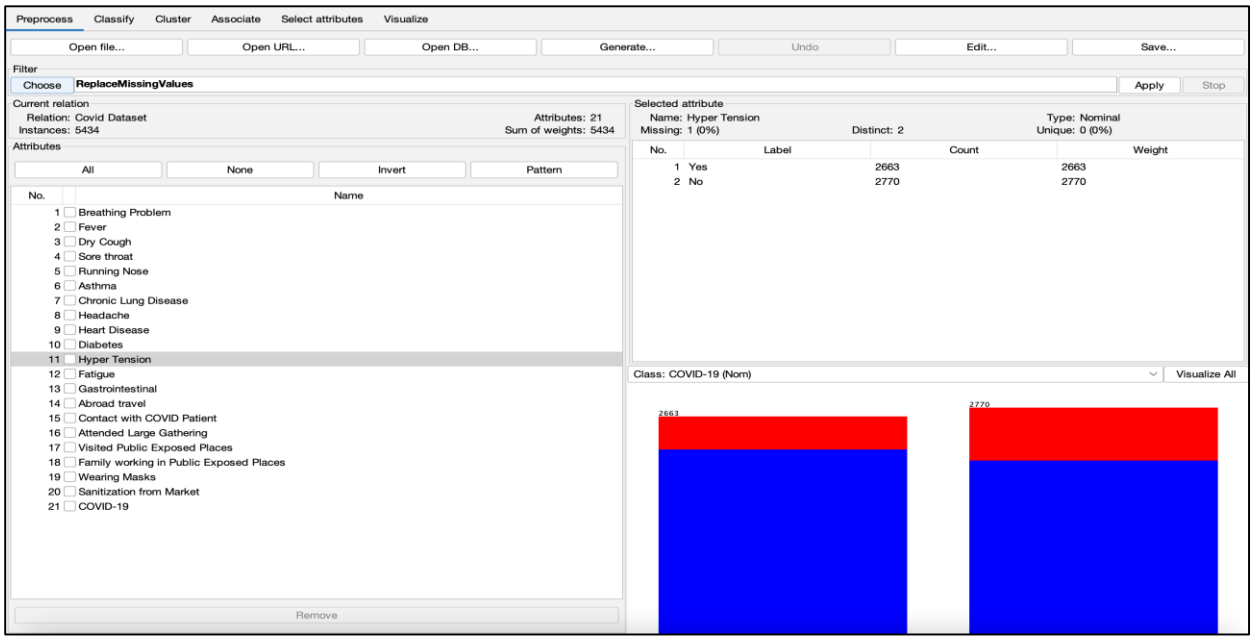

Figure 4 WEKA's Preprocess tab shows the instance with a missing value under the Hyper Tension attribute and a filter named “ReplaceMissingValues” was chosen to remove the missing value from the dataset

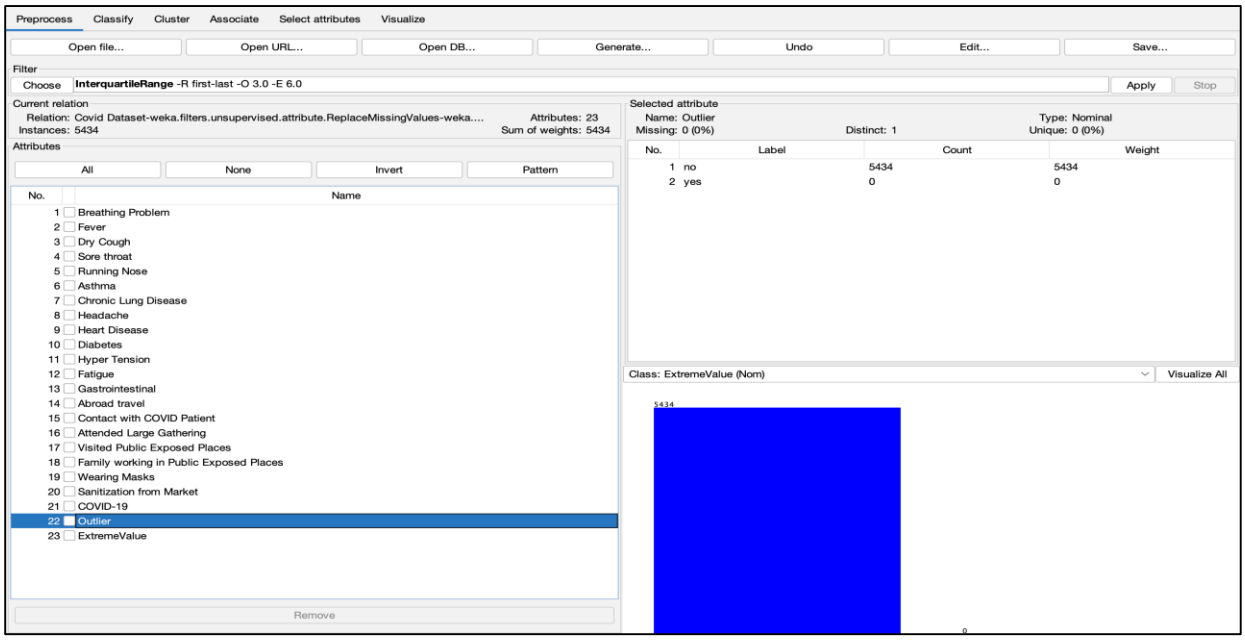

Figure 5 The preprocess tab of WEKA shows "InterquartileRange" filter used to handle outliers and extreme values which will provide additional attributes indicating if the values of instances are outliers or extreme values

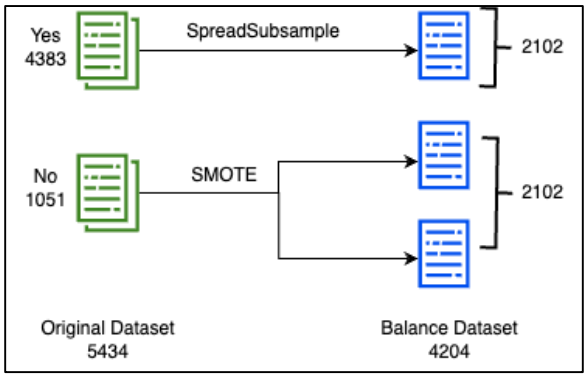

Figure 6 Dataset statistic before and after data balancing

3.3Feature selection

After data pre-processing was done, the task continued with selecting the attribute subsets. This process is called feature selection. In a data set, it contains relevant features which are prominent in contributing to high accuracy results, irrelevant features which will hurt the performance of the model with unnecessary data, and redundant features which are irrelevant in the presence of other features. Thus far, by applying the feature selection process, it will help in reducing overfitting of the data, improving the accuracy, and reducing the training time because less data means the training process is more efficient [38,39]. In WEKA, several feature selection methods are offered. Feature selection is divided into two components which are Attribute Evaluator and Search Method. In this experiment, two types of feature selection methods were chosen to test the data which are learner-based feature selection (WrapperSubsetEval) and correlation-based feature subset evaluation (CfsSubsetEval). The Search method applied in this experiment was BestFirst.

The WEKA machine learning library comprises the feature selection method, WrapperSubsetEval [40]. It is a "wrapper" method, which means that it uses a machine learning algorithm to evaluate the usefulness of each feature in the dataset. WrapperSubsetEval is a scheme-dependent attribute subset evaluator. It will create all possible subsets from the feature vector and will consider the subset of features with which the classification algorithms performed the best[41]. It uses a specific classifier to estimate the merit of a set of attributes. WrapperSubsetEval does the internal cross-validation technique to evaluate the classification accuracy of a specific group of attributes [42, 43]. It is generally considered to be more computationally expensive than other feature selection methods, but it can also be more effective at selecting the best features, because it considers the interaction between the features and the model. WrapperSubsetEval+classifier means that in these experiments, a feature selection method called WrapperSubsetEval was implemented to extract a subset of the dataset's features, and then a classifier

(such as J48 or KNN) was applied to that subset to evaluate its performance in terms of accuracy. The objective of this approach is to improve the classifier's effectiveness by identifying and extracting the most important and relevant features from the given dataset. So, in the WrapperSubsetEval method, the classifier serves as an evaluation criterion for identifying the optimal feature subset.

