



Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy

Nehal A. Mansour¹ · Ahmed I. Saleh² · Mahmoud Badawy² · Hesham A. Ali²

Received: 22 September 2020 / Accepted: 23 December 2020

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

The outbreak of Coronavirus (COVID-19) has spread between people around the world at a rapid rate so that the number of infected people and deaths is increasing quickly every day. Accordingly, it is a vital process to detect positive cases at an early stage for treatment and controlling the disease from spreading. Several medical tests had been applied for COVID-19 detection in certain injuries, but with limited efficiency. In this study, a new COVID-19 diagnosis strategy called Feature Correlated Naïve Bayes (FCNB) has been introduced. The FCNB consists of four phases, which are; Feature Selection Phase (FSP), Feature Clustering Phase (FCP), Master Feature Weighting Phase (MFWP), and Feature Correlated Naïve Bayes Phase (FCNBP). The FSP selects only the most effective features among the extracted features from laboratory tests for both COVID-19 patients and non-COVID-19 people by using the Genetic Algorithm as a wrapper method. The FCP constructs many clusters of features based on the selected features from FSP by using a novel clustering technique. These clusters of features are called Master Features (MFs) in which each MF contains a set of dependent features. The MFWP assigns a weight value to each MF by using a new weight calculation method. The FCNBP is used to classify patients based on the weighted Naïve Bayes algorithm with many modifications as the correlation between features. The proposed FCNB strategy has been compared to recent competitive techniques. Experimental results have proven the effectiveness of the FCNB strategy in which it outperforms recent competitive techniques because it achieves the maximum (99%) detection accuracy.

Keywords COVID-19 · Classification · FCNB · Feature selection · Correlation

1 Introduction

Coronavirus is highly threatening for both animal and human life. Many types of coronavirus can transfer from animals to the human population (Shaban et al. 2020; Barstugan et al. 2020). Humans have not previously identified COVID-19 because it is a new species that appeared in 2019. COVID-19 is a global epidemic problem that can spread rapidly among people (Shaban et al. 2020; Li et al. 2020a). On the 7th of January 2020, COVID-19 has been identified by the World Health Organization (WHO) and the Chinese government as a global pandemic (Kang et al. 2020). COVID-19 has typical symptoms that involve shortness of breath, fever, headache,

cough, fatigue, sore throat, and muscle pain (Huang et al. 2020). Physical contact is the main reason of the spread of COVID-19 disease among people. The infections are transferred from the infected COVID-19 person to the healthy person through hand contact, mucous contact, or breathe contact. Because of the rapid spread of COVID-19 around the world, it causes a destructive impact on issues like public health, the global economy, and daily activities. Moreover, COVID-19 infection takes less than 4 weeks to quash the medical system once it begins to spread (Shaban et al. 2020). To this end, early detection of COVID-19 especially with the lacking of specific cures or vaccines, is an essential process for treating and controlling the disease from spreading.

A real-time Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is the most preferable test that is currently used for detecting COVID-19 patients (Zu et al. 2020). Although RT-PCR tests are sensitive, fairly quick, and reliable, these tests suffer from the risk of eliciting false-negative and false-positive results. Consequently, the spread of COVID-19 infection has been increased because

✉ Nehal A. Mansour
nehal.anees.mansour@gmail.com

¹ Nile Higher Institute for Engineering and Technology,
Mansoura, Egypt

² Computers and Control Department, Faculty of Engineering,
Mansoura University, Mansoura, Egypt

RT-PCR tests cannot immediately distinguish the infected people (Zu et al. 2020). Chest radiological imaging such as Computed Tomography (CT) images and X-rays play an important role in the early detection and treatment of COVID-19 patients. Despite the advantages of CT images for detecting COVID-19 patients, misclassification may occur between the imaging features of COVID-19 and other types of diseases (Shaban et al. 2020; Li et al. 2020a, b). With increasing demand toward providing accurate tests, the dependency on CT images or RT-PCR tests as accurate tools for the detecting of COVID-19 patients is decreased dramatically. To this end, fast and accurate detection of COVID-19 patients is very important to prevent the sources of infection. Recently, machine learning is an adjunct tool for clinicians. Machine learning can automatically support medical diagnosis as a helping tool for identifying and detecting the novel coronavirus.

Machine learning is an application of Artificial Intelligence (AI) that is used for the concept of software that automatically learns how to execute a task or solve a problem (Rabie et al. 2015; Rabie et al. 2019a). Machine learning techniques become more and more accurate over time, and they work on the same principle. Firstly, they receive some input training data. Then, build the mathematical model depending on this training input data. Finally, the mathematical model is used to solve the problem at hand. Many methods have been provided for COVID-19 detection based on machine learning techniques (Zhong et al. 2020; Rustam et al. 2020; Alazab et al. 2020). Despite the efficiency of these methods, they suffer from many limitations such as low diagnosis accuracy, high complexity, and long prediction time.

Naïve Bayes (NB) is a simple, popular classifier, and powerful machine learning technique. It has been verified as the highly professional probabilistic classifier that has solid mathematical fundamentals (Kumar et al. 2019; Rabie et al. 2015, 2019a). NB has worked very well in several complex real-world applications such as; medical diagnosis, real-time prediction, spam filtering, and weather forecasting despite its oversimplified assumptions and its Naïve design (Dada et al. 2019; Ali and Ali 2020; Hewage et al. 2020; Lei et al. 2020). Thus, NB can be considered as one of the best classifiers that can be applied for COVID-19 detection. This is due to many reasons, which can be summarized as follows; (1) NB can provide fast predictions rather than other classification algorithms because the training time has an order $O(N)$ with the dataset, (2) it can be easily trained with small amount of input training dataset and it can be used also for large datasets as well, (3) the simplicity and easy implementation with the ability of real-time training for new items, (4) the implementation of this classifier has no required adjusting parameters or domain knowledge, (5) It handles both continuous and discrete data, (6) NB is less sensitive to missing data,

(7) NB has high capability to handle the noise in the dataset, (8) NB is an Incremental learning approach because its functions work from an approximation of low-order probabilities which are extracted from the training data. Hence, these can be quickly updated as new training data are obtained, (9) If the Naïve Bayes conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression, (10) NB can be used for both binary and multiclass classification problems and (11) NB is sufficient for real-time applications such as diseases diagnoses because it relies on a set of pre-computed probabilities that make the classification done in a very short time (Khotimah et al. 2020; Kaur and Oberoi 2020). Although NB has proven efficiency with real-time applications, its performance is sometimes thumping in many cases because of the unrealistic assumption that all features have the same degree of importance and are independent of the given class value. Hence, this unrealistic assumption should be mitigated to overcome such hurdles. Recently, there have been extensive researches to provide solutions for this issue such as feature selection and weighting. However, the desired performance of NB has not been introduced yet. More efforts should be performed to enhance the performance of NB to match real-world conditions.

The contributions of the proposed work are listed as follows.

- A novel Feature Correlated Naïve Bayes (FCNB) classification strategy for accurate detection of Covid-19 patients has been proposed.
- The FCNB consists of two stages, namely; (1) Pre-Processing Stage (P^2S) and (2) Classification Stage (CS).
- The P^2S contains the first three phases of the FCNB strategy called Feature Selection Phase (*FSP*), Feature Clustering Phase (*FCP*), and Master Feature Weighting Phase (*MFWP*). Moreover, the CS contains the Feature Correlated Naïve Bayes Phase (*FCNBP*) that represents the last phase of the FCNB strategy.
- In the P^2S , the collected historical data on both COVID-19 patients and non- COVID-19 people are represented in suitable form after performing many essential processes to enable the diagnostic model in the next stage to accurately diagnose COVID-19 patients.
- During P^2S , the most significant features will be selected by using the Genetic Algorithm (*GA*) in *FSP*, and then these selected features will be put into groups or clusters in the *FCP* by using a new clustering technique in which each group is called Master Feature (*MF*) that includes a set of dependent or related features. Then, the *MFWP* will assign a weight value to each *MF* by using a new weight calculation method based on the number of features in *MF*, the correlation between features, and the summation of weights for each feature in *MF*.

- During the second stage (CS), the FCNBP tries to provide a fast and accurate diagnosis for COVID-19 patients based on the data received from the P²S by using a new classification model.
- The main objective of the proposed classification model is to overcome the problems of traditional weighted NB for improving its performance by (1) taking into consideration the correlation between features, and (2) reduces the classification time because it considers the weights of the used MFs rather than the weights of many individual features.

The paper is organized as follows: In Sect. 2, the main problem of this study is formulated. In Sect. 3, the diagnosing methodologies of COVID-19 are presented. In Sect. 4, related work is reviewed. In Sect. 5, the weighted Naïve Bayes is explained. In Sect. 6, the proposed Feature Correlated Naïve Bayes (FCNB) classification strategy is elaborated. An illustrative example is introduced in Sect. 7. The experiments are presented and the results are analyzed in Sect. 8. In Sect. 9, the paper is concluded and the future work is presented.

2 Problem definition

Due to the unavailability of a specific vaccine against COVID-19 infections with no drug has proven a high clinical efficacy, the early detection of COVID-19 disease is essential for disease cure and control. Undoubtedly, the management of COVID-19 will place considerable pressure on health-care systems (Wang et al. 2020; Li et al. 2020a, c). Moreover, the low availability of appropriate personal protective equipment for front-line health-care staff causes these key staff to be disproportionately affected by COVID-19. Nowadays, fast detection and isolation of the infected people is an effective method for the healthcare system protection from becoming overwhelmed because it will flat the epidemic curve as depicted in Fig. 1. Otherwise, with no protective measures, the capacity of the health-care systems will be broken. Disruption or complete breakdown of health-care systems would result in high mortality since the care of all illnesses will be degraded. Due to the unavailability of the diagnosis system everywhere, the detection of COVID-19 is currently a tedious task, which will cause panic. Because of the limited availability of COVID-19 testing kits especially in developed countries, there is a critical need to rely on other diagnosis strategies (Li et al. 2020a, c).

Rapid and accurate detection of COVID-19 is an increasingly vital issue since the infected people may not be recognized and get suitable treatment on time. The infected people will spread the virus to healthy people due to the communicable nature of COVID-19. Although several COVID-19

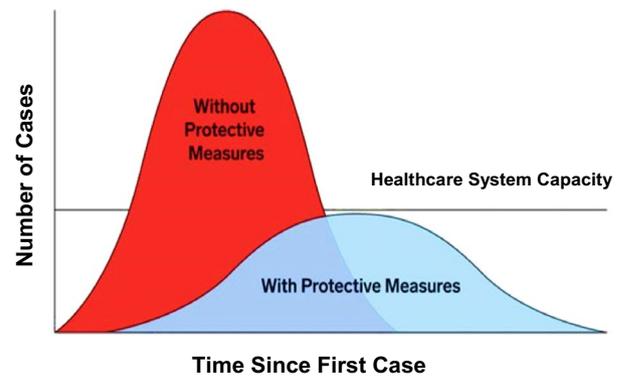


Fig. 1 COVID-19 epidemic curve with and without protective measures

diagnosis strategies based on data mining and artificial intelligence have been recently introduced, the desired diagnose accuracy to flatten the COVID-19 epidemic curve has not been reached yet (Li et al. 2020a, b; Jamshidi et al. 2020). The aim objective of this paper is to introduce an accurate, fast, and reliable COVID-19 diagnosis strategy, called FCNB, which inherits the advantages of NB with several modifications.

3 COVID-19 diagnose methodologies

Generally, the diagnosing of COVID-19 can be achieved using three different methodologies as depicted in Fig. 2. These three methodologies are (1) Real-Time reverse transcriptase- Polymerase Chain Reaction (RT-PCR) (Tahamtan and Ardebili 2020; Waller et al. 2020; Li et al. 2020a, d), (2) chest CT imaging scan (Mishra et al. 2020; Li et al. 2020a, e; Kovács et al. 2020), and (3) numerical laboratory tests (Brinati et al. 2020; Kukar et al. 2020; Cabitza et al. 2020; Qiu et al. 2020). RT-PCR tests are fairly quick, sensitive, and reliable. The sample is collected from a person's throat or nose; adding some chemicals for removing any proteins, fats, and other molecules, leaving behind only the existing Ribonucleic Acid (RNA) (Huang et al. 2020). The separated RNA is a mixture of a person's RNA and the coronavirus's RNA if exists. Despite its popularity, the RT-PCR test suffers from the risk of false-negative and false-positive results (Chen et al. 2020a, b; Kasteren et al. 2020).

Although several studies had observed that the sensitivity of Chest CT in the diagnosing of COVID-19 is higher than that of RT-PCR, the American College of Radiology (ACR) has issued guidance that CTs and X-rays are not accurate tools for diagnosing COVID-19 (Gietema et al. 2020). There are three significant reasons for ACR's recommendation, which are; (1) both chest CT and

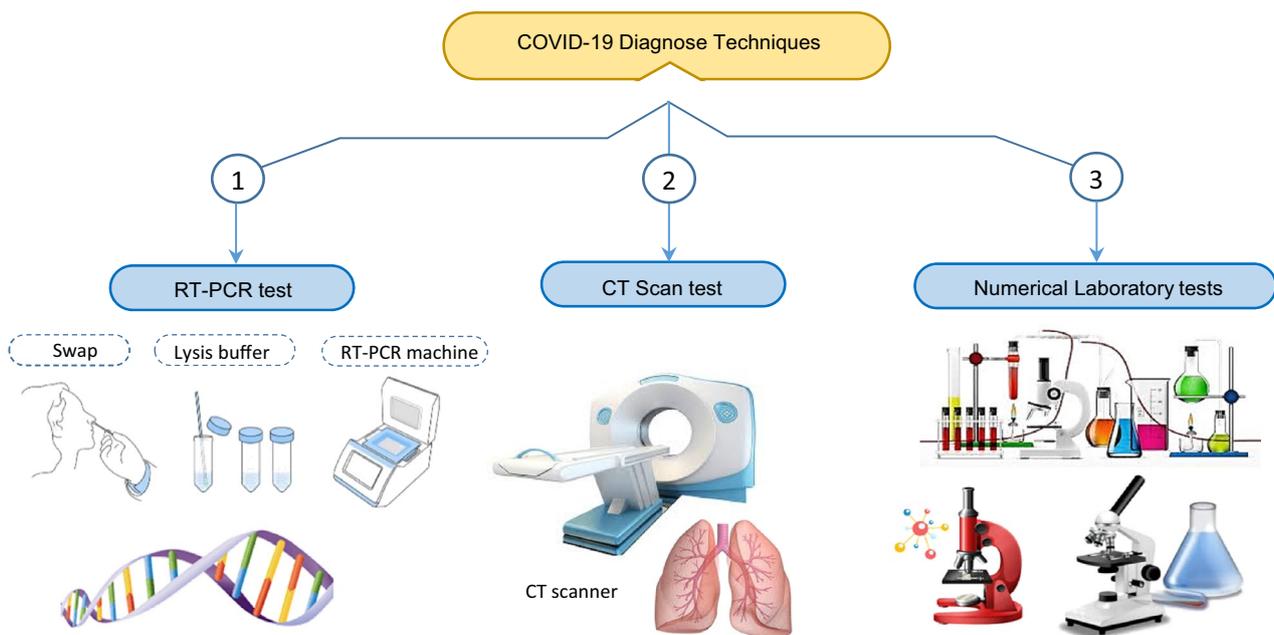


Fig. 2 Different COVID-19 diagnosis techniques

X-ray cannot accurately distinguish between COVID-19 and other respiratory infections. They can only point to signs of an infection, which could be due to other reasons such as seasonal flu. (2) A huge number of patients infected with COVID-19 have normal chest CTs, which wrongly convince them that they are healthy. Those convince patients can easily spread the virus to others. (3) The usage of the imaging equipment on COVID-19 patients is a critical hazard for doctors and other patients. CT scanners are complex and large machinery pieces (Gietema et al. 2020). They need to be carefully cleaned after each potential COVID-19 patient.

However, even with precise cleaning, there is a high risk that the virus could remain on the surface of the CT scanner room. Moreover, the movement of COVID-19 patients to and from a CT scanner room increases the risk of spreading the virus inside of the healthcare system. On the other hand, the use of accurate Numerical Laboratory Tests (NLTs) can be considered as the most accurate method for diagnosing COVID-19. Recently, the use of NLTs is the only method that the Centers for Disease Control (CDC) currently endorse. Hence; it makes perfect sense that the use of NLTs will provide more accurate diagnosis with less waiting time. The work in this paper is concentrated on providing a new COVID-19 diagnosis system based on NLTs, which have proven to be the most effective methodology for COVID-19 diagnosis. A new diagnosis strategy called FCNB will be introduced, which is based on the weighted Naïve Bayes algorithm with several modifications.

4 Related work

Recently, there has been extensive research on COVID-19 patients detection.

A Textual Clinical Reports Classification (TCRC) model was provided by Khanday et al. (2020) for detecting COVID-19, Severe Acute Respiratory Syndrome (SARS), Acute Respiratory Distress Syndrome (ARDS), and both (COVID-19, ARDS) by using different classical and ensemble machine learning methods. The experimental results showed that the logistic regression and multinomial Naïve Bayes provided the best results compared to other machine learning algorithms.

Ozturk et al. (2020a) developed a Deep Learning (DL) model to detect COVID-19. The proposed model was implemented on the dataset that consists of three classes called; COVID-19, pneumonia, and normal X-ray imagery. This study passed through two main steps, which are; pre-processing step and the classification step. In the pre-processing step, the fuzzy coloring method was used to re-structure the data classes and the structured images were stacked. In the classification step, deep learning models (MobileNetV2, SqueezeNet) were trained and then the social mimic optimization technique was used to obtain a set of efficient features. These efficient features were combined to provide the classification by using Support Vector Machines (SVM) as a classifier. The experimental results proved that the proposed classification model could efficiently detect the COVID-19 disease.

Maghdid et al. (2020) introduced a Convolution Neural Network (CNN) model to detect COVID-19 cases based on chest X-ray and CT images dataset. The proposed CNN model contained two main algorithms called CNN architecture and AlexNet as a transfer-learning algorithm. Although the simplicity of this proposed model, its accuracy is not enough for the diagnosing of COVID-19 patients. The experimental results illustrated that the maximum accuracy of the utilized models was provided by using a pre-trained network, but the minimum accuracy was provided by using the modified CNN.

Chen et al. (2020a, b) introduced a COVID-19 Diagnostic Model (CDM) based on radiological semantic and clinical features without the need for the nucleic acid test. The experimental results demonstrated the effectiveness of the proposed CDM technique for the diagnosing of COVID-19 cases in which CDM provided better diagnostic performance and more considerable net benefits.

Waheed et al. (2020) proposed an Auxiliary Classifier Generative Adversarial Network (ACGAN) based GAN called CovidGAN to produce synthetic Chest X-Ray (CXR) images. The synthetic images generated from CovidGAN were utilized to enlarge the dataset and to enhance the performance of Convolutional Neural Network (CNN) for COVID-19 detection. The experimental results proved that the accuracy of the usage of CNN based on the synthetic images generated from CovidGAN was better than the accuracy of using CNN alone. Although the proposed detection model provided the best accuracy, it depended on a small dataset. Additionally, the quality of the synthetic samples needed to be improved by adding more labeled data, which increased the learning process of GAN.

An Automatic COVID-19 Detection Model (ACDM) based on using the DarkNet model as a classifier was provided by Ozturk et al. (2020b). The proposed ACDM method was used as a new detection method based on using chest X-ray images. This model represented the development of deep learning techniques to be able to perform both binary and multi-class classification. The experimental results demonstrated that the effectiveness of ACDM to perform binary tasks was better than its effectiveness in performing

multi-class tasks as the accuracy of binary was higher than in multi-class.

Sun et al. (2020) presented an Adaptive Feature Selection guided Deep Forest (AFS-DF) based on using chest CT images was introduced to classify COVID-19 patients. For learning a high-level representation of features, the AFS-DF method used a deep forest model. Based on the trained forest, an adaptive feature selection operation was used to decrease the redundancy of the features for improving the performance of the classification process. The experimental results showed that the AFS-DF model outperformed several existing methods in which it could efficiently classify COVID-19 cases based on CT images. Table 1 illustrates a comparative study of the previous efforts on COVID-19 patients detection methods.

5 Weighted Naïve Bayes

No doubt, Naïve Bayes (NB) is a popular classifier that had been applied in several domains such as; weather forecasting, bioinformatics, image and pattern recognition, and medical diagnosis. NB allows each feature to contribute towards the classification decision both equally and independently of other features. Although such simplicity increases computational efficiency, it sometimes makes NB insufficient with real-world conditions.

Consider $F = \{f_1, f_2, f_3, \dots, f_n\}$ to be a set of feature vectors of a new item IC to be classified and $C = \{c_1, c_2, c_3, \dots, c_m\}$ be set of target classes. The probability of a new item being in class c_j using NB is given by (1) (Berrar 2018; Taha et al. 2013)

$$Target(IC) = \operatorname{argmax}_{c_j \in C} [P(c_j|F)] = \operatorname{argmax}_{c_j \in C} \left[\frac{P(F|c_j) \times P(c_j)}{P(F)} \right] \tag{1}$$

where, $P(c_j|F)$ is the conditional probability of class c_j given the feature vector F (also called posterior probability), $P(F|c_j)$ is the conditional probability of class F given the class c_j (also called likelihood), and $P(c_j)$ is the prior probability of class c_j . Since features are independent, this yields;

$$P(F) = P(f_1, f_2, f_3, \dots, f_n) = P(f_1) \times P(f_2) \times P(f_3) \times \dots \times P(f_n) = \prod_{i=1}^n P(f_i)$$

$$P(F|c_j) = P(f_1, f_2, f_3, \dots, f_n|c_j) = P(f_1|c_j) \times P(f_2|c_j) \times P(f_3|c_j) \times \dots \times P(f_n|c_j) = \prod_{i=1}^n P(f_i|c_j)$$

Table 1 A comparative study of the previous efforts on COVID-19 patients detection methods

Used technique	Description	Advantages	Disadvantages
Textual Clinical Reports Classification (TCRC) model (Khanday et al. 2020)	TCRC was provided for detecting COVID-19, SARS, ARDS, and both (COVID-19, ARDS) by using different classical and ensemble machine learning methods	The technique used to distinguish between four viruses, and many ensemble machine learning methods are used	The dataset that is used is small. More feature engineering is needed in order to get better results, and the best accuracy for the detection is not achieved
Deep Learning (DL) model (Ozturk et al. 2020a)	DL Models (<i>MobileNetV2</i> and <i>SqueezeNet</i>) were trained on the chest images dataset, pneumonia chest images, and normal chest images after preprocessing them. In the preprocessing step, the fuzzy technique and stacking technique were performed to reconstruct the dataset. Then, SVM was used to classify the dataset after trained with DL models	The parameters in the used DL models (<i>MobileNetV2</i> and <i>SqueezeNet</i>) are few compared to other DL models. Thus, these models take low time and increase the speed of the process by using the Social Mimic Optimization (SMO) algorithm and CNN algorithm. DL provides effective features with stacking technique and decreases the interference in every image in the dataset Easy to implement	The resolution dimensions of the original images and the structured images should be the same in the Stacking technique, and a complete success cannot be achieved because the size of the input images in the dataset is different
Convolution Neural Network (CNN) model (Maghdid et al. 2020)	CNN was introduced to detect COVID-19 cases based on using chest X-ray and CT images dataset. The proposed CNN model contains two main algorithms called CNN architecture and AlexNet as a transfer learning algorithm		The accuracy of the proposed models is not enough to diagnose COVID-19
COVID-19 Diagnostic Model (CDM) (Chen et al. 2020a, b)	CDM was introduced based on radiological semantic and clinical features without the need for the nucleic acid test	The clinical and radiological semantic models give a better detection performance and more considerable net benefits	Patient selection has a potential bias, and sample size is small
CovidGAN model (Waheed et al. 2020)	CovidGAN model was designed for the automatic detection of COVID-19 by producing synthetic Chest X-Ray (CXR) images to classify them into two classes namely; COVID-CXR and Normal-CXR	The generated synthetic CXR images from CovidGAN improved the classification performance of CNN, and synthetic data augmentation increase the variability to the dataset by enlarging it	The dataset is small, and the quality of the synthetic samples need to be improved by adding more labeled data which increases the learning process of GAN
Automatic COVID-19 Detection (ACDM) model (Ozturk et al. 2020b)	ACDM method was used as a new detection method based on using chest X-ray images. This model represented a development of deep learning techniques to be able to perform both binary and multi-class Classification	Chest X-ray images are classified without using feature extraction techniques. Expert radiologist evaluates the heatmaps generated by the model that focus on localizing effective areas on chest X-ray images	The COVID-19 public image data has limited data
Adaptive Feature Selection guided Deep Forest (AFS-DF) model (Sun et al. 2020)	AFS-DF based on using chest CT images was introduced to classify COVID-19 patients. For learning high level representation of features, AFS-DF method used a deep forest model. Based on the trained forest, an adaptive feature selection operation was used to decrease the redundancy of features for improving the performance of classification process	The size of dataset is large	The features are extracted depending on the prior knowledge in the current work. To enhance the performance, this can be done by using a deep learning model

Substitute in (1), this yield (2) (Jabeen et al. 2019)

$$Target(IC) = \operatorname{argmax}_{c_j \in C} \left[\frac{P(c_j) \times \prod_{j=1}^n P(f_i|c_j)}{\prod_{i=1}^n P(f_i)} \right] \tag{2}$$

Since the denominator in (2) remains constant for a given input for all target classes, it can be removed as illustrated in (3) (Zhang et al. 2021; Subramanian and Prabha 2020; Abellán and Castellano 2017)

$$Target(IC) = \operatorname{argmax}_{c_j \in C} \left[P(c_j) \times \prod_{i=1}^n P(f_i|c_j) \right] \tag{3}$$

However, the performance of NB is sometimes low due to the unrealistic assumption that all features are independent and equally important given the class value. The performance of NB can be increased by mitigating this assumption. Many improvements have been proposed to resolve this problem including feature selection and feature weighting. Generally, feature selection can be applied to enhance the performance of the traditional Naïve Bayes classifier. Hence, the target class can be identified by (4) (Lee et al. 2011).

$$Target(IC) = \operatorname{argmax}_{c_j \in C} \left[P(c_j) \times \prod_{i=1}^n P(f_i|c_j)^{S_i} \right] \tag{4}$$

where $S_i \in \{0,1\}$

However, assigning an equal value of weight to all considered features breaks the nature of real-world applications. Accordingly, different weights can be assigned to each feature as a generalization of feature selection as illustrated in (5) (Yu et al. 2019; Jiang et al. 2019)

$$Target(IC) = \operatorname{argmax}_{c_j \in C} \left[P(c_j) \times \prod_{i=1}^n P(f_i|c_j)^{W_i} \right] \tag{5}$$

where $W_i \in \mathbb{R}^+$

As depicted in (5), unlike traditional NB, each feature f_i has its weight w_i , which can be a positive number representing the significance of the feature. However, both traditional and Weighted Naïve Bayes (WNB) classifiers are based mainly on probabilities, namely; the conditional probabilities of the input features given the considered target classes as well as the classes prior probabilities. From another point of view, promoting the performance of the WNB classifier can be achieved by compensating its performance with another heuristic besides conditional and prior probabilities.

6 The proposed Feature Correlated Naïve Bayes (FCNB) classification strategy

According to the rapid growth of COVID-19, the detection of this virus is an important process for healthcare organizations. Fast and accurate COVID-19 detection will be more helpful to decrease the alarming effect of this pandemic and will support in designing good strategies and taking productive decisions (Shinde et al. 2020). As illustrated in Fig. 3, the FCNB strategy composes of two stages, which are; (1) Pre-Processing Stage (P²S) and (2) Classification Stage (CS).

During P²S, three main processes are performed on the collected data by applying data mining techniques to provide a meaningful pattern of data. These three processes are called Feature Selection Phase (FSP), Feature Clustering Phase (FCP), and Master Feature Weighting Phase (MFWP). Thus, P²S gives only the most informative data that enables the next stage which is called CS to detect early and accurately COVID-19 cases. On the other hand, during CS, fast and accurate COVID-19 diagnosis is provided by using Feature Correlation Naïve Bayes Phase (FCNBP) that uses a new weighted NB with many modifications. Finally, the FCNB strategy consists of four phases called FSP, FCP, MFWP, and FCNBP in which the first three phases are included in P²S while the last phase is presented in CS. In the next sections, there will be a detailed description of the P²S, CS stages, and a related discussion of the key algorithms.

6.1 Pre-processing stage (P²S)

Data pre-processing plays an important role in providing fast, useful, and accurate decisions for detecting COVID-19 cases. Accordingly, the clinical features of this pandemic must be known and well understood. In P²S, three main phases called FSP, FCP, and MFWP are performed on the collected data to provide the most informative data that helps the detection method to quickly and accurately detect COVID-19 patients. The FSP as the first phase in P²S aims to select the most effective features on COVID-19 diagnosis. The FCP as the second phase in P²S aims to put the selected features into groups. Finally, the MFWP aims to assign a weight value to each master feature for the next COVID-19 classification stage.

6.1.1 Feature selection phase (FSP)

Usually, records of patients contain many features used to support the medical diagnosis. However, for the early COVID-19 detection task, not all of these features have the same importance. The performance of the diagnostic

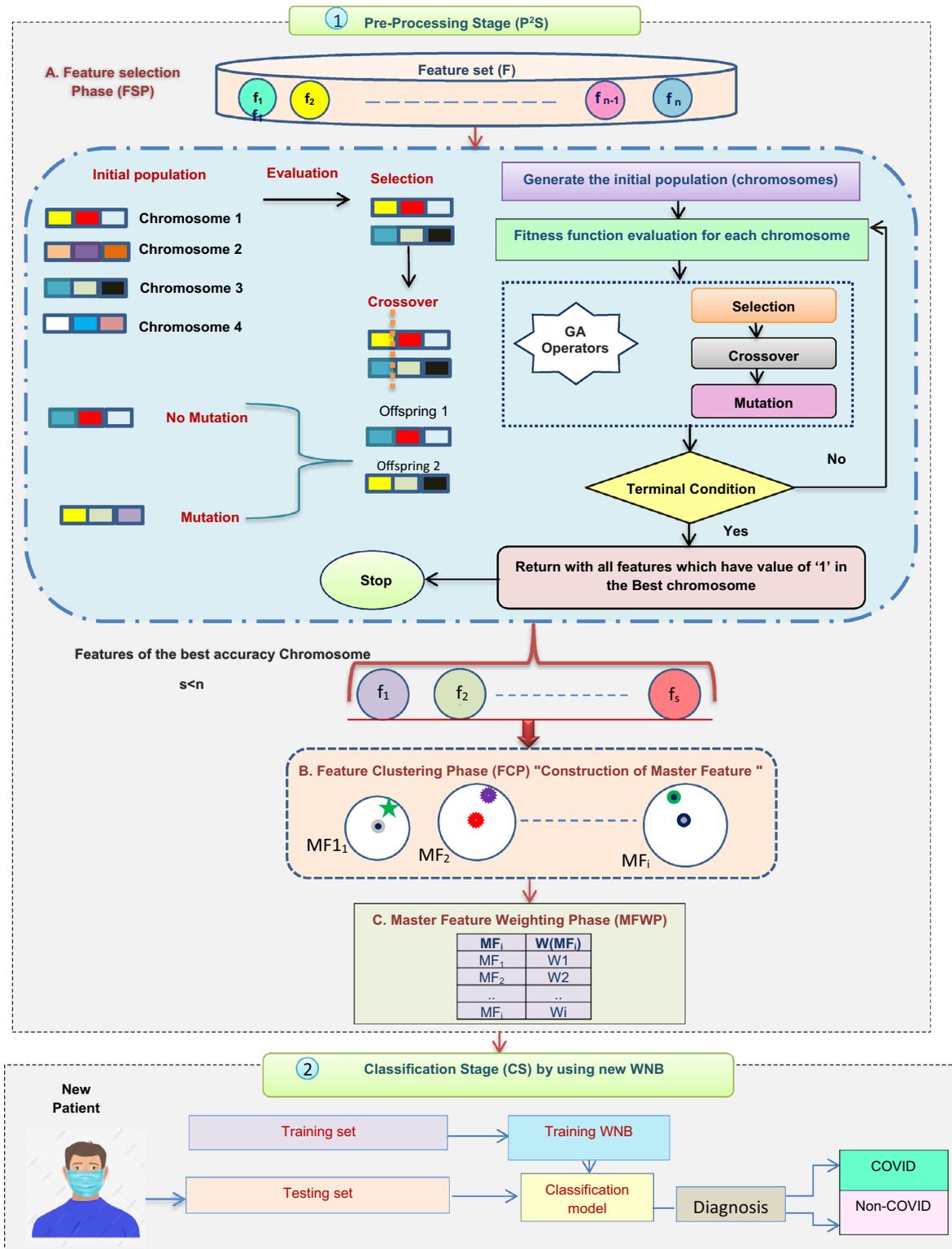


Fig. 3 The proposed FCNB classification strategy

operation may rely on the selected features in all phases of the FCNB. Hence, the main objective of FSP is to eliminate the irrelevant features and select the best features before using the diagnostic model. Selecting the best features will improve the performance of the machine learning algorithm, decrease the time of processing, increase the computational efficiency, minimize the storage requirement, and increase the convergence of learning (Wosiak and Zakrzewska 2018; Saleh et al. 2016). In this paper, the considered methodology to select the most effective subset of features on COVID-19 is Feature Selection based on Genetic Algorithm (FSGA) methodology. FSGA is a wrapper method used to select the most important features depending on specific evaluation metrics.