CfsSubsetEval on the other hand, is a "filter" method, which means that it uses a statistical measure to evaluate the significance of each feature in the dataset. It is a scheme-independent selection. It takes into consideration each attribute's predictive value as well as the degree of inter-redundancy [44]. Attribute sets that have a high level of correlation among their own attributes and a low level of correlation with the attributes of other sets are deemed to be suitable [43]. Filter methods are generally less computationally expensive than wrapper methods, but they may not be as effective at selecting the best features since they do not consider the interaction between the model and the features [45]. *Figure 8* shows how the feature selection process takes place in the WEKA Explorer Select attributes tab. After the best attributes were selected, the other attributes will be removed in the preprocess tab.

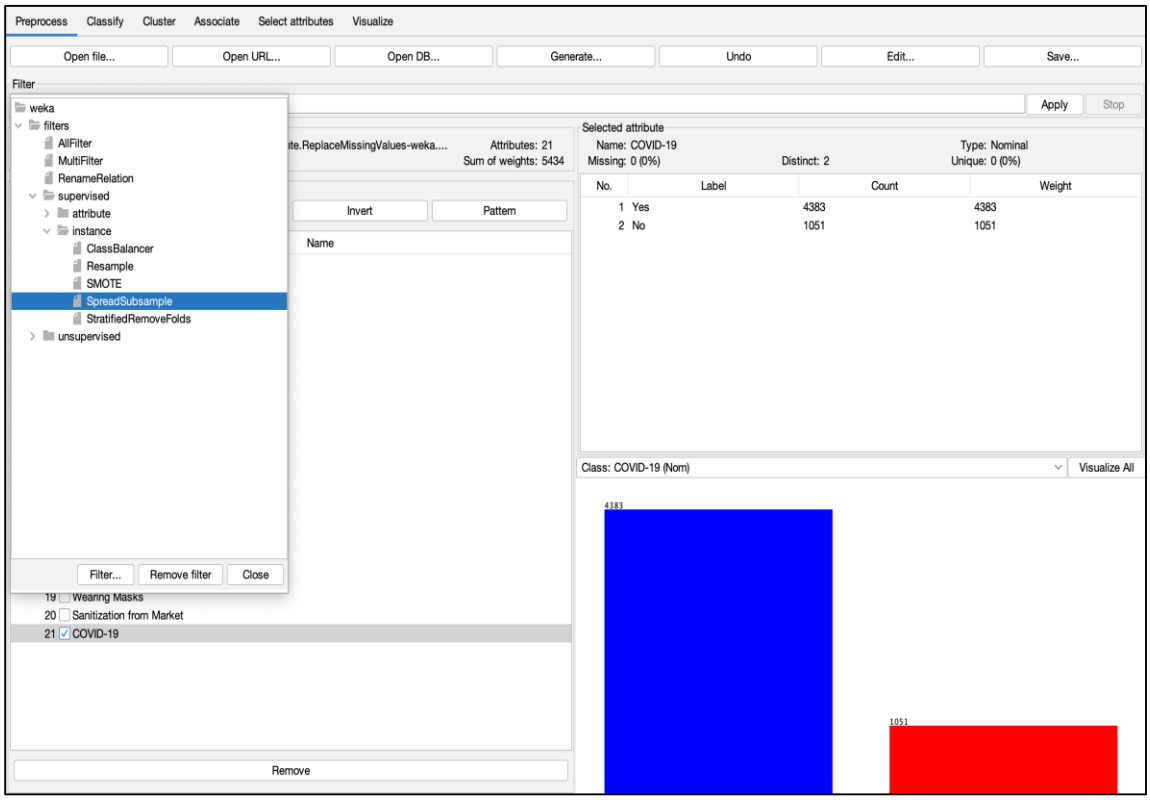

Figure 7 The filter dialog box shows SMOTE and SpreadSubsample filters that can be found under the instance section

Figure 9 shows the algorithm that outlines a process for analyzing the COVID-19 dataset using different feature selection methods and classification models. The first step involves loading the COVID-19 dataset into the algorithm. In step 2, a feature selection method is applied to select a subset of relevant features from the dataset. This step is crucial as it can significantly improve the accuracy of the model by identifying and selecting only the most important and relevant features from the dataset. A classifier (n) is also selected as a hyperparameter applied to the feature selection in this step. In step 3, for each feature selection choose, a classification model (N) is trained using only the selected features. The performance of the model is then evaluated to determine its accuracy. This process is repeated for each feature selection method being considered. In step 4, the accuracy results of classification models for each feature selection method are determined and compared to identify the best performing model.

Step 2 and step 3 are particularly important as they can increase the accuracy of the model by selecting only the most relevant features from the dataset and training a classification model using those features. By reducing the dimensionality of the dataset and focusing only on the most important features, these steps can improve the performance of the classification model and increase its accuracy in predicting COVID-19 cases. This approach is novel as it uses feature selection methods with classifiers act as a hyperparameter to improve the accuracy of COVID-19 prediction.

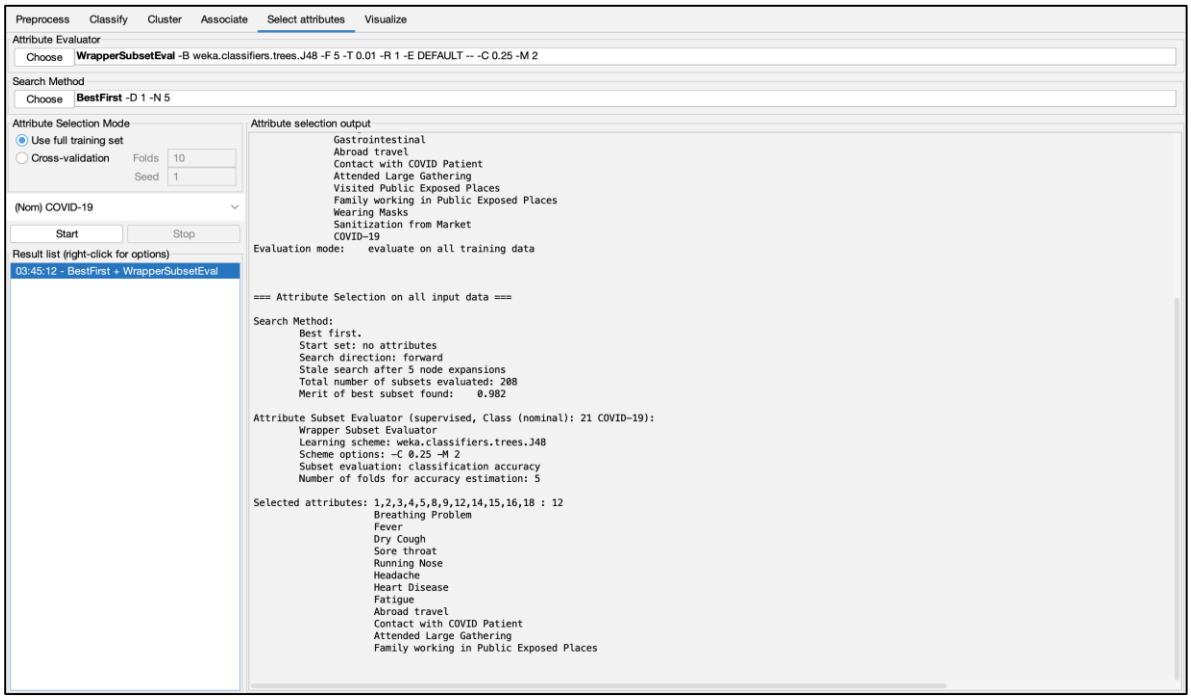

Figure 8 WEKA's Select attributes tab, where feature selection techniques for the dataset were chosen