Unlike classical selection methods which search from a single point and can deal poorly with large search spaces, FSGA depends on GA that can discover the global optimal solution and prevent the trapping in local optimal solution (Sivanandam and Deepa 2008). To implement FSGA, consider that the Feature set (F) of 'n' features can be expressed by $F = \{f_1, f_2, f_3, f_4, \dots, f_n\}$, where the input training data set of 'k' patients can be expressed by $I = \{I_1, I_2, I_3, I_4, \dots, I_k\}$. Additionally, the testing dataset of 'q' patients can be expressed by $G = \{G_1, G_2, G_3, G_4, \dots, G_q\}$. Each patient of $Y_j \in I$ and $R_i \in G$ is expressed as an ordered set of 'n' features; $Y_j(f_1, f_2, f_3, f_4, \dots, f_n) = [f_{1j}, f_{2j}, f_{3j}, f_{4j}, \dots, f_{nj}]$ and $R_i(f_1, f_2, f_3, f_4, \dots, f_n) = [f_{1i}, f_{2i}, f_{3i}, f_{4i}, \dots, f_{ni}]$. Accordingly, each training patient Y_j and testing patient R_i can be expressed in an 'n' dimensional space of features. For the considered COVID-19 detection problem, it is important to use FSGA as a suitable feature selection methodology to reduce or

eliminate the irrelevant features to enhance the performance of the classifier.

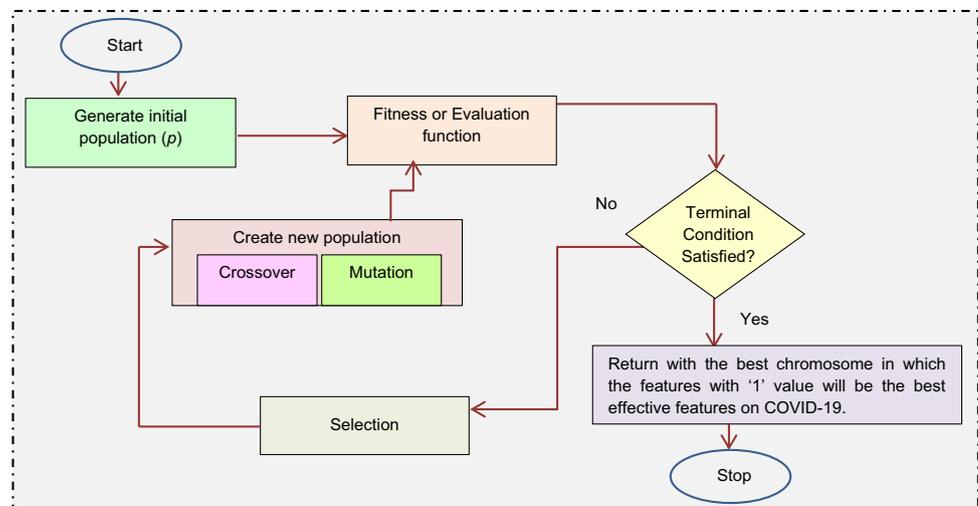
After extracting the features from laboratory tests for both COVID-19 patients and non -COVID-19 people, the collected dataset should be passed to FSGA for selecting the most effective features on COVID-19 cases. The FSGA depends on applying GA as it is an optimization technique and adaptive search heuristic algorithm that followed the process of natural evolution. The GA starts with a population of potential solutions, and then it employs the concept of survival of the fittest to generate the closest optimal solutions according to a fitness function of an optimization problem (Saleh et al. 2016; Oluleye et al. 2014). Hence, FSGA begins with an initial population, which is a group of candidate solutions, or chromosomes in which every chromosome composes of series of genes. The value '1' of a gene denotes that the feature is selected in the particular subset. Otherwise, the value '0' of a gene denotes that the feature is eliminated from the particular subset (Saleh et al. 2016; Kaviani and Dhotre 2017). Consider that a single chromosome has 'n' genes (i.e., the same number of features in the dataset), hence; $F = \{f_1, f_2, f_3, f_4, \dots, f_n\}$. Assume that "n = 15 features", thus, a single chromosome can be represented as; $\{f_1, f_2, f_3, f_4, \dots, f_{15}\}$ as shown in Table 2. The biological functions (three operators of FSGA) such as selection, crossover, and mutation are applied to these chromosomes to produce a new generation of the population. These three operators are repeated until a termination condition has been satisfied.

The accuracy of NB is evaluated to be used as a fitness function in FSGA for choosing the best chromosome that

Table 2 The representation of a single chromosome

f ₁	f ₂	f ₃	f ₄	f ₅	f ₆	f ₇	f ₈	f ₉	f ₁₀	f ₁₁	f ₁₂	f ₁₃	f ₁₄	f ₁₅
0	1	0	1	1	0	1	0	1	0	1	0	0	1	0

Fig. 4 The steps of FSGA implementation



includes the most effective features on COVID-19. The main objective of selecting the best subset of the features is to achieve the highest accuracy of the used COVID-19 detection model. Finally, there are many steps to implement FSGA as presented in Fig. 4. At first, the initial population (p) of FSGA is represented by many candidate solutions, which are called chromosomes. Each chromosome consists of genes; each gene represents a feature in COVID-19's dataset. The existence or absence of a feature is determined by the value of the gene, where the value equals '1' means the feature is existing, and '0' means the feature is eliminated. Secondly, NB's accuracy as a fitness function is calculated for each chromosome (candidate solution) in p to provide a fitness value that indicates the goodness of the solution. The optimal solution is the solution that maximizes the fitness function.

Based on the fitness values, the selection of parent members (chromosomes) for reproduction is done according to the probability of selection (p_{sel}). After that, the crossover between the parent members is done to produce the offspring according to the probability of crossover (p_{cross}). According to the probability of mutation (p_{mut}), the mutation is performed for each offspring. Loops over these steps are repeated from the selection until the size of the next population equals the size of the initial population. If the terminal condition is not satisfied, the previous steps will be repeated from the fitness function. In the end, when the terminal condition is satisfied, the chromosomes in the population will be evaluated as the final results by using only the fitness function. Then, the chromosome that provides the highest fitness value contains the best subset of features denoted by '1' value. The steps to implement FSGA are illustrated in Algorithm 1.

Features selection based on Genetic Algorithm (FSGA)

- **Inputs:**
 P_s = size of population (no. of candidate solutions or chromosomes).
 $C = \{C_1, C_2, C_3, \dots, C_{ps}\}$; Set of chromosomes in population.
 P_{sel} = probability of selection.
 P_{cross} = probability of crossover.
 P_{mut} = probability of mutation.

- **Output:**
 BC = the best chromosome with the highest accuracy.

• **Steps:**

- // **Generate the initial population.**
 1. Randomly generate the initial population (chromosomes).
- // **Evaluate the fitness function for each chromosome.**
 2. Determine the fitness for each individual $\in p$.
- // **Use Roulette wheel method to estimate the selection of population.**
 3. Define the probability of distribution (P) over the individual of P
 Where $P_d(C) \propto F(C)$.
- 4. Select parent members C_i, C_j according to P_d, P_{sel} where $i, j \in P_s, i \neq j$.
- // **Perform crossover by using "single point crossover."**
 5. Produce a new population members C_i^*, C_j^* according to P_{cross} .
- // **Perform mutation using "flip bit mutation."**
 6. Apply mutation to C_i^*, C_j^* according to P_{mut} .
- 7. Put C_i^* and C_j^* in the next population (P_{next}).
- 8. If (no. of chromosomes in $P_{next} < P$) then
- 9. | Go to step 4.
- 10. Else
- 11. | Let $p \leftarrow P_{next}$.
- 12. End if

Algorithm Parameters	
P	Initial population.
P_s	Size of population.
C	Set of chromosomes in population ; $C = \{C_1, C_2, C_3, \dots, C_{ps}\}$.
P_{sel}	Probability of selection.
P_{cross}	Probability of crossover.
P_{mut}	Probability of mutation.
BC	Best chromosome in the population.
$F(C)$	Fitness function of the chromosome C .
P_d	Probability of distribution.
$P_d(C)$	Probability distribution of the chromosome C .
P_{next}	Next generation of the population.
C^*	Set of new chromosome in next population ; $C^* = \{C_1^*, C_2^*, \dots, C_{ps}^*\}$.

- // **Check the termination condition.**
 13. If (there are more generations) then
- 14. | Go to step 2.
- 15. Else
- 16. | Return with all features which have value of '1' in BC .
- 17. End if

Table 3 A brief descriptions about the features which have an effect on COVID-19 patients

Feature	Abbreviations	Description	Unit
White blood cells counts	WBC	WBC is used to determine the number of the white blood cell to detect the infections in the body	$\times 10^9$ cells/L
Neutrophil Count	NC	Neutrophil is a part of WBC which occupy (50%-75%) of it. NC provides the important information about the health status of the patient	$\times 10^9$ cells/L
Eosinophil Count	EC	Eosinophil is a type of white blood cells that help in curb infections and boost inflammation, which can help the immune system fight off a disease	$\times 10^9$ cells/L
Monocytes Count	MC	Monocytes are a kind of white blood cell. Monocytes test measures the amount of monocytes that circulating in the blood,	$\times 10^9$ cells/L
Platelet Count	PC	PC is a blood test that determines the average number of platelets in the blood. Platelets help the blood cure wounds and stop the over bleeding	$\times 10^9$ cells/L
Lymphocyte Count	LC	LC determines the count of lymphocyte which is the part of WBC	$\times 10^9$ cells/L
Basophils Count	BC	Basophils are white blood cells from the bone marrow that play an important role in keeping the immune system working correctly	$\times 10^9$ cells/L
Lactate Dehydrogenase	LDH	LDH test is used for detecting any damage in the tissue	U/L
aspartate aminotransferase	AST	AST is an enzyme that make by the liver, the high levels of AST indicates damage in the liver. AST test is a blood test that checks for liver disease	U/L
alkaline phosphatase	ALP	ALP is an enzyme exists in many tissues throughout the body. Abnormal level of ALP is a sign for a problem with liver, gallbladder, or bones. ALP test measures the amount of the enzyme in the blood	U/L
Gamma-glutamyl -transpeptidase	GGT	GGT a common enzyme that exists in many parts of body. High levels of GGT in the blood may be an indication of damage to the bile ducts or liver disease; GGT test determines the amount of GGT in the blood	U/L
Alanine aminotransferase	ALT	ALT test is a blood test that detect if there is a failure in the liver	U/L
C-reactive protein	CRP	CRP is a protein produced by the liver, CRP test used to detect or monitor conditions that cause inflammation	mg/L
Age	Age	To determine the age of the person	Years
Gender	Gender	To determine Male or Female	–

Generally, records of patients can be used to represent the data in supervised learning. Each record is described by a set of features. These features take one of two types, which are; “nominal” or “numeric” values. While nominal values represent members of an ordered set, and numeric values represent real numbers. In fact, the number of features which affected COVID-19 patients is “15” features ($n = 15$) as described in Table 3. As an illustrative example, a nominal dataset of 25 patients as well as the features, which affected them, are represented in Table 4. For simplicity, each patient in Table 4 has been described by a subset of features presented in Table 3. This subset of features contains ‘6’ features, which are; Platelet Count (PC), White Blood cell (WBC), Monocytes Count (MC), Aspartate aminotransferase (AST), Basophils Count (BC), and Lactate Dehydrogenase (LDH). Hence, the features in Table 4 are represented as; $F = \{f_1, f_2, f_3, f_4, f_5, f_6\} = \{PC, WBC, MC, AST, BC, LDH\}$ in which nominal values of each feature is presented in Table 5. Each patient has a class label, which indicates one of the two target classes “True, False”. True indicates COVID-19 Patient and False indicates non-COVID-19 People. The

first 15 records in Table 4 represent the training dataset (e.g., $k = 15$); $I = \{I_1, I_2, I_3, I_4, \dots, I_{15}\}$ and the last 10 records represent the testing dataset (e.g., $q = 10$); $G = \{G_1, G_2, G_3, G_4, \dots, G_{10}\}$.

To implement the NB classifier, it is essential to create the frequency distribution tables (also called “contingency tables”) to construct the relationships between the features and the class categories (Huang et al. 2020; Saleh et al. 2016). Tables 6a–f are the frequency distribution tables that represent the relationships between the features and the class categories in the considered “COVID-19” dataset. Tables 6a–f, are used to calculate the probabilities which are used to apply the NB equation.

The FSGA is a feature selection method that is used on the “COVID-19” dataset in Table 4 to choose the most effective features on COVID-19 patients. The accuracy of the NB classifier is used as a fitness function to evaluate each chromosome in the population of FSGA in which NB’s accuracy can be calculated by using the confusion matrix (Saleh et al. 2016; Visa et al. 2011). There are many assumptions to implement FSGA on the considered “COVID-19” dataset as presented in Table 7. Based on the previous assumptions,

Table 4 The “COVID-19” dataset with nominal values

Patient #	Features of COVID-19						Diagnosis
	PC	WBC	MC	AST	BC	LDH	
1	Low	Low	Low	High	Normal	Normal	True
2	Low	Low	Normal	High	Normal	High	True
3	Low	High	Normal	High	Normal	Normal	False
4	Low	High	Normal	High	High	Normal	True
5	Low	Normal	High	High	Normal	Normal	False
6	Low	Normal	Normal	High	Normal	High	True
7	Normal	Low	Low	High	Normal	Normal	True
8	Normal	High	Normal	High	Normal	Normal	False
9	Normal	High	Normal	High	High	High	True
10	Normal	Normal	High	High	Normal	Normal	False
11	Normal	Normal	High	High	Normal	High	True
12	High	Low	Low	Normal	Normal	Normal	True
13	High	Normal	High	Normal	Normal	Normal	False
14	High	Normal	High	Normal	High	High	True
15	High	High	Normal	Normal	Normal	High	True
16	Low	Normal	High	High	High	Normal	False
17	Normal	Normal	High	High	High	Normal	False
18	High	Low	Low	Normal	Normal	High	True
19	Normal	Normal	Normal	High	Normal	Normal	False
20	Normal	High	Normal	High	Normal	High	True
21	Normal	Low	Normal	High	Normal	High	True
22	Low	High	Normal	High	High	High	True
23	Low	Low	Low	High	High	High	True
24	High	High	Normal	Normal	Normal	Normal	True
25	High	Normal	Normal	Normal	Normal	Normal	False

Table 5 Nominal values of each feature

Feature of COVID-19	Corresponding nominal values
PC	Low
	Normal
	High
WBC	Low
	Normal
	High
MC	Low
	Normal
	High
AST	High
	Normal
BC	High
	Normal
LDH	High
	Normal

best chromosome contains ‘3’ features; $F = \{WBC, AST, LDH\}$. After applying FSGA methodology on the original set of features ($n = 15$) in the considered “COVID-19” dataset, the number of the selected features becomes “10” features. Thus, a new set of features that contains “10” features instead of “15” features is represented as; $F = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}\} = \{WBC, AST, LDH, LC, ALT, CRP, EC, ALT, NC, GGT\}$.

6.1.2 Feature clustering phase (FCP)

FCP is the second phase in the P²S that is used to cluster the selected features from the FSP into many groups where each group contains similar features. Clustering is the main analytical technique in data mining in which data clustering is a procedure that is used to classify the data into homogenous groups based on similarity. Thus, the data in the same cluster are similar, but it should be different as much as possible according to different clusters (Bano and Khan 2018; Ayed et al. 2015). Clustering methods can be categorized into several techniques, which are; partitioning based algorithms, model-based algorithms,

after employing FSGA, the steps followed in the first and the second iterations are illustrated in Figs. 5 and 6 respectively. Finally, the best subset of features according to the

Table 6 The frequency distribution tables compiled from “COVID-19” dataset

Feature		Diagnoses		Σ	Probability in class	
		True	False		True	False
<i>(a) Platelet count (PC)</i>						
PC	Low	4	2	6	4/10	2/5
	Normal	3	2	5	3/10	2/5
	High	3	1	4	3/10	1/5
Σ		10	5	15		
<i>(b) White blood cell (WBC)</i>						
WBC	Low	4	0	4	4/10	0/5
	Normal	3	3	6	3/10	3/5
	High	3	2	5	3/10	2/5
Σ		10	5	15		
<i>(c) Monocytes count (MC)</i>						
MC	Low	3	0	3	3/10	0/5
	Normal	5	2	7	5/10	2/5
	High	2	3	5	2/10	3/5
Σ		10	5	15		
<i>(d) Aspartate aminotransferase (AST)</i>						
AST	High	7	4	11	7/10	4/5
	Normal	3	1	4	3/10	1/5
Σ		10	5	15		
<i>(e) Basophils count (BC)</i>						
BC	High	3	0	3	3/10	0/5
	Normal	7	5	12	7/10	5/5
Σ		10	5	15		
<i>(f) Lactate dehydrogenase (LDH)</i>						
LDH	High	6	0	6	6/10	0/5
	Normal	4	5	9	4/10	5/5
Σ		10	5	15		

Table 7 The assumptions for employing FSGA

No.	Assumption	Value
1	No. of generation to process	2
2	Population size	4 “no. of chromosomes”
3	Probability of selection “ P_{sel} ”	various value for each selected chromosome
4	Probability of crossover “ P_{cross} ”	0.93
5	Probability of mutation “ P_{mut} ”	0.15
6	Chromosome size “C”	6 “no. of features” (n)

density-based algorithms, hierarchical-based algorithms, and grid-based algorithms (Benabdellah et al. 2019). In this vein, many applications used clustering techniques such as pattern recognition, image processing, disease detection, etc. The similarity between data items can be measured by using distance metrics. Thus, there are many distance functions, which are used to define a distance between items or elements. These distance functions such as; Cosine similarity (Shirchorshidi et al. 2015), Jaccard

distance (Fletcher and Slam 2018), Manhattan distance (Pandit and Gupta 2011), Euclidean distance (Dokmanic et al. 2015), etc.