Step 1: Load COVID-19 dataset
Step 2: Apply feature selections method
  i. select classifier (n) as hyperparameter
  ii. select a subset of relevant features from the dataset
Step 3: For each feature selection, do
  i. train a classification models (N) using only selected features
  ii. Evaluate the performance of N
  iii. End
Step 4: Determine the accuracy result of N for each feature selection
Step 5: End

Figure 9 The algorithm of the proposed system

Figure 10 presents the hyperparameter turning for the WrapperSubsetEval feature selection method. In this figure, the hyperparameter for the WrapperSubsetEval method are specified for various classifiers, including DT, J48, SVM, NB, KNN, and SMO. The default values are used for the number of folds (internal cross-validation), the evaluation measure used, the seed value for randomization, and the threshold value, indicating that the default settings were applied during the hyperparameter tuning process except for the variation of the classifiers used. *Figure 11*, on the other hand, focuses on the hyperparameter tuning for the CfsSubsetEval feature selection method. In this figure, only the default setting is specified for the hyperparameters of the method, indicating that no specific tuning was performed for this feature selection method.

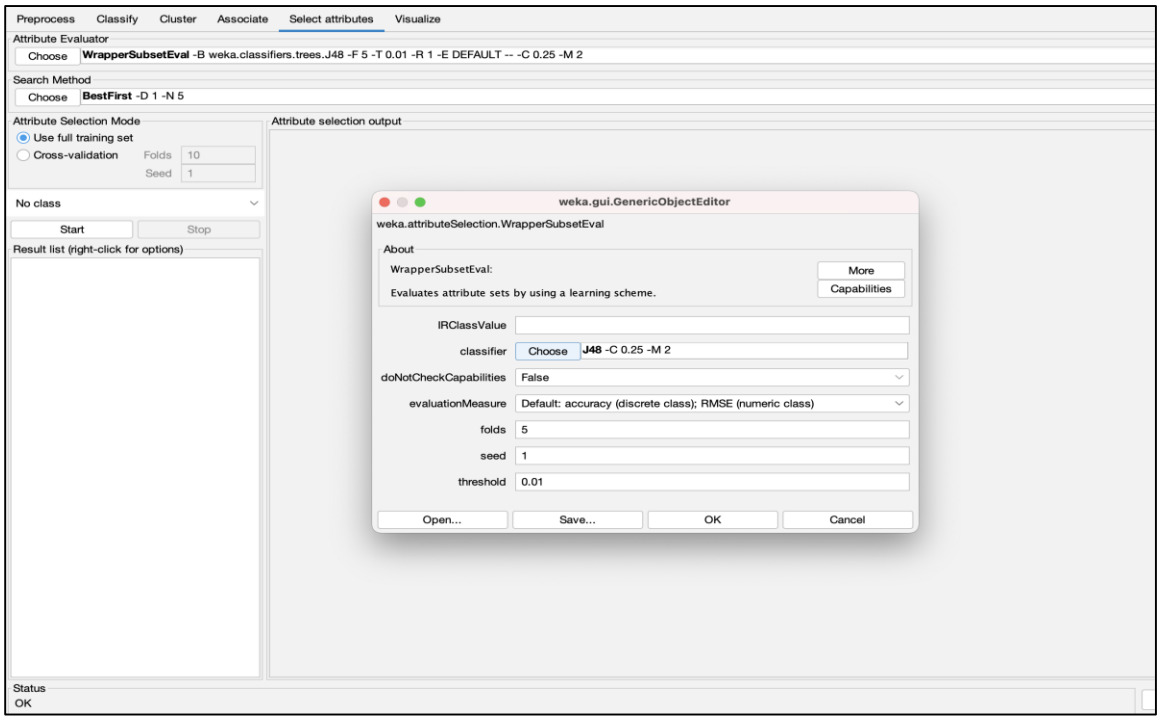

Figure 10 WEKA's Select attributes tab shows the hyperparameter tuning for WrapperSubsetEval feature selection

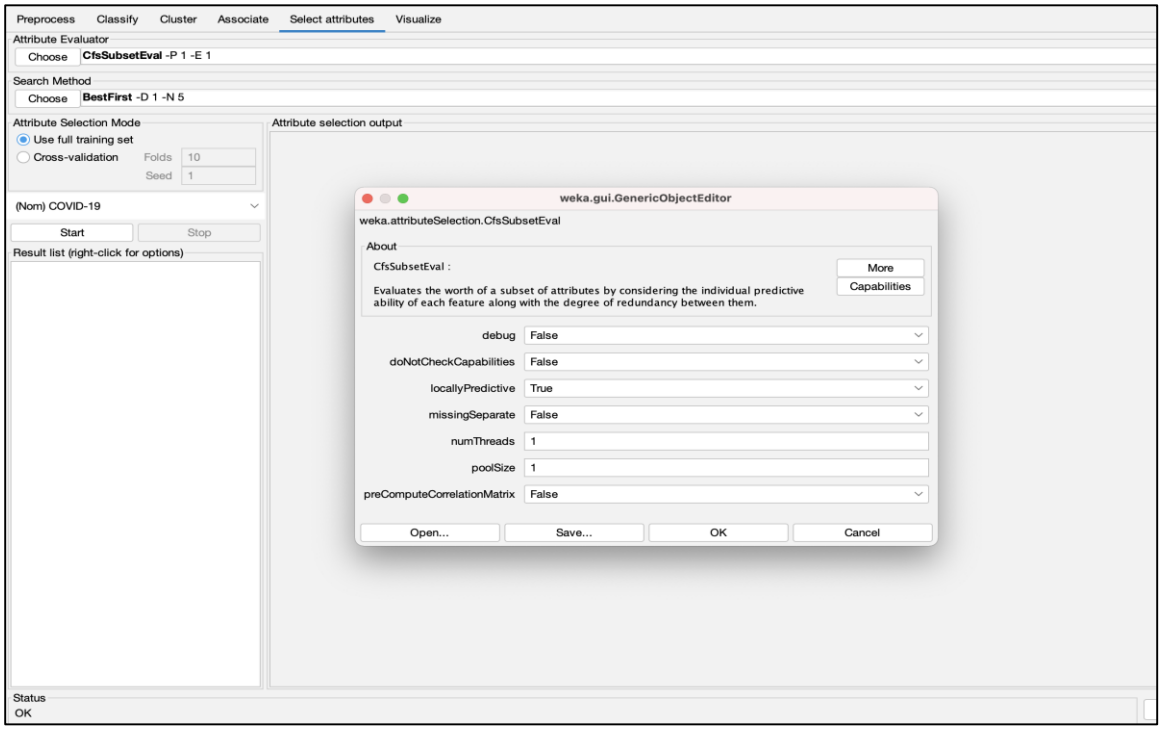

Figure 11 WEKA's Select attributes tab shows the hyperparameter tuning for CfsSubsetEval feature selection

3.4Modelling

After the feature selection methods were completed, a variety of supervised machine learning algorithms, including J48, SVM, NB, KNN and SMO were utilized through the WEKA Explorer Classify tab to construct multiple models. These models were built by

using a 10-fold cross-validation technique. Cross-validation can include each sample in the testing because no two test sets overlap as a result of the sampling. During the k-fold cross-validation, the original training set is divided into k separate and non-overlapping subsets of an equivalent size, where the term "fold" denotes the proportion of resulting subsets, while the value k indicates the number of folds used in the process [46]. The classifier's parameter used in this experiment was set to the default parameter. The classifier output part presented the generated model's performance for each training as depicted in *Figure 12*.