The cluster is constructed in a way that any two data items associated with the same cluster have the minimum value of distance and any two data items associated with different clusters have the maximum value of distance (Zhu et al. 2019). Although the simplicity of clustering techniques, these techniques suffer from many challenges such as; determining the number of clusters, selecting the centroid of each cluster, and choosing the similarity measurement. Thus, it is essential to introduce a new clustering method to overcome these pre-mentioned challenges. In the FCP, there are three main steps to cluster the features, which are; (1) Construct actual clusters, (2) Isolated feature assignment, and (3) Weighting features. The proposed clustering method will be implemented on the selected features by FSGA in FSP; $F = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}\} = \{WBC, AST, LDH, LC, ALT, CRP, EC, ALT, NC, GGT\}$.

To construct clusters of similar features, assume that the radius of each cluster is called Zone Half Diameter (ZHD)

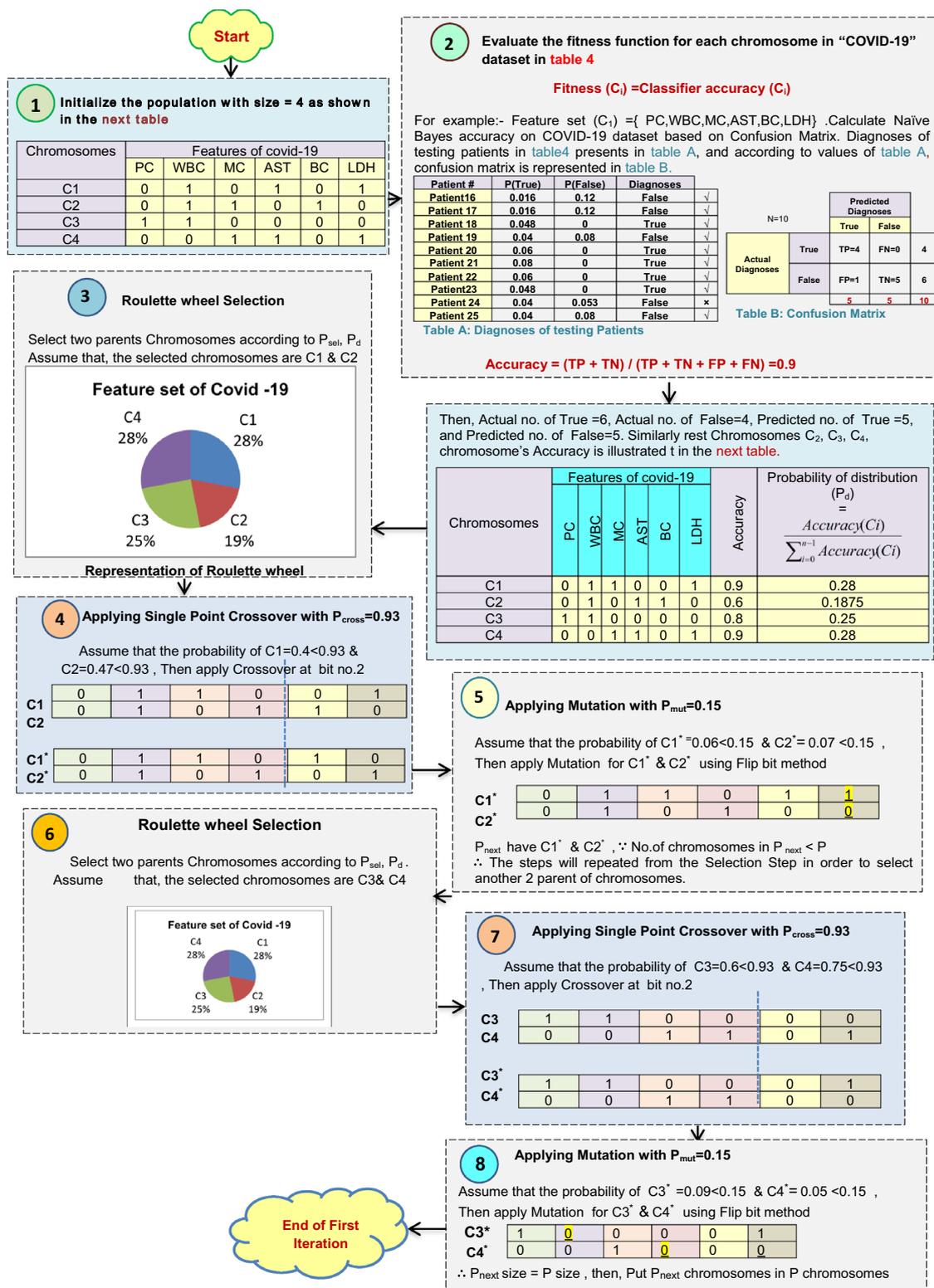


Fig. 5 The followed steps for the first iteration of FSGA

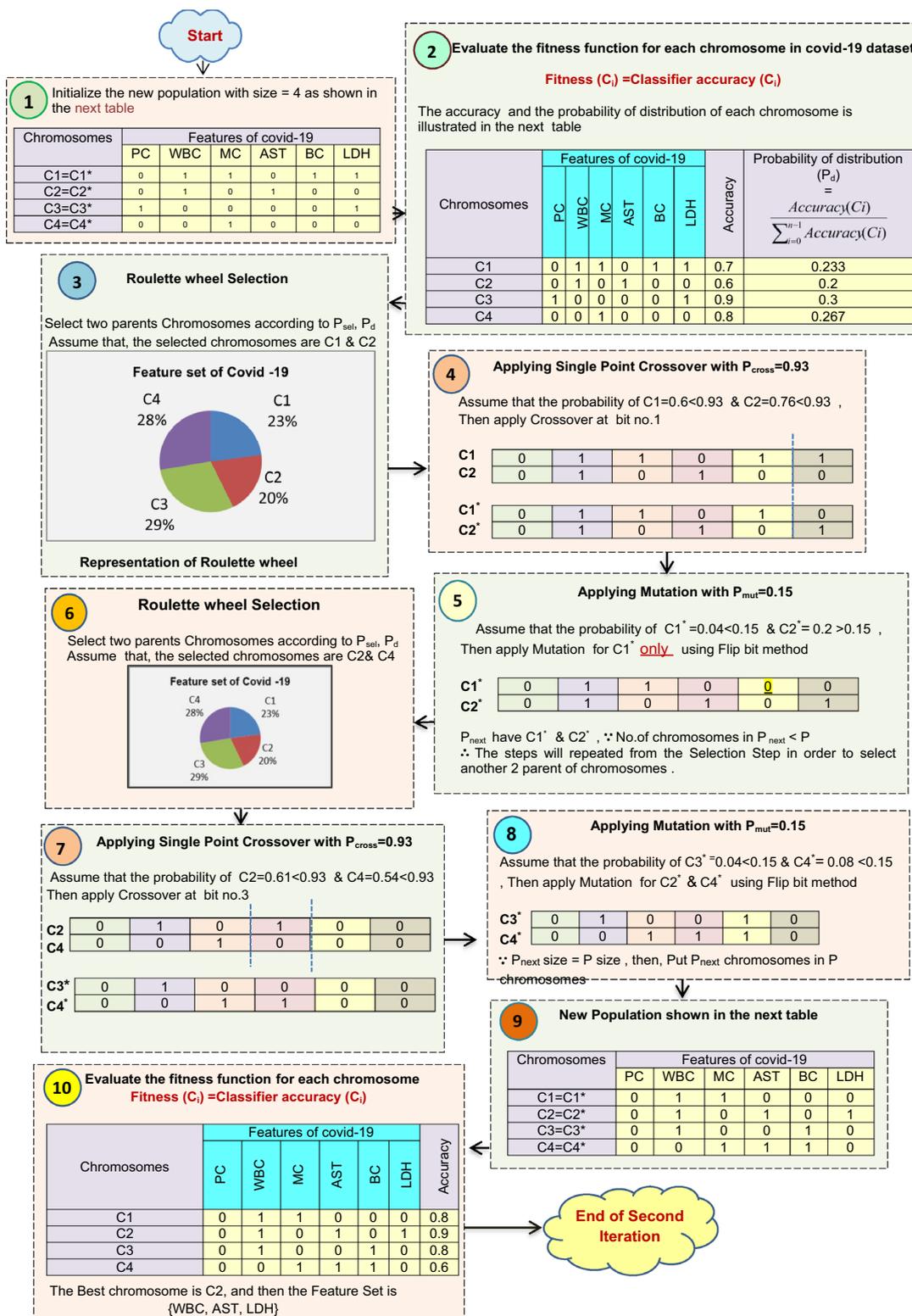


Fig. 6 The followed steps for the second iteration of FSGA

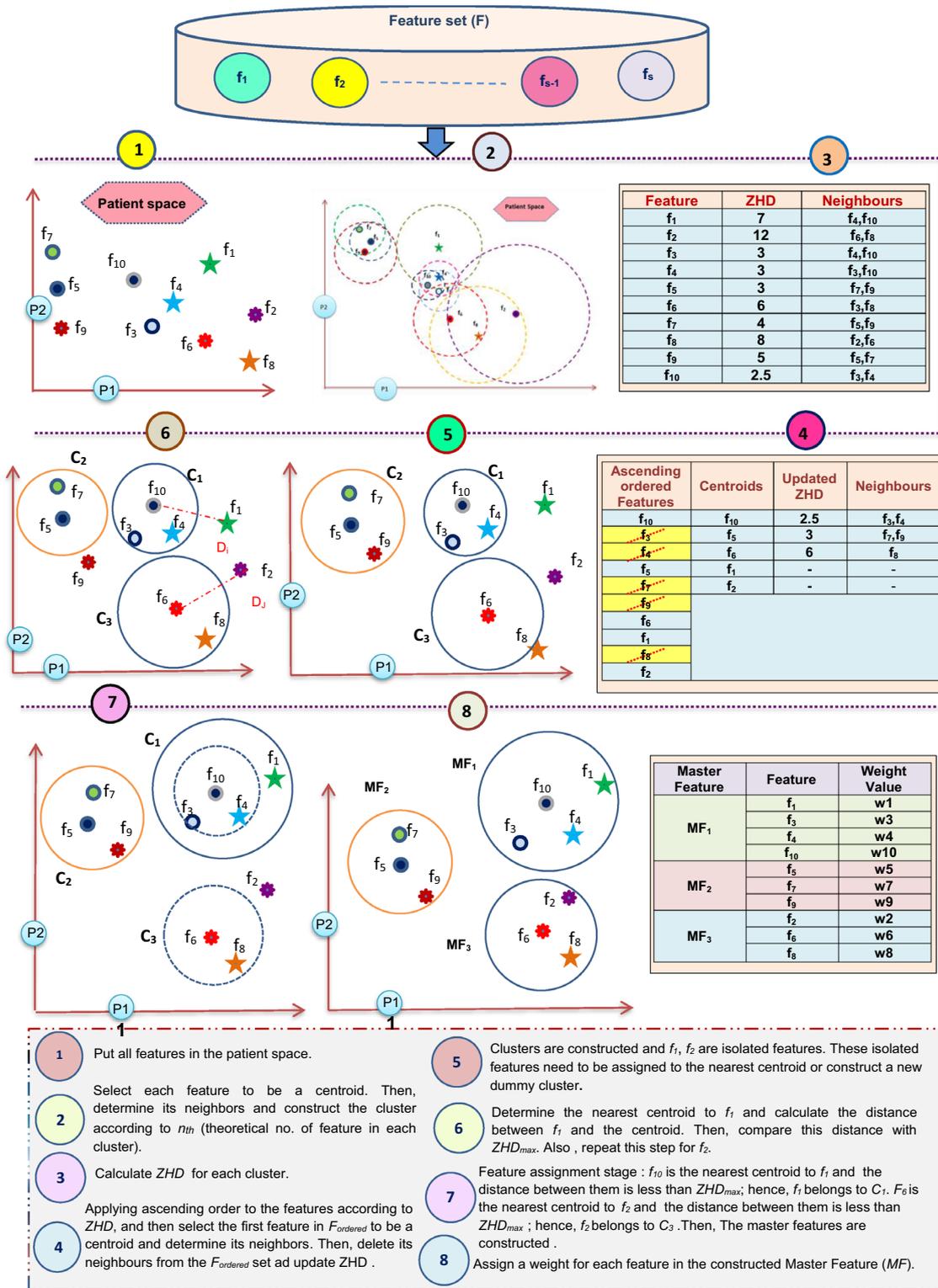
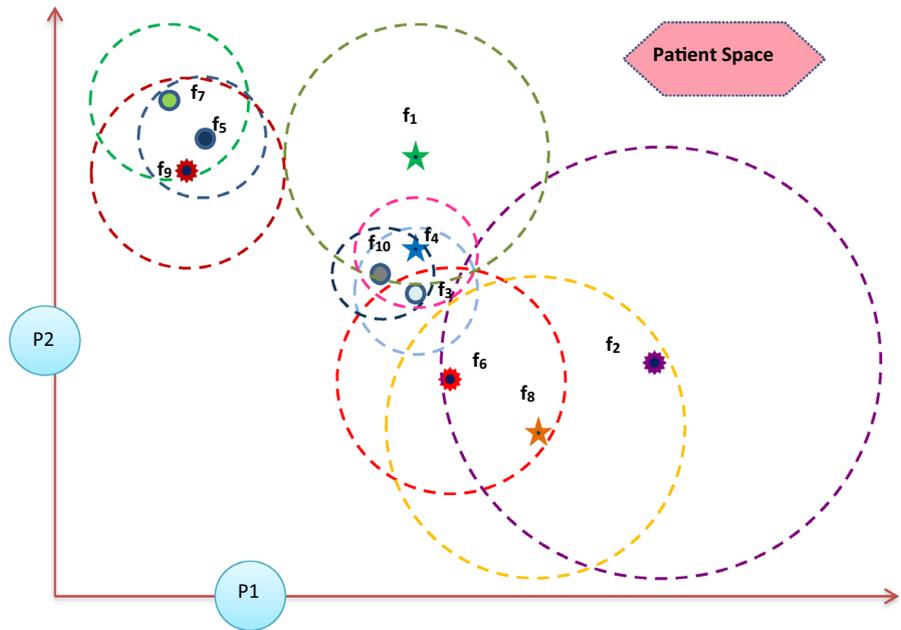


Fig. 7 The main steps for clustering the features to construct Master Features

Fig. 8 The initial construction of the clusters



and the largest radius of all clusters is called the maximum ZHD (ZHD_{max}). The theoretical number of features in each cluster is denoted by (n_{th}) while the actual number of features in each cluster is denoted by (n_{act}). The ascending order of features according to ZHD is denoted by ($F_{ordered}$).

To assign the isolated features, the distance between each feature f and the nearest centroid of all clusters (c_i) should be calculated by using Euclidean distance ($D(f, c_i)$) (Dokmanic et al. 2015). The steps of constructing the clusters of similar features are illustrated in algorithm 2. Accordingly,

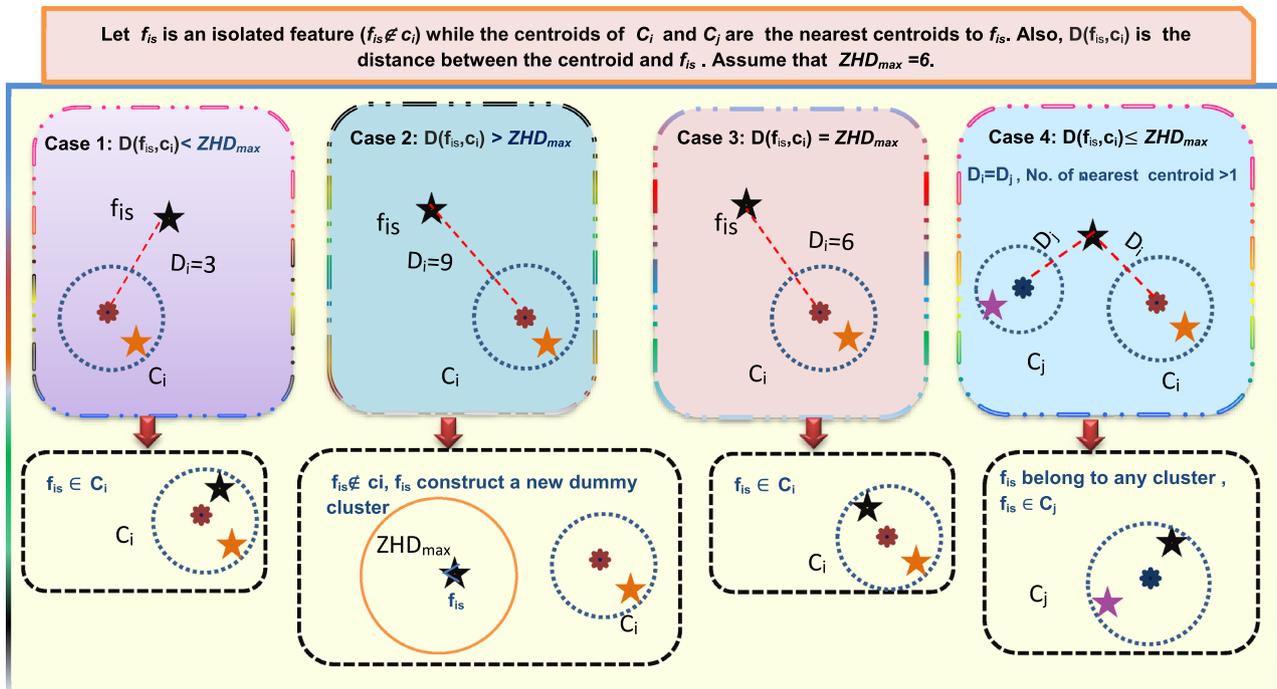
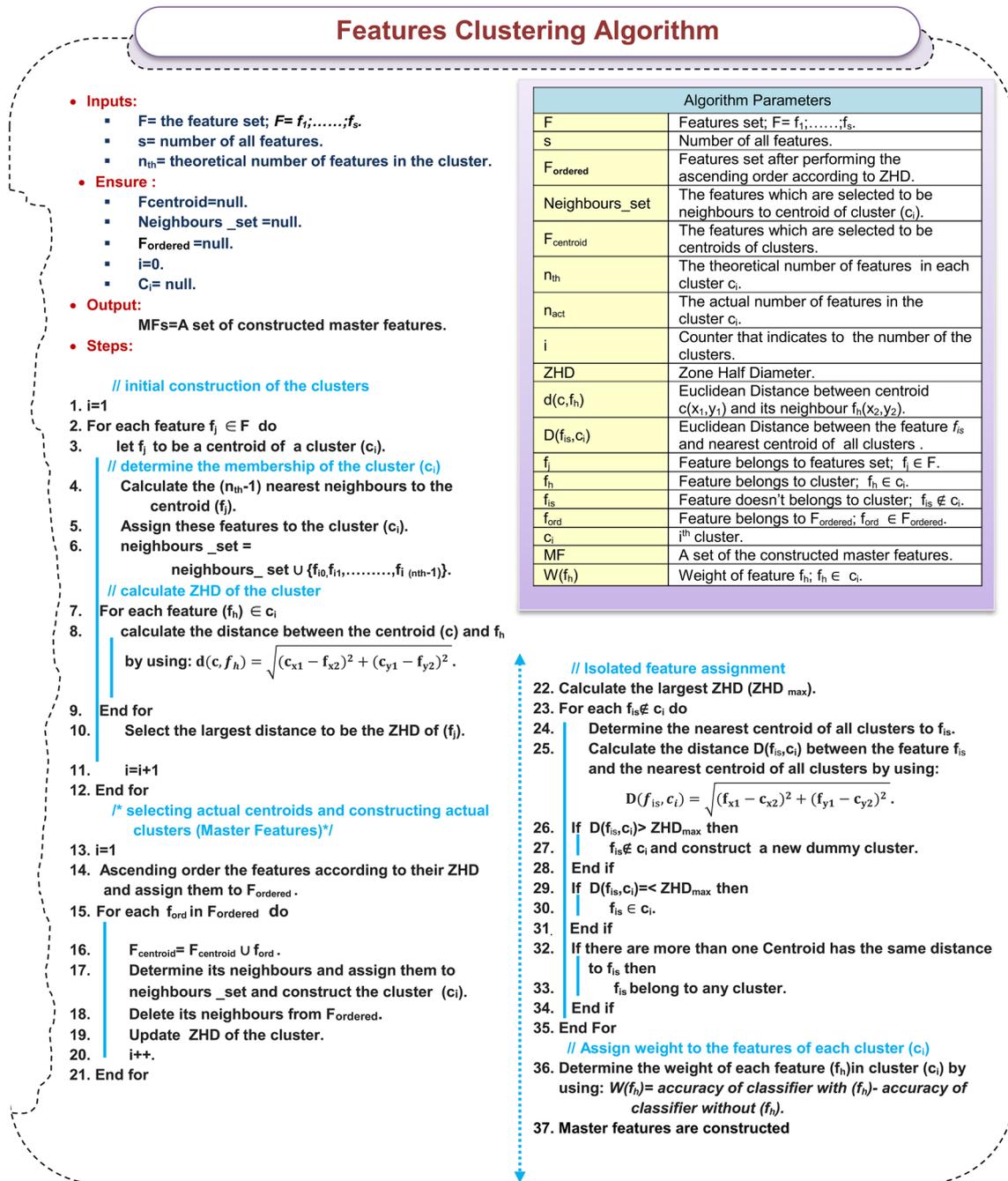


Fig. 9 The steps of assigning each isolated feature to its nearest cluster or to a new dummy cluster

the main steps of FCP to construct the clusters of similar features are presented in Fig. 7. In Fig. 7, each cluster is represented as a big circle, but each feature is represented as a small circle or star.

neighbours = 2), hence, $n_{th} = 3$. For simplicity, assume that the Euclidean distance $d(c, f)$ will be implemented between



6.1.2.1 Construct actual clusters According to the first step called construct actual clusters, each feature in the Features set is considered as a centroid of the cluster as illustrated in Fig. 8. Then, the distance between each centroid and their neighbours should be calculated by using Euclidean distance to determine their nearest neighbours (e.g. no. of nearest

the centroid c and one of its neighbour features f in 2-dimension; $c(x_1, y_1)$ and $f(x_2, y_2)$ by using (6) (Dokmanic et al. 2015; Liu et al. 2020).

$$d(c, f) = \sqrt{(c_{x1} - f_{x2})^2 + (c_{y1} - f_{y2})^2} \tag{6}$$

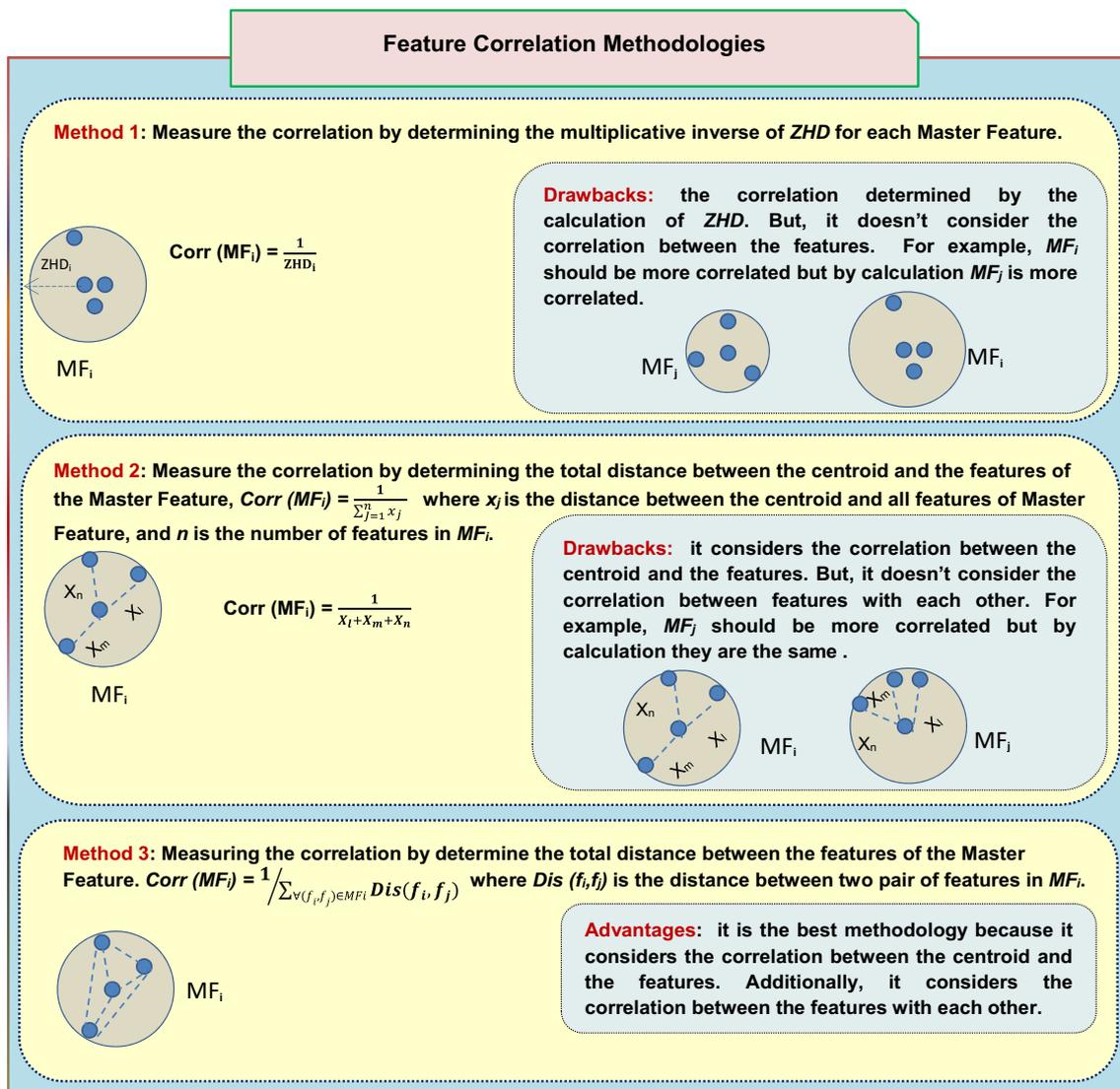


Fig. 10 The steps of implementing the proposed feature correlation methodology

where $d(c, f)$ is the distance between the centroid c and one of its neighbours f , x and y are the coordinates of both the centroid and the feature. Distance calculation should be performed between the centroid and all neighbors of features. According to the smallest distances, the centroid can determine their nearest neighbors (e.g. no. of nearest neighbors = 2). On the other hand, the ZHD can be determined for the cluster in which its value is the largest distance between the centroid and its neighbours as illustrated in the third step of Fig. 7. Then, actual clusters are constructed by placing the features in F_{ordered} in an ascending order based on their

ZHD. After that, the feature which has the smallest ZHD is selected to be a centroid of the first actual cluster.

The centroid neighbours should be determined and assigned to their corresponding cluster, and then these neighbours should be removed from F_{ordered} as illustrated in the fourth step of Fig. 7. The same steps will be repeated according to the current F_{ordered} until all actual clusters have been constructed. Although actual clusters that include similar features have been constructed, many isolated features do not belong to any actual cluster such as f_1 and f_2 as illustrated in the fifth step of Fig. 7. Thus, these isolated features need to be assigned to the nearest cluster or need to construct a

new dummy cluster that includes them. The next subsection describes how to assign the isolated features.

6.1.2.2 Isolated feature assignment After creating all actual clusters, there are many isolated features. An isolated feature is a feature that doesn't belong to any actual cluster. This feature needs to be assigned to the nearest cluster or needs to construct a new dummy cluster that includes it (Arunadevi et al. 2019). There are many steps to solve this problem as shown in Fig. 9. Figure 9 illustrates the assignment of the isolated features to the nearest cluster or a new dummy cluster. At first, the largest radius of all actual clusters (ZHD_{max}) should be determined. Then, the distance between each isolated feature and the nearest centroid should be calculated by using Euclidean distance and then compared to ZHD_{max} . If the distance is more than ZHD_{max} , this means that the isolated feature will construct a new dummy cluster. Otherwise, the isolated feature will belong to the cluster of the nearest centroid (C_i). If there is more than one centroid has the same distance to the isolated feature, the isolated feature will belong to any one of these nearest centroids. Finally, the Master Features (MFs) are constructed to represent the final clusters after assigning the isolated features to their corresponding clusters.

6.1.2.3 Weighting features After the construction of actual clusters has been performed and then the isolated features have been assigned to their corresponding cluster, the features should be weighted. Weighting features is a significant process in P²S because it can decrease the complexity, increase the performance of the machine-learning algorithm, and increase the resource efficiency of the used classifier. Usually, many classification algorithms suppose that all features have the same importance (same weights) or neglect the consistency of weights assigned to features. To solve this problem, it is important to calculate the feature weight value in which the largest weights should be assigned to the most effective features on COVID-19. Hence, different features can have different levels of importance in class prediction (Arunadevi et al. 2019). The last step in the FCP is to calculate the weight of each feature in the constructed Master Feature (MF_i). The weight of each feature f_h can be calculated by using (7).

$$W(f_h) = \text{Accuracy of classifier } (+f_h) - \text{Accuracy of classifier } (-f_h) \quad (7)$$

where $W(f_h)$ is the weight value of feature f_h , *Accuracy of classifier* $(+f_h)$ is the accuracy of the used classifier when the feature f_h is included in the feature set, and *Accuracy of classifier* $(-f_h)$ is the accuracy of the used classifier when f_h is eliminated.

6.1.3 Master feature weighting phase (MFWP)

MFWP is the third and final phase in the P²S stage that is used to assign a weight to each MF. Indeed, the correlation between features is very important before assigning weights to them. Hence, it is an essential process to determine the correlation between features by using a suitable correlation method. Correlation analysis is one of the well-known widely used techniques that identifies; (1) the relationship between the features and the predicted class and (2) the relationship between the features with each other. Mathematically, the relationship between features can be determined by a decimal value called the correlation coefficient. The positive sign of the coefficient indicates that the two features are positively correlated, the negative sign means negative correlation, and the '0' value means no correlation (Li et al. 2016). In this paper, the proposed feature correlation method based on the distance measurement has been introduced to calculate the relationship between the features of MF_i. The steps of implementing the proposed feature correlation methodology are illustrated in Fig. 10.

Figure 10 shows three proposed feature correlation methods to measure the correlation between the features of the MF_i. The first method measures the correlation by determining the multiplicative inverse of ZHD, but it does not take into consideration the correlation between features. For example, as shown in Fig. 10, if "ZHD of MF_i = 10" and "ZHD of MF_j = 5" then, the correlation of MF_i and MF_j are 0.1 and 0.2 respectively. This means that MF_j is more correlated than MF_i, but it is not correct. In the second method, the correlation is measured by determining the total distance between the centroid and all features of the master feature. The limitation of this method is that it does not take into consideration the correlation between the features with each other. In the third and final method, the correlation is measured by determining the total distance between the features of the master feature. This method considers the correlation between the features and the centroid, and also the correlation between the features with each other. Accordingly, the third method is the best correlation method used to measure the correlation between features. After implementing the third correlation method to calculate the relationship between the features of the MF, MFWP can be implemented to calculate the weights of MFs.

NB classifier is a common that is used in machine learning and data mining. It is crucial to use NB for solving different data classification problems because it is simple to be trained, easy to implement, and can provide fast and accurate predictions. However, it assumes that all features are conditionally independent which is often harming the performance of classification. This is not correct

in real-world applications because the features don't have the same importance (Taheri et al. 2014). To improve the performance of NB, many modified methods based on NB have been proposed. One of these modified methods is to assign a weights value to each feature. In this paper, the weight of each master feature (MF_i) depends on three parameters, which are; the number of features in this master feature, the correlation value between features in MF_i , and the summation of weights values for each feature in MF_i . The weight of master feature MF_i can be calculated by using (8).

$$W(MF_i) = N_i \times corr(MF_i) \times \sum_{\forall f_j \in MF_i} w_i \tag{8}$$

where $W(MF_i)$ is a weight of master feature MF_i , N_i is the number of features in the master feature MF_i , and $corr(MF_i)$ is the correlation value between features in the master feature MF_i . w_i is the weight value of each feature f_j that belongs to the master feature (MF_i). After calculating the weights of all MFs, the weighted MFs will be used in the next stage (CS) to implement the weighted NB in FCNBP. In the next section, FCNBP will be explained in detail to classify the COVID-19 patients by implementing the weighted NB classifier on the weighted MFs.