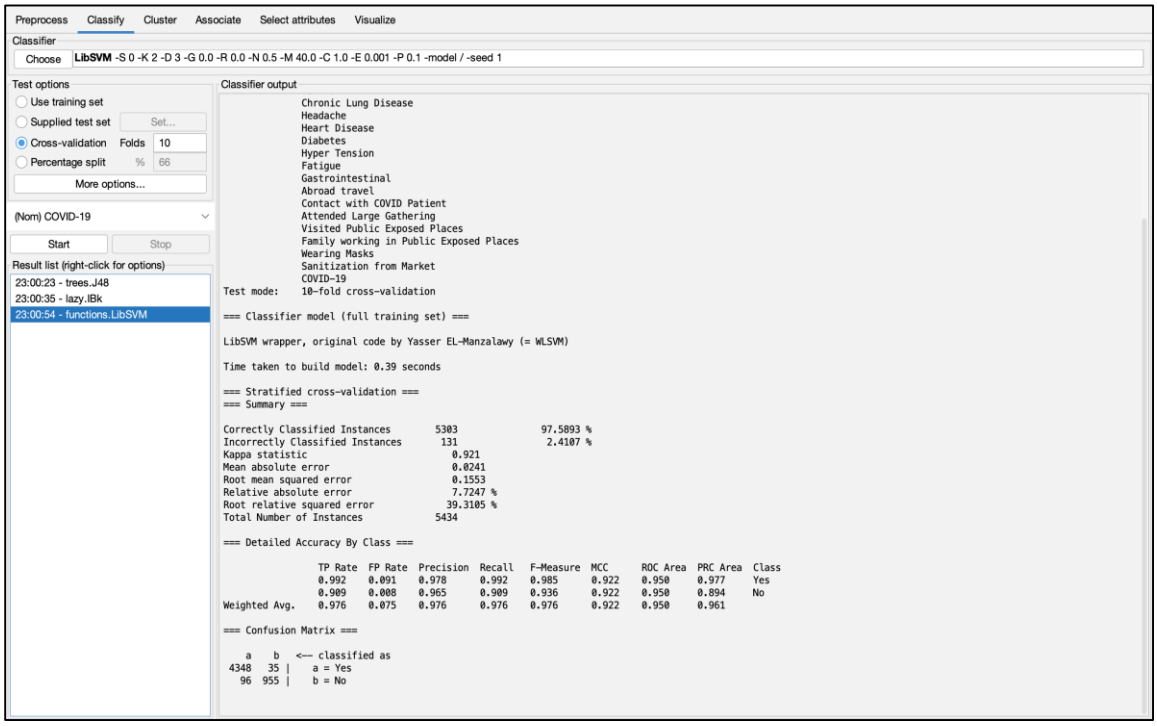

Figure 12 The results of the classification models can be viewed through WEKA's Classify function when the 10-fold cross-validation technique is implemented

In this experiment, the following are the requirements to create a COVID-19 presence predictor. These requirements will be used to determine the model which is going to serve as the best algorithm for machine learning.

- Highest accuracy;
- Lowest feature selection processing time.

3.5Evaluation metric

To assess the performance of the classification model, various evaluation metrics can be employed. One commonly used metric is the confusion matrix, which provides a detailed breakdown of the model's predictions. The confusion matrix, as shown in *Table 3*, presents four key elements: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).

TP refers to the instances correctly predicted and classified as positive. It signifies the model's ability to accurately identify positive cases. TN represents the instances correctly predicted and classified as negative, demonstrating the model's effectiveness in identifying negative cases. FP occurs when the model incorrectly predicts positive cases that are negative. On the other hand, FN represents instances that are truly positive but are incorrectly predicted as negative [47].

In addition to the confusion matrix, another widely used evaluation metric is classification accuracy. It measures the proportion of correct predictions made by the classification model out of the total number of predictions made. The accuracy, as defined in Equation 1, considers both TP and TN predictions and provides an overall measure of the model's predictive performance.

$$Accuracy, acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Table 3 Confusion matrix

	Actual Yes		Actual No	
Predicted Yes	True (TP)	Positive	False (FN)	Negative
Predicted No	False (FP)	Positive	True (TN)	Negative

4.Results

The proposed work is using the WrapperSubsetEval and CfsSubsetEval methods for the COVID-19 dataset. In this section, we will interpret and discuss the results presented in the previous section.

Table 4 shows the accuracy performance of five different classifiers using a 10-fold cross-validation approach named J48, SVM, NB, KNN and SMO. In the absence of data pre-processing and feature selection (unprocessed data set), J48 outperformed the other four classifiers in terms of classification accuracy of 98.18%. The KNN model, which is not significantly different from J48, achieved the second-best accuracy at 98.09%. The NB model attained the lowest accuracy at 96.54%.

After applying data preprocessing, KNN achieved the highest accuracy at 98.69%, followed by J48 at 98.45% as shown in *Table 5*. This indicates that there was an increment in accuracy value of 0.27% – 0.60% after the preprocessing method. The accuracy of SVM, NB and SMO dropped to 97.24%, 93.98% and 95.48% respectively. Upon comparing both sets of results, the accuracy scores for the classifiers have changed slightly. However, the overall trend of which classifiers perform the best is consistent. It is also noteworthy that while the accuracy scores have changed, the top performers are still J48 and KNN, suggesting that these classifiers are generally robust and perform well on COVID-19 dataset.

Table 4 Analysis of the accuracy of different classifiers without data pre-processing and feature selection (unprocessed data set)

Classifier	J48	SVM	NB	KNN	SMO
Accuracy (%)	98.18	97.59	96.54	98.09	96.74

Table 5 Analysis of the accuracy of different classifiers without feature selection/attribute selection (after data pre-processing)

Classifier	J48	SVM	NB	KNN	SMO
Accuracy (%)	98.45	97.24	93.98	98.69	95.48
Number of Features	21				

Table 6 portrays the result after the application of feature selection by using CfsSubsetEval and WrapperSubsetEval with different classifiers used when estimating the subset accuracy to a COVID-19 dataset. The feature selection methods used are WrapperSubsetEval+SVM, WrapperSubsetEval+J48, WrapperSubsetEval+NB, WrapperSubsetEval+KNN, WrapperSubsetEval+SMO, and CfsSubsetEval. The classifiers used are J48, SVM, NB, KNN, and SMO. In this article, the evaluation metric employed is the accuracy of the classifier. This metric measures the percentage of accurate predictions made by the classifier, without any bias towards incorrect predictions. In addition to that, the duration required to execute the feature selection method is also denoted as the processing time. It is also included the reduced number of features in the table. The application of feature selection to the COVID-19 dataset resulted in enhanced classification accuracy for some classifiers with the reduced feature set.

From this table, it appears that using WrapperSubsetEval+KNN feature selection methods with the respective classifiers results in the highest accuracy scores. Based on the WrapperSubsetEval+KNN feature selection, the best performing classifier based on accuracy is KNN, which had an accuracy of 98.81%. The WrapperSubsetEval+J48 feature selection method with J48 classifier is also a top performer with 98.74% accuracy and relatively low processing time of 8 seconds. The WrapperSubsetEval+SMO and WrapperSubsetEval+SVM feature selection methods had relatively high accuracy scores, but with higher processing times of 147 and 927 seconds respectively. The WrapperSubsetEval+NB shows that the accuracy achieved by the classifiers are the lowest due to the underfitting with only 7 features that were evaluated. CfsSubsetEval feature selection method had relatively lower accuracy scores and a low processing time of 1 second.

Figure 13 demonstrates that most of the algorithms exhibited satisfactory performance during the training process, as determined by data pre-processing and feature selection. The findings suggest that these approaches are effective for achieving desirable outcomes in the analysed data. Compared to the accuracy of the unprocessed data set, most of the models gave an improvement in the accuracy percentage. The x-axis shows the feature selection method and the y-axis shows the accuracy performance of classifiers. This figure is the summarization of the *Table 6* analysis.

Table 6 Analysis of different feature selection applied on a COVID-19 data set

Feature Selection	Reduced no. of features	Processing time (S)	Classifier accuracy (%)				
			J48	SVM	NB	KNN	SMO
WrapperSubsetEval+SVM	14	927	98.62	97.45	94.50	98.72	94.87
WrapperSubsetEval+J48	14	8	98.45	96.46	94.12	98.74	94.87
WrapperSubsetEval+NB	7	2	95.24	95.15	95.15	95.24	94.31
WrapperSubsetEval+KNN	14	512	98.36	97.15	94.24	**98.81**	95.58
WrapperSubsetEval+SMO	11	147	98.00	96.77	94.24	98.29	95.91
CfsSubsetEval	10	1	97.36	96.70	94.15	97.62	95.22

Based on the grouped bar graph, the WrapperSubsetEval+KNN and WrapperSubsetEval+J48 feature selection methods had the highest accuracy scores across all the classifiers. These feature selection methods can achieve accuracy scores of around 98.8% and 98.7% respectively. The accuracy scores for WrapperSubsetEval+SMO and WrapperSubsetEval+SVM are also relatively high, around 98.0% and 98.6% respectively, but still lower than the previous two methods. The WrapperSubsetEval+NB and CfsSubsetEval feature selection methods had the lowest accuracy scores, around 95.2% and 97.4% respectively. It is also notable that KNN classifiers had consistent high accuracy scores across all feature selection methods, while the other classifiers (J48, SVM, NB, and SMO) had relatively lower accuracy scores.

It can be noticed that from the visualization bar graph, WrapperSubsetEval+SVM, WrapperSubsetEval+J48 and WrapperSubsetEval+KNN might result in approximately the same accuracy for J48, SVM, NB, SMO and KNN classifiers but had a very significant difference in processing time.

Figure 14 shows that there were significant differences in execution time between different feature selection methods. When the reduced number of features varies, there is a discrepancy in the classifier accuracy. Among the three methods as in WrapperSubsetEval+SVM, WrapperSubsetEval+J48 and WrapperSubsetEval+KNN, the processing time of WrapperSubsetEval+SVM was higher than the two methods which took 927s. In addition, WrapperSubsetEval+KNN consumed more time to execute the subset than WrapperSubsetEval+J48 which is 512 seconds, while WrapperSubsetEval+NB had a processing time of only 2 seconds. CfsSubsetEval was the fastest method to execute the feature subset which only took 1 second however, their classifier accuracy was not as good as compared to the WrapperSubsetEval method. Different methods, on the contrary, have different strengths in classification data analysis. This indicates that some feature selection methods may be more computationally expensive than others.

In terms of feature reduction, different feature selection methods also resulted in different levels of reduction in the number of features. As can be seen in the *Figure 14*, WrapperSubsetEval+SVM reduced the number of features to 14, while WrapperSubsetEval+NB reduced the number of features to only 7. This suggests that different feature selection methods can result in different levels of dimensionality reduction.

Figure 15 demonstrate a summary of the confusion matrix for each classifier trained using a different feature selection method on the COVID-19 dataset. Different feature selection methods result in different levels of performance in terms of TP, FN, FP, and TN. For example, the WrapperSubsetEval+KNN method had the highest number of TP and the lowest number of FN among all the methods. This suggests that this method is effective at correctly identifying positive cases and minimizing false negatives. On the other hand, the CfsSubsetEval method had a relatively lower number of TP and a higher number of FN compared to other methods.

The confusion matrix in *Figure 15* corresponds to the WrapperSubsetEval+KNN feature selection, shows that this method resulted in 2052, 2033, 1990, 1990, and 2012 TP for KNN, J48, NB, SMO, and SVM respectively. This indicates that these models correctly identified positive cases, with the KNN classifier contributing the highest number of correctly classified instances. The true negatives were 2102 (KNN), 2102 (J48), 1972 (NB), 2028 (SMO), and 2072 (SVM), meaning that these models correctly identified these cases as negative. The “Incorrect_No” and the “Incorrect_Yes” refer to the false negative and false positive cases, respectively. The other confusion matrices in *Figure 15* can be interpreted in a similar manner. Each matrix provides information on the performance of a specific feature selection method in

terms of its ability to correctly identify positive and negative cases. By comparing the confusion matrices, the classification accuracy of different feature selection methods can be evaluated.

In general, feature selection methods that result in a greater reduction in the number of features require longer execution times, whereas methods that result in less reduction tend to have shorter execution times. For example, the WrapperSubsetEval+SVM method decrease the number of features to 14 and had a much longer execution time of 927 seconds, whereas the WrapperSubsetEval+NB method reduces the number of features to only 7 and had a much shorter execution time of 2 seconds. The findings indicate that the WrapperSubsetEval+SVM technique requires a longer processing time as it examines a larger dataset to identify the most significant subset of features. On the other hand, the WrapperSubsetEval+NB method processes a smaller dataset, resulting in a shorter execution time.

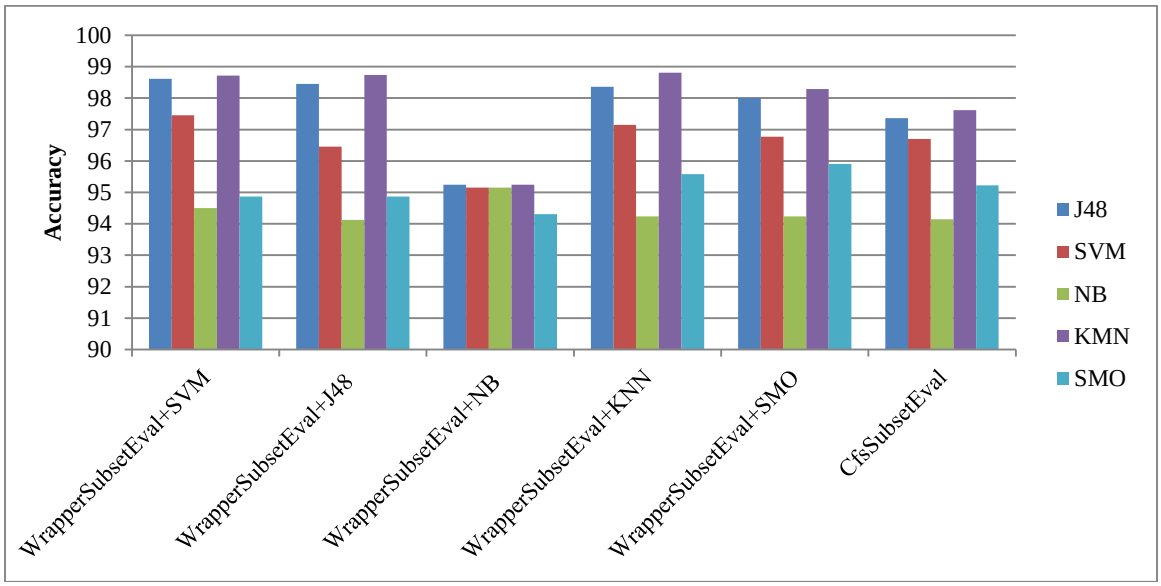

Figure 13 The visualization of bar chart shows the accuracy measurements of the developed model by utilizing different feature selection methods and algorithms