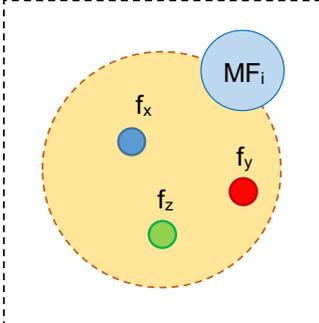
6.2 Feature Correlated Naïve Bayes phase (FCNBP)

NB is known to be an effective, robust, and efficient classification algorithm. NB is a promising solution as it only requires a little amount of training data to estimate the parameters required for classification and able to accommodate new incoming data for training both efficiently and incrementally. Although NB had received extensive attention due to its excellent classification performance and simplicity, it sometimes has a degraded performance due to the naïve assumption that features are independent and equally weighted. To compensate the performance of the traditional NB, a new classifier is proposed in this

phase, which is called Feature Correlated Naïve Bayes (FCNB). The proposed FCNB enhances the performance of the traditional NB by clustering the selected features into groups called master features (MFs) in which each MF includes a set of dependent or related features. Moreover, each MF is weighted based on the importance of the features it includes as well as the correlation among the included features. FCNB operates just like weighted NB; however, it replaces the employed features with a set of constructed MFs. Also, it considers the weights of the used MFs rather than the weights of the individual features. This has a positive effect in (1) promoting the performance of the traditional weighted NB as it considers the correlation among features and (2) minimizes the classification time as it considers a smaller number of MFs rather than many individual features.

To explain how FCNB operates, consider a diagnosis database that includes 'Ca' cases in which 'A' cases are infected with COVID-19 and 'B' cases are not, hence; $Ca = A + B$. Consider 's' selected features labeled as; f_1, f_2, \dots, f_s , which are clustered into mm master features labeled as MF_1, MF_2, MF_3, \dots and MF_{mm} , where $s > mm$. Like any supervised learning-based classifier, FCNB operates in two sequential phases; namely training and testing. The training of the proposed FCNB is accomplished by constructing a Conditional Probability Table (CPT) for each master feature MF_i as illustrated in Table 8 based on the input diagnosis database. As depicted in Table 8, for simplicity, it is assumed that MF_i includes three dependent features, namely; f_x, f_y , and f_z in which each feature takes 'L' or 'H' value, which corresponds to "Low" or "High" respectively. Accordingly, MF_i has 8 distinct values labeled as; $X_{ij} \forall j \in \{1, 2, 3, \dots, 8\}$. However, each MF can include more dependent features in which each feature can take one of many values rather than 'L' and 'H' only. For illustration, a feature may take a value $\forall \in \{VL, L, M, H, VH\}$, which indicates "Very Low", "Low", "Medium", "High", or "Very High" respectively. Table 8 illustrates CPT of MF_i in which the conditional probability for each

Table 8 Conditional probabilities for master feature MF_i



MF _i Values	Description			Classes		P(f ₁ T)	P(f ₁ F)
	f _x	f _y	f _z	T	F		
X ₁₁	L	L	L	V _{T1}	V _{F1}	V _{T1} /S _T	V _{F1} /S _F
X ₁₂	L	L	H	V _{T2}	V _{F2}	V _{T2} /S _T	V _{F2} /S _F
X ₁₃	L	H	L	V _{T3}	V _{F3}	V _{T3} /S _T	V _{F3} /S _F
X ₁₄	L	H	H	V _{T4}	V _{F4}	V _{T4} /S _T	V _{F4} /S _F
X ₁₅	H	L	L	V _{T5}	V _{F5}	V _{T5} /S _T	V _{F5} /S _F
X ₁₆	H	L	H	V _{T6}	V _{F6}	V _{T6} /S _T	V _{F6} /S _F
X ₁₇	H	H	L	V _{T7}	V _{F7}	V _{T7} /S _T	V _{F7} /S _F
X ₁₈	H	H	H	V _{T8}	V _{F8}	V _{T8} /S _T	V _{F8} /S _F
Total				S _T	S _F	100%	100%

value $X_{ij} \forall j \in \{1, 2, 3, \dots, 8\}$ of MF_i for each target class (e.g., T or F) is calculated given the input diagnose database. It is assumed that the weight of MF_i is W_i while the prior probabilities of the considered target classes are; $P(T)=A/Ca$ and $P(F)=B/Ca$.

On the other hand, the task during the testing phase of FCNB is to diagnose the input case to indicate whether the case is infected with COVID-19 or not. Assuming an input case IC who has the following feature vector $\langle f_1, f_2, f_3, \dots, f_{s-1}, f_s \rangle$ with the corresponding values $\langle L, L, H, \dots, h, L \rangle$. Initially, the input features are clustered into the corresponding master features (e.g., $MF_1, MF_2, \dots, MF_{mm}$) with the corresponding values. Considering the CPT of each employed MF , it will be easy to find the conditional probability for each value of the employed master features. Hence, it will be easy to diagnose the new case by estimating the posterior probability that the case is belonging to each class (T or F) as shown in (9) (Yearwood et al. 2014).

$$P(c_i|IC) \propto \left[P(c_i) \times \prod_{j=1}^n P(MF_j|c_i)^{W_j} \right] \tag{9}$$

where $P(c_i|IC)$ is the posterior probability that the case IC belongs to class c_i , $P(c_i)$ is the prior probability of class c_i , $P(MF_j|c_i)$ is the conditional probability of the master feature MF_j given the target class c_i , and W_j is the weight of MF_j . Considering two target classes (e.g., T and F), this yields (10) and (11) (Lee et al. 2011).

$$P(T|IC) \propto \left[P(T) * \prod_{j=1}^n P(MF_j|T)^{W_j} \right] \tag{10}$$

$$P(F|IC) \propto \left[P(F) * \prod_{j=1}^n P(MF_j|F)^{W_j} \right] \tag{11}$$

Finally, the target class for the input case IC can be calculated by using (12) (Ji et al. 2019).

$$Target(IC) = \arg \max_{c_i \in \{T,F\}} \left[P(c_i) \times \prod_{j=1}^n P(MF_j|c_i)^{W_j} \right] \tag{12}$$

where $W_j \in \mathbb{R}^+$

7 Illustrative example

In this section, an illustrative example showing how the diagnosis decision can be taken in the Classification Stage (CS) of the proposed Feature Correlated Naïve Bayes (FCNB) classification strategy. As illustrated in Table 9, consider a COVID-19 diagnosis database for 100 persons in which 40 persons are infected by COVID-19 while the other 60 persons are not. For simplicity, Considering 8 selected features labeled f_1, f_2, \dots, f_8 , which are clustered into three master features labeled MF_1, MF_2 , and MF_3 , as well as two target classes, namely; “True” and “False” diagnose. The symbols ‘L’, ‘M’, and ‘H’ represents “low”, “medium”, and “high” respectively, while ‘T’ and ‘F’ represents “true” or “false” diagnose of the COVID-19 virus. The weight of each master feature is also reported in the last row of Table 9. On the other hand, the conditional probability for each feature value given different classes as well as the prior probability for each class are illustrated in Tables (10, 11, 12).

Now, it is required to diagnose a new case IC who has the following feature vector $\langle f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8 \rangle$ with the corresponding values $\langle L, L, H, H, L, L, h, L \rangle$. Initially, the input features are clustered into the corresponding master features (e.g., MF_1, MF_2, MF_3) with the corresponding values. Considering the input values of the selected features, it is found that; $MF_1 = X_{1,2}, MF_2 = X_{2,5}, MF_3 = X_{3,3}$. From Tables 10, 11, 12, it will be easy to find the conditional probability for each value of the employed master features, which are; $P(X_{1,2}|T) = 0.145, P(X_{1,2}|F) = 0.098, P(X_{2,5}|T) = 0.077, P(X_{2,5}|F) = 0.143, P(X_{3,3}|T) = 0.212$, and $P(X_{3,3}|F) = 0.194$.

Table 9 Medical database used for the illustrative example

Case number	MF ₁			MF ₂			MF ₃		Diagnose
	f ₁	f ₂	f ₃	f ₄	f ₅	f ₆	f ₇	f ₈	
1	L	L	H	L	L	H	M	L	T
2	H	H	H	H	H	H	M	L	F
3	M	L	L	M	L	L	L	H	T
↓ ↓ ↓ ↓ ↓									
99	M	H	L	M	H	L	H	L	T
100	H	L	L	H	L	L	L	H	T
Master Feature Weight	0.42			0.38			0.59		

On the other hand, the weights of the employed master features are illustrated at the bottom of Table 9, numerically; 0.42, 0.38, and 0.59 for MF_1 , MF_2 , and MF_3 respectively. Since the employed database has 40 infected persons with COVID-19, while the remaining persons are not, the prior probability for the target classes (e.g., T and F) are; 0.4 and 0.6 respectively. Hence, it will be easy to diagnose the new

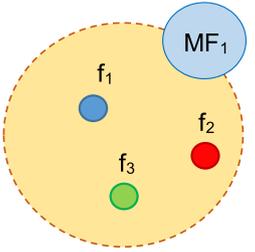
case by estimating the posterior probability that the case is belonging to each class (T or F) as shown below.

$$QT = \left[P(T) \times \prod_{j=1}^n P(MF_j|T)^{W_j} \right] = 0.4 \times P(X_{1,2}|T)^{W_1} \times P(X_{2,5}|T)^{W_2} \times P(X_{3,3}|T)^{W_3}$$

$$QT = 0.4 * (0.145)^{0.42} \times (0.077)^{0.38} \times (0.212)^{0.59} = 0.02687$$

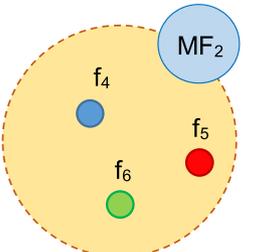
$$QF = \left[P(F) * \prod_{j=1}^n P(MF_j|F)^{W_j} \right] = 0.6 \times P(X_{1,2}|F)^{W_1} \times P(X_{2,5}|F)^{W_2} \times P(X_{3,3}|F)^{W_3}$$

Table 10 Conditional probabilities for master feature MF_1



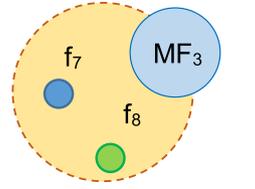
MF ₁ Values	Description			Classes		P(MF ₁ T)	P(MF ₁ F)
	f ₁	f ₂	f ₃	T	F		
X _{1,1}	L	L	L	13	2	≈ 0.188	≈ 0.065
X _{1,2}	L	L	H	10	3	≈ 0.145	≈ 0.098
X _{1,3}	L	H	L	11	2	≈ 0.159	≈ 0.065
X _{1,4}	L	H	H	6	5	≈ 0.087	≈ 0.161
X _{1,5}	H	L	L	9	3	≈ 0.130	≈ 0.098
X _{1,6}	H	L	H	7	5	≈ 0.101	≈ 0.161
X _{1,7}	H	H	L	8	4	≈ 0.116	≈ 0.129
X _{1,8}	H	H	H	5	7	≈ 0.072	≈ 0.226
Total				69	31	100%	100%

Table 11 Conditional probabilities for master feature MF_2



MF ₂ Values	Description			Classes		P(MF ₂ T)	P(MF ₂ F)
	f ₄	f ₅	f ₆	T	F		
X _{2,1}	L	L	L	2	9	≈ 0.031	≈ 0.257
X _{2,2}	L	L	H	4	7	≈ 0.065	≈ 0.200
X _{2,3}	L	H	L	6	5	≈ 0.096	≈ 0.143
X _{2,4}	L	H	H	10	2	≈ 0.154	≈ 0.057
X _{2,5}	H	L	L	5	5	≈ 0.077	≈ 0.143
X _{2,6}	H	L	H	9	2	≈ 0.138	≈ 0.057
X _{2,7}	H	H	L	12	3	≈ 0.185	≈ 0.086
X _{2,8}	H	H	H	17	2	≈ 0.261	≈ 0.057
Total				65	35	100%	100%

Table 12 Conditional probabilities for master feature MF_3



MF ₃ Values	Description		Classes		P(MF ₃ T)	P(MF ₃ F)
	f ₇	f ₈	T	F		
X _{3,1}	L	L	14	7	≈ 0.424	≈ 0.104
X _{3,2}	L	H	9	11	≈ 0.272	≈ 0.164
X _{3,3}	H	L	7	13	≈ 0.212	≈ 0.194
X _{3,4}	H	H	3	36	≈ 0.100	≈ 0.537
Total			33	67	100%	100%

Table 13 The applied parameters with the corresponding used values

Parameter	Description	Applied value
P_{sel}	Probability of selection	Random ($0 \leq P_{sel} \leq 1$)
P_{cross}	Probability of Crossover	Random ($0 \leq P_{cross} \leq 1$)
P_{mut}	Probability of Mutation	Random ($0 \leq P_{mut} \leq 1$)
n_{th}	Theoretical number of feature in each cluster	3

$$QF = 0.6 \times (0.0.98)^{0.42} \times (0.143)^{0.38} \times (0.194)^{0.59} = 0.04104$$

where QT indicates the degree of confidence that *IC* is infected with COVID-19 and QF indicates the degree of confidence that *IC* is not infected with COVID-19. Hence, since $QT < QF$, then the input case *IC* is not infected with COVID-19.

8 Experimental result

In this section, the evaluation of the proposed FCNB classification strategy is investigated. As mentioned in Sect. 6. In fact, FCNB consists of two main stages, which are; P²S and CS. The P²S stage is composed of the first three phases of the FCNB strategy called FSP, FCP, and MFWP while the CS stage contains FCNBP that represents the last phase of the FCNB strategy. To this end, the experimental results have many ordered steps. Firstly, the historically collected data on both COVID-19 patients and non-COVID-19 people will be sent to FSP for selecting the meaningful features by using FSGA. Secondly, in FCP, the selected features will be grouped into clusters according to their correlation. Then, MFWP will assign a weight value to each MF that includes a set of dependent or related features by using a new weight calculation method. Finally, the output of P²S will be passed to FCNBP in CS for providing a fast and accurate diagnosis of COVID-19 patients by using the weighted NB classifier.

In this vein, there are two main scenarios are followed to implement the proposed FCNB classification strategy. In the first scenario, FSGA is applied to select informative features from the COVID-19 dataset comparing to other recent state-of-the-art feature selection methods. The main aim of the first scenario is to illustrate the effectiveness of FSGA against other methods. During the second scenario, the whole FCNB classification strategy is implemented to accurately diagnose COVID-19 patients. Our implementation is based on COVID-19 dataset (Ferrari et al. 2020; Brinati et al.2020).The dataset is divided into two sets called; training and testing. The model can be learned by using the training data and then the performance of the model can be measured by using the testing data. Many tunable parameters have been used in FSGA and FCP in

which these parameters with the corresponding implemented values are described in Table 13.

8.1 Description of COVID-19 dataset

COVID-19 dataset is a real dataset that is used to detect COVID-19 patients. This real dataset contains results of routine blood tests collected from different cases who were admitted to San Raffaele Hospital (Milan, Italy) (Ferrari et al. 2020; Brinati et al. 2020). Additionally, this dataset contains personal information of cases like age and gender (Male or Female). The total number of cases in this real dataset is 207. The dataset is divided into training and testing sets where the number of cases in training data is 140 and the number of cases in testing data is 67. According to this real dataset, it is considered two class categories called; COVID patients and Un-COVID people as shown in Table 14. The distribution of the used cases in the collected dataset has been represented according to “Age”, “Gender” as shown in Figs. 11, 12, 13.

8.2 Evaluation metrics

During the next experiments, the evaluation parameters such as accuracy, error, recall, and precision will be calculated. Then, F-measure, micro average and macro average related to precision and recall will be measured. The confusion matrix is used to calculate the values of these parameters. A confusion matrix is applied as presented in Table 15. Various formulas are used as a summarization of the confusion matrix as depicted in Table 16. Finally, the speed of COVID-19 detection algorithms should be measured by using the second unit.