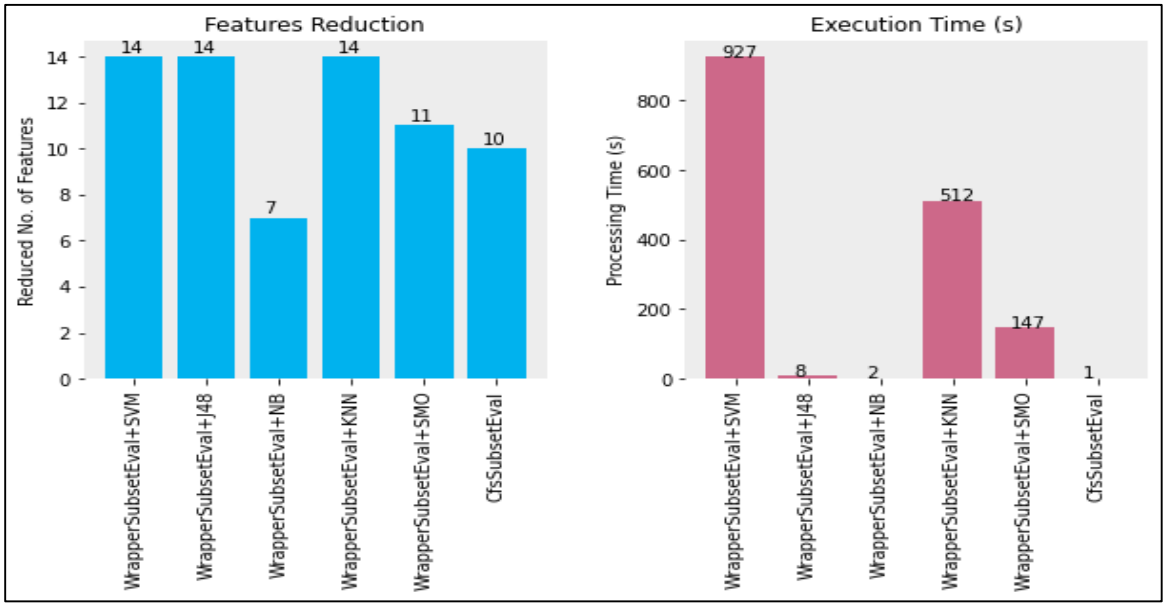

Figure 14 The bar charts of the execution time and features reduction

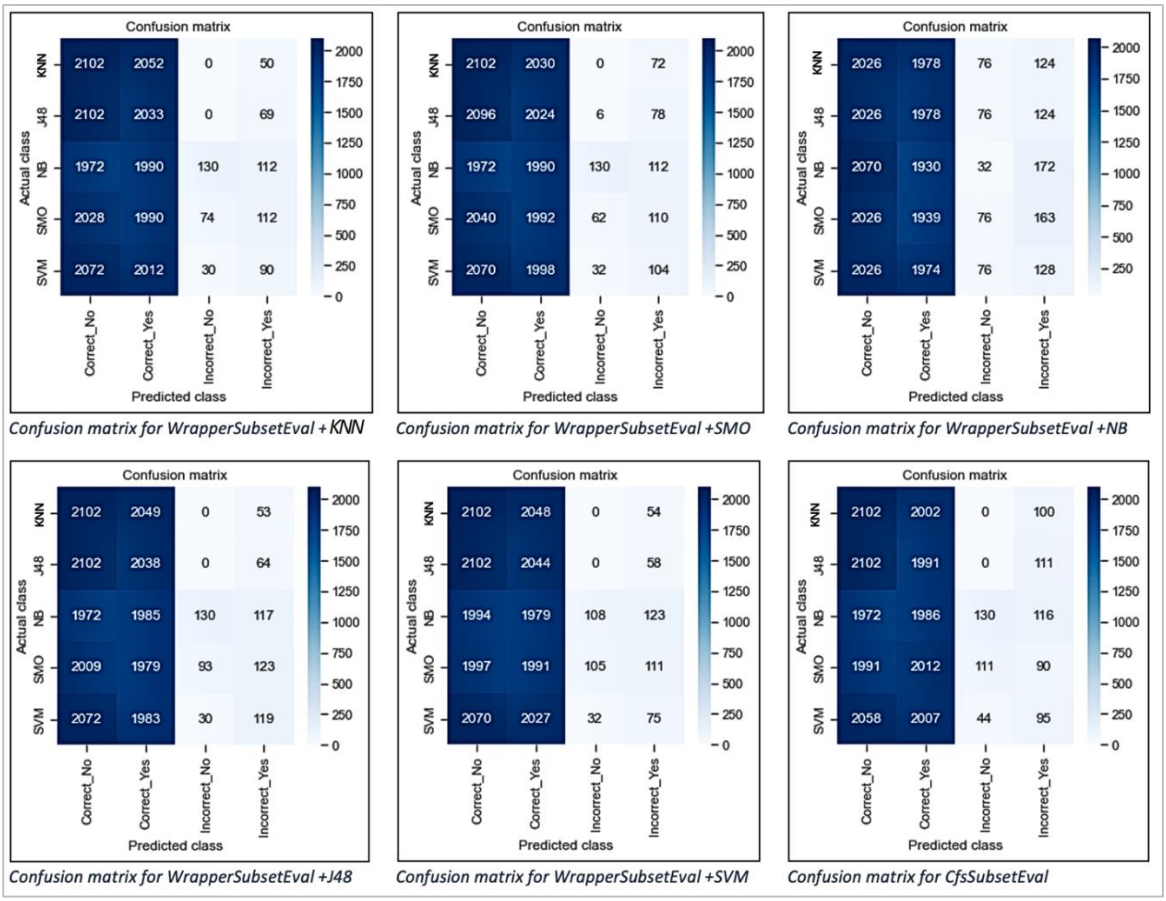

Figure 15 Confusion matrix of feature selection method

5.Discussion

Based on the results, KNN classifier using WrapperSubsetEval+KNN method was found to be the top-performing classifier, boasting a remarkable accuracy rate of 98.81%. Additionally, it had a reduced feature subset of 14 attributes than 21 attributes with the processing time of 512 seconds which consider lowest. The mechanism behind KNN algorithm itself is likely the reason why KNN produced the highest accuracy score in the result because KNN is a very simple and robust algorithm. It is a lazy learning algorithm, meaning that it does not perform any training on the data. Instead, the algorithm stores the entire training set and uses it directly to make predictions at test time which can be very effective at capturing complex decision boundaries in the data. J48 classifier came in second with accuracy ranging from 95.24% to 98.62%, followed by SVM classifier with accuracy ranging from 95.15% to 97.45%.

It is also important to note that in machine learning pipeline, selecting features is a crucial stage. It can significantly enhance the classifier's performance and reduce the dimensionality of the dataset, making it computationally more efficient. Wrapper methods like WrapperSubsetEval are generally considered more effective than filter methods like CfsSubsetEval, as they consider the classifier performance when selecting features, but they are also computationally more expensive. From the results, WrapperSubsetEval performed well compared to CfsSubsetEval in terms of accuracy even it was more time-consuming than CfsSubsetEval. The key findings of the study include that using feature selection methods can help improve the accuracy of classifiers when applied to a COVID-19 dataset. The study found that different feature selection methods and classifiers performed differently when applied to the dataset, and when using an appropriate feature selections and classifiers can help to improve the performance of predictive models for COVID-19

The primary results of the research indicate that WrapperSubsetEval+KNN is the best performing Wrapper Subset Evaluation method, while KNN is the

best performing classifier. These findings can be useful in developing accurate and reliable predictive models for COVID-19, and contribute to the fight against the pandemic. The implications of these findings suggest that the Wrapper Subset Evaluation and KNN classifier could be a useful tool in the early prediction of COVID-19 data, which would have significant implications for the field of study in the healthcare sector.

5.1 Limitations

The evaluation metrics used in this study may not offer a clear image of the classifiers' performance. Other metrics including area under the receiver operating characteristic (ROC) curve, specificity, recall, or precision could also be considered. The study's utilization of specific machine learning algorithms is also a constraint that should be considered. Other algorithms not tested in the study may perform better or worse than those used. Another limitation that can be highlighted is the dataset used in this study is limited to COVID-19 cases only and may not be representative of other types of respiratory diseases or infections. This could limit the generalizability of the findings to other populations or diseases.

A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

This research experiment was conducted to compare the different feature selection algorithms and analyze the attribute selection and their classification accuracies that are widely used in data mining for the COVID-19 dataset. The research question or hypothesis of this research was to evaluate the performance of different classifiers in predicting the COVID-19 data using feature selection methods. The predictor model for the COVID-19 dataset was built by employing five supervised machine learning techniques, such as J48, SVM, NB, KNN, and SMO. The performance of the model was assessed through a comparison analysis using the machine learning tool WEKA in a 10-fold cross-validation process. Different types of feature selection approaches like CfsSubsetEval and WrapperSubsetEval and algorithms used to evaluate the feature subsets such as J48, NB, SMO, KNN and SVM were discussed. When doing the feature selection method, the dataset becomes reduced thus plead to the occurrence of higher accuracy results. Using feature selection algorithms in this work, it can determine which attributes are beneficial and which ones should not be used for the prediction model.

The statistical outcomes were scrutinized based on the accuracy of classification. Based on the obtained results, KNN classifiers performed consistently well in terms of accuracy, both with and without pre-processing and feature selection. However, applying feature selection methods did result in some improvement in accuracy scores for some classifiers. The WrapperSubsetEval+KNN feature selection method with the KNN classifier was found to be the best performing method, achieving an accuracy of 98.81%. The significance of feature selection in the machine learning was highlighted, as it had the potential to enhance classifier performance and reduce dataset dimensionality to a great extent. Future research could extend these findings by exploring different feature selection methods to determine the best methods for accurate COVID-19 prediction, such as embedded methods and hybrid methods, to compare their performance with WrapperSubsetEval and CfsSubsetEval. Additionally, further investigation can be conducted to determine the reasons for the performance differences observed among the classifiers and feature selection methods. For example, precision, F1 score and confusion metrics which can further provide insights into the performance of the classifiers and help identify specific areas where improvements can be made.

In summary, the study provides valuable insights into the performance of different classifiers in predicting COVID-19 data, and highlights the importance of feature selection methods in improving the accuracy of predictive models. The results of this study can serve as a starting point for further research in this area and contribute to the development of more effective predictive models for COVID-19. The results also provide insights into the strengths and weaknesses of different classifiers, which can be useful in selecting appropriate classifiers for similar prediction tasks in the future.

Acknowledgment

This research was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2020/ICT06/UNISZA/02/1).

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Fauzan Iliya Khalid: Conceptualization, methodology, framework draft, investigation, analysis, interpretation of results, writing original draft. **Mokhairi Makhtar:** Validation of the models, interpretation of the results, framework of methodology, project administration,

reviewing and editing. **Rosaida Rosly:** Validation of the analytics model, project administration, reviewing and editing. **Aceng Sambas:** Reviewing, analysis, and editing.

References

[1] Podder P, Mondal MR. Machine learning to predict COVID-19 and ICU requirement. In 11th international conference on electrical and computer engineering 2020 (pp. 483-6). IEEE.

[2] Silahudin D, Holidin A. Model expert system for diagnosis of covid-19 using naïve Bayes classifier. In IOP conference series: materials science and engineering 2020 (pp. 1-7). IOP Publishing.

[3] Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN Computer Science. 2021; 2:1-3.

[4] Shanmugam SK. A study on the performance of classification models for COVID-19 datasets. Turkish Journal of Computer and Mathematics Education. 2021; 12(10):1123-7.

[5] Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: a review. Egyptian Informatics Journal. 2018; 19(3):179-89.

[6] Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-turjman F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisciplinary Sciences: Computational Life Sciences. 2021; 13:103-17.

[7] Rahman MM, Usman OL, Muniyandi RC, Sahran S, Mohamed S, Razak RA. A review of machine learning methods of feature selection and classification for autism spectrum disorder. Brain Sciences. 2020; 10(12):949.

[8] Al JKB, Kadhim R. Data reduction techniques: a comparative study for attribute selection methods. International Journal of Advanced Computer Science and Technology. 2018; 8(1):1-13.

[9] Venkatesh B, Anuradha J. A review of feature selection and its methods. Cybernetics and Information Technologies. 2019; 19(1):3-26.

[10] Richhariya B, Tanveer M, Rashid AH. Alzheimer's disease neuroimaging initiative diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE). Biomedical Signal Processing and Control. 2020; 59:101903.

[11] Senan EM, Al-adhaileh MH, Alsaade FW, Aldhyani TH, Alqarni AA, Alsharif N, et al. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering. 2021; 2021:1-10.

[12] Gnanambal S, Thangaraj M, Meenatchi VT, Gayathri V. Classification algorithms with attribute selection: an evaluation study using WEKA. International Journal of Advanced Networking and Applications. 2018; 9(6):3640-4.

[13] Elgamal ZM, Yasin NB, Tubishat M, Alswaitti M, Mirjalili S. An improved Harris hawks optimization algorithm with simulated annealing for feature selection in the medical field. IEEE Access. 2020; 8:186638-52.

[14] Gárate-escamila AK, El HAH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked. 2020; 19:1-13.

[15] Zaini NA, Awang MK. Hybrid feature selection algorithm and ensemble stacking for heart disease prediction. International Journal of Advanced Computer Science and Applications. 2023; 14(2):158-65.

[16] Wah YB, Ibrahim N, Hamid HA, Abdul-rahman S, Fong S. Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika Journal of Science & Technology. 2018; 26(1):329-40.

[17] Alaika L, Alamsyah A. Optimization of accuracy to autism spectrum disorder identification for children using support vector machine and correlation-based feature selection. Journal of Advances in Information Systems and Technology. 2022; 4(1):1-2.

[18] Reddy KV, Elamvazuthi I, Abd AA, Paramasivam S, Chua HN, Pranavanand S. Prediction of heart disease risk using machine learning with correlation-based feature selection and optimization techniques. In 7th international conference on signal processing and communication 2021 (pp. 228-33). IEEE.

[19] Kar M, Dewangan L. Classification of epileptic EEG signals based on J48 classifier and correlation based feature selection. International Journal for Research in Applied Science & Engineering Technology. 2018; 6:2557–60.

[20] Khaniabadi PM, Bouchareb Y, Al-dhuhli H, Shiri I, Al-kindi F, Khaniabadi BM, et al. Two-step machine learning to diagnose and predict involvement of lungs in COVID-19 and pneumonia using CT radiomics. Computers in Biology and Medicine. 2022; 150:106165.

[21] Effrosynidis D, Arampatzis A. An evaluation of feature selection methods for environmental data. Ecological Informatics. 2021; 61:101224.

[22] Zhang R, Nie F, Li X, Wei X. Feature selection with multi-view data: a survey. Information Fusion. 2019; 50:158-67.

[23] Omuya EO, Okeyo GO, Kimwele MW. Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications. 2021; 174:114765.

[24] Almugren N, Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access. 2019; 7:78533-48.

[25] Shaban WM, Rabie AH, Saleh AI, Abo-elsoud MA. A new COVID-19 patients detection strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowledge-Based Systems. 2020; 205:106270.