8.3 Testing the proposed feature selection technique

The effectiveness of the proposed feature selection method called FSGA is evaluated and compared with other existing approaches, which are; FSJaya (Das et al. 2020), MGOA (Sehgal et al. 2020), SDS (Shanthi and Rajkumar 2020), and ACO (Sowmiya and Sumitra 2020) by using the considered COVID-19 dataset. These feature selection approaches are described in Table 17. To prove the effectiveness of the feature selection method, the NB classifier is applied as a standard classifier (Rabie et al. 2019a, b, 2020; Ayyad et al. 2019). The obtained results show that the FSGA outperforms other approaches which are presented in Table 17. The results are shown in Figs. 14, 15, 16, 17, 18, 19, 20, 21, 22, 23.

As illustrated in Figs. 14, 15, 16, 17, FSJaya, SDS, ACO, MGOA, and FSGA techniques introduced accuracy with 0.82, 0.84, 0.88, 0.87, and 0.98 respectively with 140 training patients. The best accuracy is achieved by FSGA because

Table 14 Dataset description

Criteria	Value/description						
Total number of cases	Male					Female	
	127					80	
Un-COVID-19 cases	Male					Female	Total
	53					49	102
COVID-19 cases	Male					Female	Total
	74					31	105
	Male	< 18	<18	< 18	45–55	55–65	> 65
		2	2	2	20	18	30
	Female	0	0	0	6	3	15

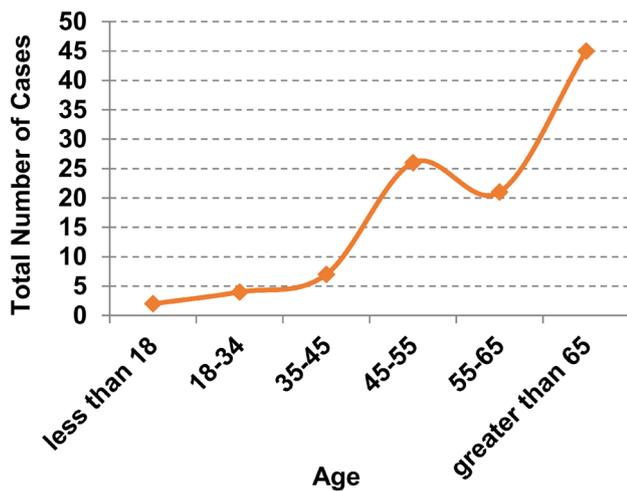


Fig. 11 The total number of cases according to age

FSGA can accurately select the most effective features of COVID-19 diagnosis. Thus, FSJaya, SDS, ACO, MGOA, and FSGA techniques give error values of 0.21, 0.19, 0.17, 0.13, and 0.01 respectively. FSGA gives a precision value that equals 0.89 while FSJaya, SDS, ACO, and MGOA give 0.65, 0.66, 0.63, and 0.68 respectively. While the recall value of FSGA is 0.84, the recall values of FSJaya, SDS, ACO, and MGOA are 0.6, 0.62, 0.63, and 0.67 respectively. Consequently, Figs. 14, 15, 16, 17 show that FSGA is better than other recent methods, which are; FSJaya, MGOA, SDS, and ACO because FSGA provides the maximum accuracy and the minimum error.

The results in Figs. 18, 19, 20, 21, 22 show that the highest macro-average precision value is provided by FSGA with a value that reaches 0.8 at a training number of 140 patients. On the other hand, the lowest macro-average precision value is introduced by FSJaya with a value reaches to 0.63. Additionally, macro-average recall for FSGA is about 0.84 which represents the highest value concerning other techniques, while the lowest one

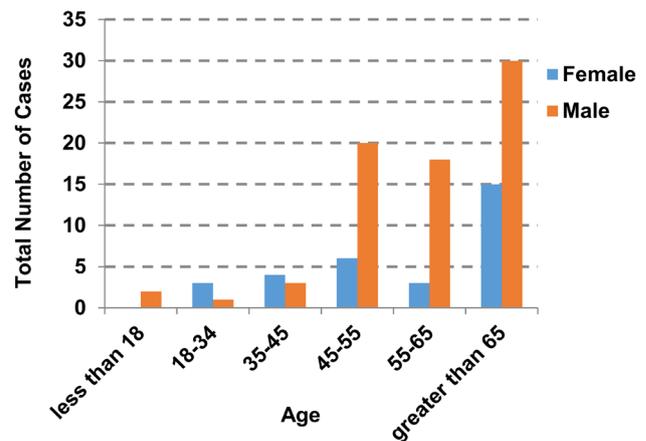


Fig. 12 The total number of COVID-19 cases according to age and gender

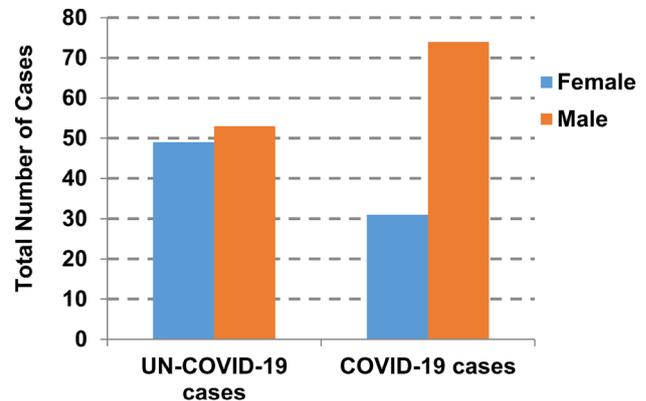


Fig. 13 The presentation of COVID-19 cases and un COVID-19 cases distribution

is FSJaya with a value of 0.61 at a training number of 140 patients. At the maximum number of training data (e.g., 140 patients), FSGA gives the highest micro-average precision value equals 0.8 while SDS introduced 0.61

which is the lowest value of micro-average precision. At a training number of 140 patients, FSGA provides a micro average recall value that equals 0.79, while FSJaya, SDS, ACO, and MGOA provide 0.59, 0.6, 0.66, and 0.67 respectively. F-measure value for FSGA is about 0.78 while it is about 0.60, 0.64, 0.65, and 0.65 for FSJaya, SDS, ACO, and MGOA respectively. In Fig. 23, the run-time of FSGA is 10 (s) that represents the highest speed while SDS introduces the lowest speed with a run-time value equals to 20 (s). In the end, FSGA outperforms other recent methods, which are; FSJaya, MGOA, SDS, and ACO because it can accurately select the most informative features with high speed.

8.4 Testing feature correlated Naïve Bayes (FCNB) classification strategy

In this section, the proposed FCNB strategy that includes four phases, which are; feature selection, feature clustering, master feature weighting, and classification phases will be evaluated. To ensure the effectiveness of the FCNB strategy, it is compared to some of the recently used COVID-19 classification strategies as presented in Table 1. Those methods are TCRC (Khanday et al. 2020), DL (Ozturk et al. 2020a), CNN (Maghdid et al. 2020), CDM (Chen et al. 2020a, b),

COVIDGAN (Waheed et al. 2020), ACDM (Ozturk et al. 2020b) and AFS-DF (Sun et al. 2020). In fact, the proposed FCNB classification strategy depends on many essential techniques which enable the classification model to provide fast and accurate classifications. These essential techniques are FSGA that is employed for selecting the best subset of features in FSP, the proposed clustering method in FCP that is applied on the selected features to put them in clusters, the proposed weighting method in MFWP that is used to weight the master feature, and the weighted NB classifier that is applied on the weighted master features in FCNBP to accurately detect COVID-19 patients. Results are shown in Figs. 24, 25, 26, 27, 28, 29, 30, 31, 32, 33. As shown Figs. 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, the proposed FCNB strategy demonstrates the best performance while its error and run-time are decreased. Really, the accuracy, precision, recall, macro-average, micro-average, and F-measure of FCNB are promoted. This proves the effectiveness of FCNB in which its phases are FSP, FCP, MFWP, and FCNBP, can efficiently work together.

As shown in Figs. 24, 25, 26, 27, TCRC, DL, CNN, CDM, COVIDGAN, ACDM, AFS-DF, and FCNB provide accuracy values reach to 0.8, 0.81, 0.87, 0.85, 0.90, 0.92, 0.96, and 0.99 respectively at the maximum number of training data (e.g., 140 patients). The best accuracy is achieved by FCNB depends on the usage of many main processes in the pre-processing stage before applying the classification model to accurately diagnose COVID-19 patients. Accordingly, TCRC, DL, CNN, CDM, COVIDGAN, ACDM, AFS-DF, and FCNB techniques introduce error values equal 0.2, 0.19, 0.13, 0.15, 0.10, 0.8, 0.4, and 0.02 respectively. FCNB gives precision value that equals 0.84 while TCRC, DL, CNN, CDM, COVIDGAN, ACDM, and AFS-DF give 0.63, 0.65, 0.63, 0.68, 0.67, 0.69, 0.73, 0.69, and 0.73 respectively.

Table 15 Confusion matrix which depicts how classification on cases

	Predicted true class	Predicted negative class
Actual true class	True positive (TP)	False negative (FN) Type 2 error
Actual False class	False positive (FP) Type 1 error	True negative (TN)

Table 16 Confusion matrix formulas

Measure	Formula	Definition
Precision (PR)	$TP/(TP + FP)$	Metric indicates the number of the correct positive class
Recall/sensitivity (RE)	$TP/(TP + FN)$	Metric indicates the number of correct positive class made out of all positive class classification
Accuracy (AC)	$(TP + TN)/(TP + TN + FP + FN)$	The classifier’s ability to classify the class label correctly
Error (ER)	1-Accuracy	The percentage of classification which is inaccurate
Macro-average	$\sum_{i=1}^n P_i/n$ “for Precision” $\sum_{i=1}^n R_i/n$ “for Recall”	The average of the precision and recall of the system on various n classes. It is used for knowing how the system acts overall across the sets of data
Micro-average	$(TP1 + TP2)/(TP1 + TP2 + FP1 + FP2)$ “for precision” $(TP1 + TP2)/(TP1 + TP2 + FN1 + FN2)$ “for Recall”	Summation of the individual true positives, false positives, and false negatives of the system for various sets and then applies them to get the statistics. It can be a useful measure when the size of dataset varies
F-measure	$2 \times \frac{P+R}{P+R}$	Metric for merge both recall and precision into a single score that take both properties

Table 17 The most recent features selection techniques used for evaluation

Feature selection techniques	Description
Feature Selection based on Jaya optimization algorithm (FSJaya) (Das et al. 2020)	Das et al. (2020) proposed a new wrapper feature selection technique based on Jaya optimization algorithm (FSJaya) to select the most important subset of features by updating the worst features. The main theoretical implications of this technique are to decrease the computational time, improve the performance of the model and solve the over fitting problem. The effectiveness of FSJaya was evaluated by using four classifiers, which are; KNN, NB, LDA, and RT. The results presented shown that the proposed FSJaya technique could effectively select an optimal subset of features comparing to other methods
The Modified Grasshopper Optimization Algorithm(MGOA) (Sehgal et al. 2020)	Sehgal et al. (2020) proposed MGOA aims to improve the traditional GOA algorithm by selecting the best subset of features that effect on Parkinson disease and minimize the computational time. In this paper three machine learning algorithms have been used to classify the datasets which are; KNN, Random Forest and decision tree. The experimental results shown that the Random Forest applied in the proposed algorithm overcomes other machine learning algorithms by achieving accuracy of 95.37% with detection rate of 99.47% and false alarm rate of 15.78%
Stochastic Diffusion Search (SDS) algorithm (Shanthi and Rajkumar 2020)	Shanthi and Rajkumar (2020) introduced a new wrapper feature selection method based on SDS algorithm. The GLCM with the method of Gabor filter feature have been used to extract the radiomic features. Then, the most significant features were selected to differentiate between different classes in efficient and accurate manner. In order to accomplish the classification task, three classifiers have been used which are; Neural network, Naïve Bayes, and decision tree. An evaluation of the performance of the proposed method proven that the model reaches to effective results and able to achieve better levels of performance compared with the another methods
Ant Colony Optimization (ACO) algorithm (Sowmiya and Sumitra 2020)	Sowmiya and Sumitra (2020) introduced an enhanced hyper approach with new feature selection for providing accurate predictions. In the first step, the Cleveland dataset is pre-processed. Then, ant colony algorithm was used to choose the most necessary features in dataset to improve the prediction performance. The hybrid KNN (HKNN) used the selected features for the classification task