[26] Torse DA, Khanai R, Pai K, Iyer S, Mavinkattimath S, Kallimani R, et al. Optimal feature selection for

COVID-19 detection with CT images enabled by metaheuristic optimization and artificial intelligence. Multimedia Tools and Applications. 2023:1-31.
[27] Danacı Ç, Tuncer SA. Incorporating feature selection methods into machine learning-based covid-19 diagnosis. Applied Computer Systems. 2022; 27(1):13-8.
[28] Hayet-otero M, García-garcía F, Lee DJ, Martínez-minaya J, España VPP, Urrutia LI, et al. Extracting relevant predictive variables for COVID-19 severity prognosis: an exhaustive comparison of feature selection techniques. Plos One. 2023; 18(4): e0284150.
[29] Ali RH, Abdulsalam WH. The prediction of covid 19 disease using feature selection techniques. In journal of physics: conference series 2021 (1-12). IOP Publishing.
[30] Yusuf R. Comparing different supervised machine learning accuracy on analyzing COVID-19 data using ANOVA test. In 6th international conference on interactive digital media 2020 (pp. 1-6). IEEE.
[31] Varzaneh ZA, Orooji A, Erfannia L, Shanbehzadeh M. A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method. Informatics in Medicine Unlocked. 2022; 28:100825.
[32] Mohammad MA, Aljabri M, Aboulnour M, Mirza S, Alshobaiki A. Classifying the mortality of people with underlying health conditions affected by COVID-19 using machine learning techniques. Applied Computational Intelligence and Soft Computing. 2022; 2022:1-12.
[33] Sardar R, Sharma A, Gupta D. Machine learning assisted prediction of prognostic biomarkers associated with COVID-19, using clinical and proteomics data. Frontiers in Genetics. 2021; 12:636441.
[34] Palattao CA, Solano GA, Tee CA, Tee ML. Determining factors contributing to the psychological impact of the COVID-19 pandemic using machine learning. In international conference on artificial intelligence in information and communication 2021 (pp. 219-24). IEEE.
[35] Mahdi AY, Yuhaniz SS. Optimal feature selection using novel flamingo search algorithm for classification of COVID-19 patients from clinical text. Mathematical Biosciences and Engineering. 2023; 20(3):5268-97.
[36] Ranganathan G. A study to find facts behind preprocessing on deep learning algorithms. Journal of Innovative Image Processing. 2021; 3(1):66-74.
[37] Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences. 2017; 12(16):4102-7.
[38] Jain N, Jhunthra S, Garg H, Gupta V, Mohan S, Ahmadian A, et al. Prediction modelling of COVID using machine learning methods from B-cell dataset. Results in Physics. 2021; 21:103813.
[39] Usman MM, Owolabi O, Ajibola AA. Feature selection: it importance in performance prediction. IJESC. 2020:25625-32.
[40] Shaikh TA, Ali R. Applying machine learning algorithms for early diagnosis and prediction of breast cancer risk. In proceedings of 2nd international conference on communication, computing and networking 2019 (pp. 589-98). Springer Singapore.
[41] Cornforth D, Jelinek H, Teich M, Lowen S. Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures. In international conference on computational intelligence for modelling, control and automation 2004 (pp. 446-55). University of Canberra.
[42] Gonçalves VP, Ribeiro EA, Imai NN. Mapping areas invaded by pinus sp. from geographic object-based image analysis (GEOBIA) applied on RPAS (Drone) color images. Remote Sensing. 2022;14(12):2805.
[43] Mishra S, Mallick PK, Tripathy HK, Bhoi AK, González-briones A. Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. Applied Sciences. 2020; 10(22):8137.
[44] Nedeva V, Pehlivanova T. Students' performance analyses using machine learning algorithms in WEKA. In IOP conference series: materials science and engineering 2021 (pp. 1-13). IOP Publishing.
[45] Biswas S, Bordoloi M, Purkayastha B. Review on feature selection and classification using neuro-fuzzy approaches. International Journal of Applied Evolutionary Computation. 2016; 7(4):28-44.
[46] Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Computational Statistics. 2021; 36(3):2009-31.
[47] Aljohani A. Machine learning techniques for COVID-19 detection: a comparative analysis. International Journal of Computer and Information Engineering. 2022; 16(12):592-7.

Fauzan Iliya Khalid completed her Bachelor's degree in Computer Science with a specialization in Software Development from Universiti Sultan Zainal Abidin (UniSZA) in 2019. She is currently pursuing a Master of Science in Computer Science at the same university. Her areas of interest revolve around Classification Models, Machine Learning, and Deep Learning Algorithms.
Email: fauzan.iliya@gmail.com

Mokhairi Makhtar is a Professor of Computing at the Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin. He obtained his Ph.D. in Computer Science from the University of Bradford in 2012. He has authored more than 100 papers in peer-reviewed journals and international conferences. His research interests encompass Machine Learning, Data Mining, and Big Data Analytics for applications in Toxicology, Education, Health, and Business.
Email: mokhairi@unisza.edu.my

Rosaida Rosly is a lecturer at the Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terrengganu (UMT). She graduated her PhD in Computer Science and MSc in Computer Science from Universiti Sultan Zainal Abidin (UniSZA). She has 6 years of research experience and published more than 10 academic papers in journals including Scopus and Conferences. Her research interests include Data Mining, Machine Learning, Deep Learning, and Classification.
Email: rosaida@umt.edu.my

Aceng Sambas is currently a Lecturer at the Universiti Sultan Zainal Abidin, Malaysia since 2023. He received his PhD in Mathematics from the Universiti Sultan Zainal Abidin (UniSZA), Malaysia in 2021. His current research focuses on Dynamical Systems, Chaotic Signals, Electrical Engineering, Computational Science, Signal Processing, Robotics, Embedded Systems, and Artificial Intelligence.
Email: acengsambas@unisza.edu.my

Appendix I

S. No.	Abbreviation	Description
1	CFS	Correlation-Based Feature Selection
2	CfsSubsetEval	Correlation-Based Feature Subset Selection Evaluator
3	COVID-19	Coronavirus Disease 2019
4	CT	Computed Tomography
5	DT	Decision Tree
6	FP	False Positive
7	FN	False Negative
8	fsvFS	Feature Selection Via Concave Minimization
9	GA	Genetic Algorithms
10	IBFSA	Improved Binary Flamingo Search Algorithm
11	J48	C4.5 Decision Tree in WEKA
12	KNN	K-Nearest Neighbor
13	LR	Logistic Regression
14	MLP	Multi-Layer Perceptron
15	NB	Naïve Bayes
16	NN	Neural Network
17	RBM	Restricted Boltzmann Machine
18	RF	Random Forest
19	RFE	Recursive Feature Elimination
20	ROC	Receiver Operating Characteristic
21	SARS-CoV-2	Coronavirus Disease COVID-19
22	SA	Simulated Annealing
23	SMO	Sequential Minimal Optimization
24	SVM	Support Vector Machine
25	TP	True Positive
26	TN	True Negative
27	WEKA	Waikato Environment for Knowledge Analysis
28	WrapperSubsetEval	Wrapper Subset Evaluator
29	WrapperSubsetEval+ KNN	Wrapper Subset Evaluation With K-Nearest Neighbors (as a feature selection technique)
30	WrapperSubsetEval+J 48	Wrapper Subset Evaluation with Decision Tree (as a feature selection technique)
31	WrapperSubsetEval+ NB	Wrapper subset evaluation with naïve bayes (as a feature selection technique)
32	WrapperSubsetEval+ SMO	Wrapper Subset Evaluation with Sequential Minimal Optimization (as a feature selection technique)
33	WrapperSubsetEval+ SVM	Wrapper Subset Evaluation with Support Vector Machine (as a feature selection technique)