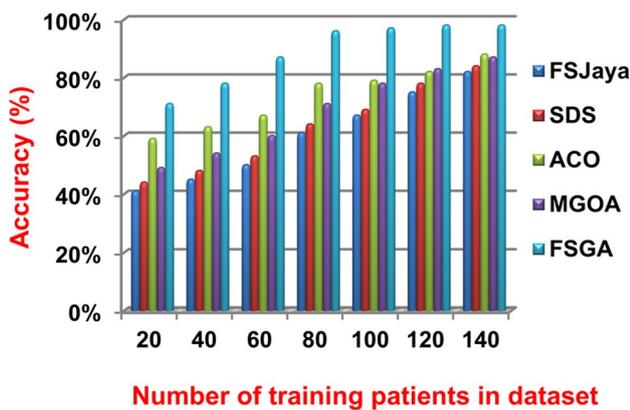


Fig. 14 Accuracy of different feature selection techniques

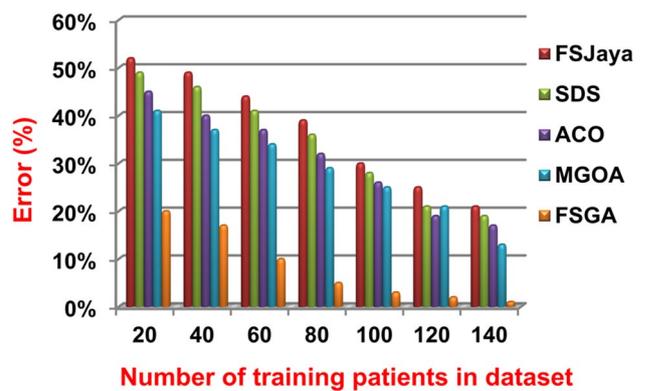


Fig. 15 Error of different feature selection techniques

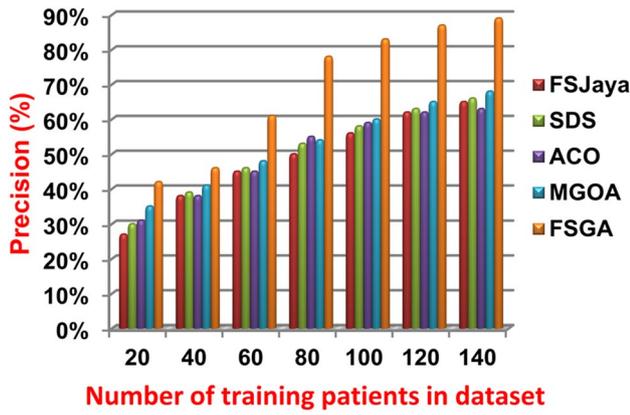


Fig. 16 Precision of different feature selection techniques

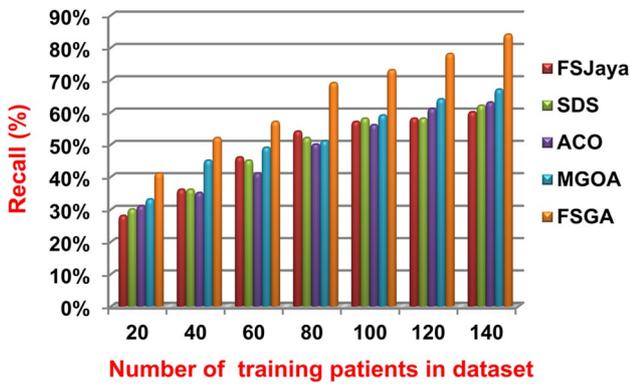


Fig. 17 Recall of different feature selection techniques

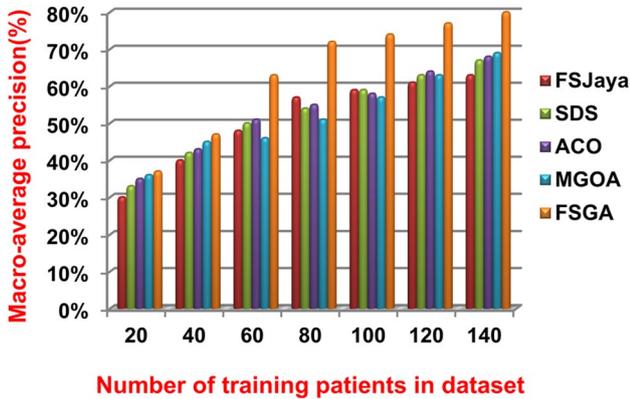


Fig. 18 Macro-average precision of different feature selection techniques

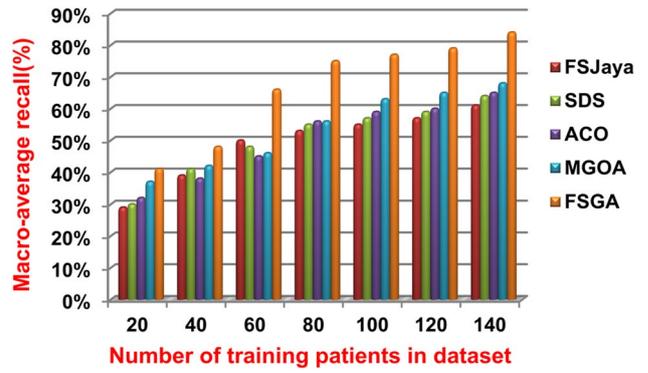


Fig. 19 Macro-average recall of different feature selection techniques

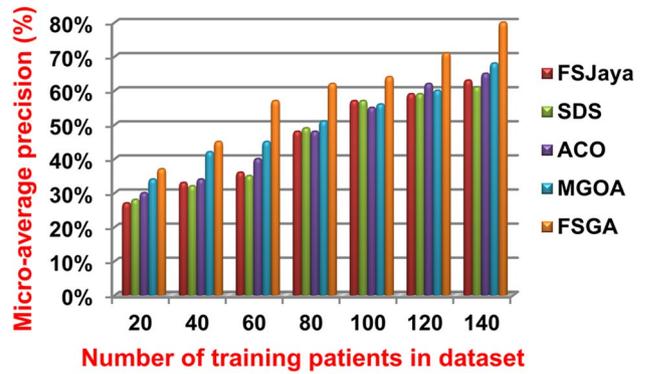


Fig. 20 Micro-average precision of different feature selection techniques

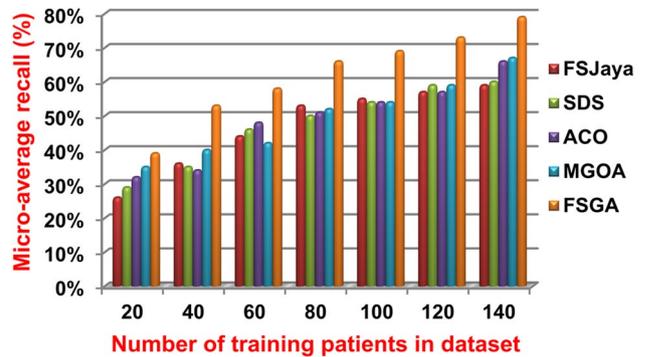


Fig. 21 Micro-average recall of different feature selection techniques

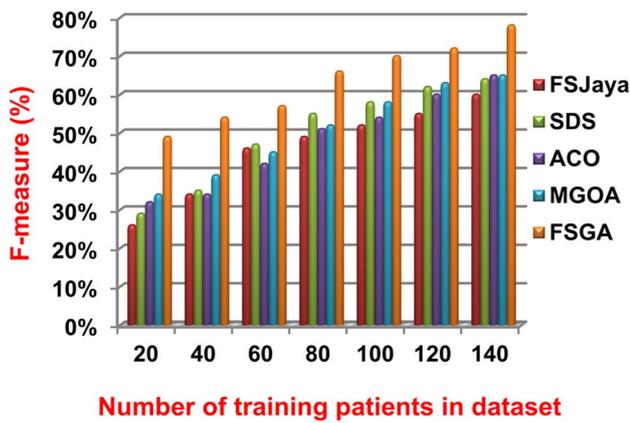


Fig. 22 F-measure of different feature selection techniques

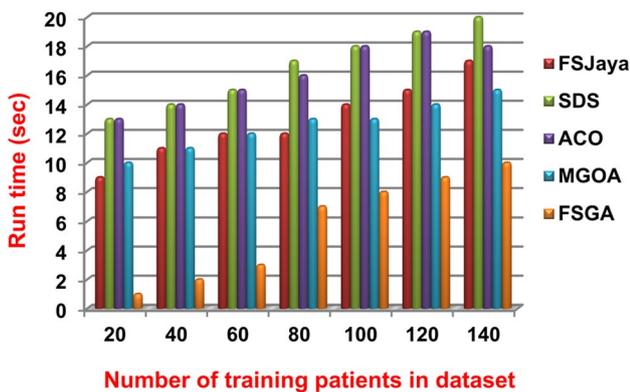


Fig. 23 Run time of different feature selection techniques

While the recall value of FCNB is 0.79, the recall values of TCRC, DL, CNN, CDM, COVIDGAN, ACDM, and AFS-DF are 0.6, 0.62, 0.63, 0.67, 0.66, 0.68, and 0.75 respectively. Hence, Figs. 24, 25, 26, 27 show that FCNB is better than other recent methods, which are; TCRC, DL, CNN, CDM, COVIDGAN, ACDM, and AFS-DF because FCNB introduces the maximum accuracy and the minimum error.

The results in Figs. 28, 29, 30, 31, 32 show that the highest macro-average precision value is provided by FCNB with value reaches to 0.78 at training number of 140 patients. On the other hand, the lowest macro-average precision value is introduced by TCRC with value reaches to 0.61 at the same training number of patients. Additionally, the macro-average recall for FCNB is about 0.77 which represents the highest value concerning techniques, while the lowest one is TCRC with a value of 0.60 at a training number of 140 patients.

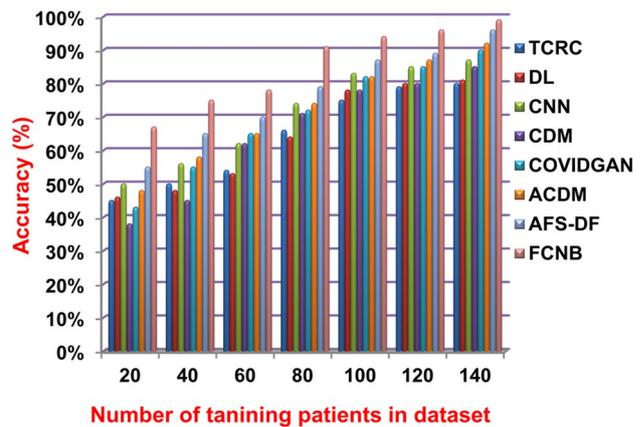


Fig. 24 Accuracy of different COVID-19 classification techniques

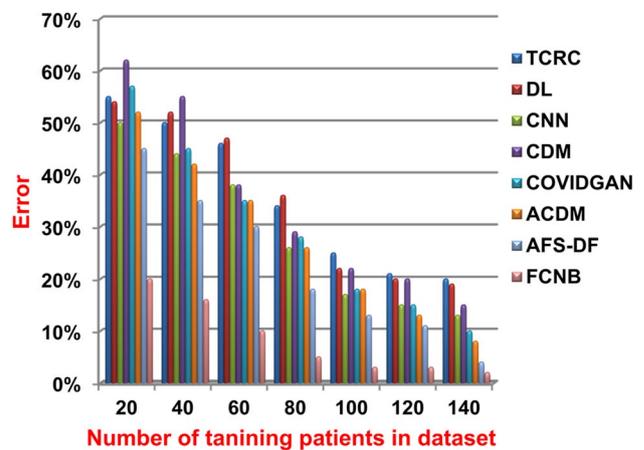


Fig. 25 Error of different COVID-19 classification techniques

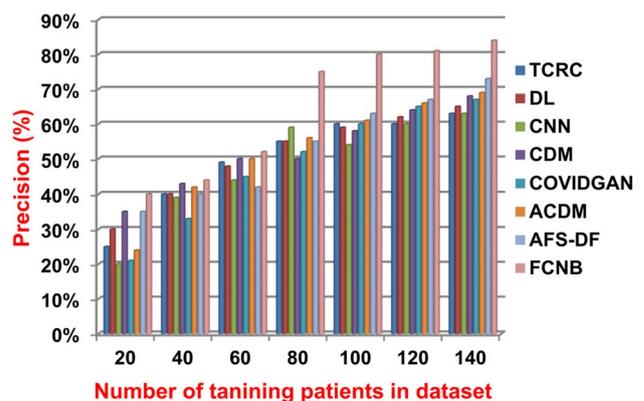


Fig. 26 Precision of different COVID-19 classification techniques

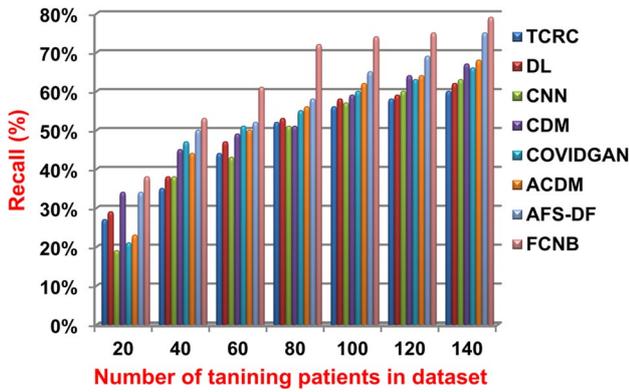


Fig. 27 Recall of different COVID-19 classification techniques

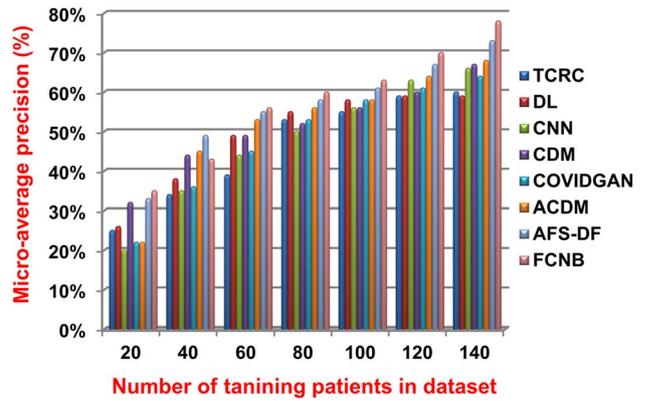


Fig. 30 Micro-average precision of different COVID-19 classification techniques

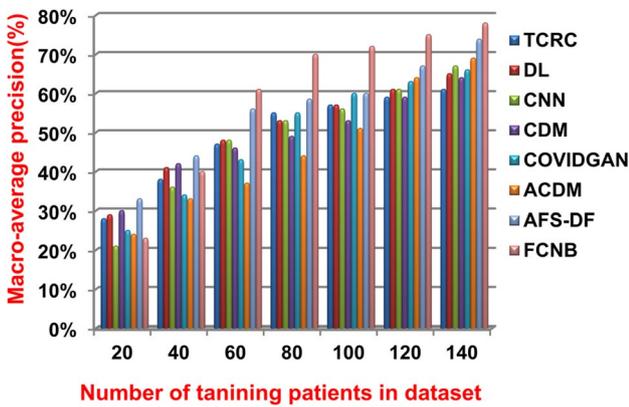


Fig. 28 Macro-average for precision of different COVID-19 classification techniques

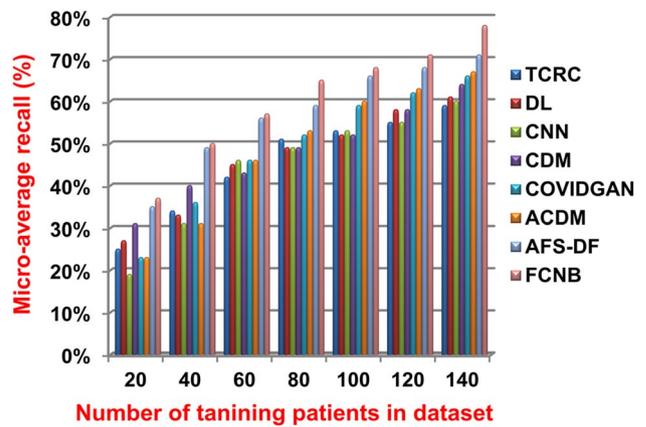


Fig. 31 Micro-average recall of different COVID-19 classification techniques

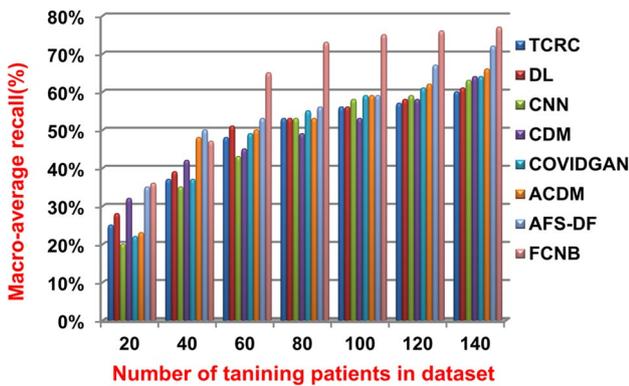


Fig. 29 Macro-average for recall of different COVID-19 classification techniques

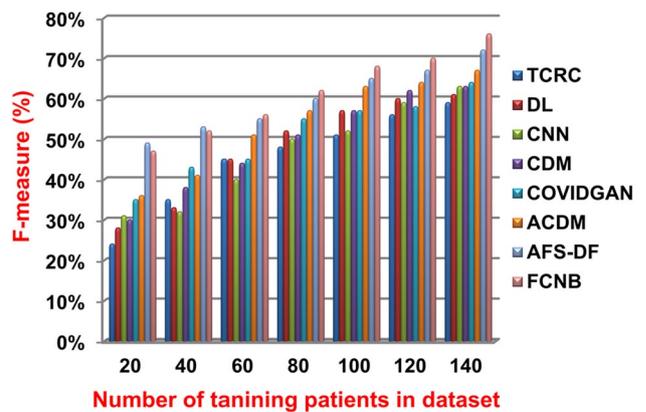


Fig. 32 F-measure of different COVID-19 classification techniques

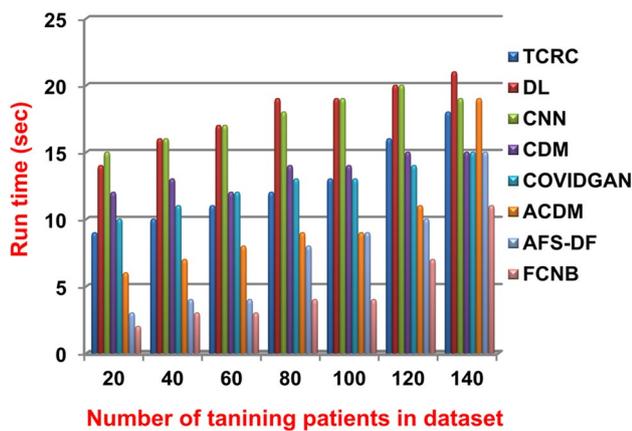


Fig. 33 Run time of the different classification techniques

FCNB gives the highest micro-average precision value equals 0.78 at the same training number of patients, while DL introduced 0.59 which is the lowest value of micro-average precision. FCNB provides micro average recall value that equals 0.78 while TCRC, DL, CNN, CDM, COVIDGAN, ACDM, and AFS-DF provide 0.59, 0.61, 0.60, 0.64, 0.66, 0.67, and 0.71 respectively. The highest F-measure value is introduced by FCNB with a value that equals 0.76, while the lowest value is introduced by TCRC with a value that equals 0.59 at the training number of patients = 140. In Fig. 33, the run time of FCNB is 11 (s) that represents the highest speed while DL introduces the lowest speed with run-time value equals to 20 (s). Finally, FCNB is better than other recent techniques which are; TCRC, DL, CNN, CDM, COVIDGAN, ACDM, and AFS-DF.

9 Conclusion and future work

It is very important to detect COVID-19 positive cases as early as possible to prevent the further spread of this pandemic and to quickly treat affected patients. In this paper, the FCNB classification strategy has been provided as a new COVID-19 diagnoses strategy to accurately diagnose COVID-19 patients with high speed. FCNB strategy is built upon two essential stages, which are; P²S and CS. P²S includes three essential phases, which are; FSP, FCP, and MFWP. In FSP, the most effective features on COVID-19 have been selected by using the FSGA method. In FCP, the selected features have been grouped into many clusters called Master Features (MFs) in which each MF includes a set of related features. In MFWP, each MF has been weighted based on the importance of the features it includes as well as the correlation among included features. On the other hand, in CS, the weighted NB has been implemented

on the weights of MFs rather than the weights of individual features to introduce fast and accurate diagnosis. Experimental results have shown that the proposed FCNB strategy increases the performance of the traditional weighted NB as it considers the correlation among features. Additionally, FCNB minimizes classification time as it considers small number of MFs rather than many individual features. In the future, we plan to apply the proposed FCNB strategy in fog on the COVID-19 dataset collected in the fog's cache server to provide fast diagnosis and to directly rehabilitate the infected people. In fact, this will greatly reduce the efforts of medical systems (e.g., hospitals) because fog depends on the Internet of Things (IoT) sensors that can automatically measure the body temperature and other symptoms to maintain social distance and to prevent spreading the infection.

References

- Abellán J, Castellano J (2017) Improving the naive bayes classifier via a quick variable selection method using maximum of entropy. *Entropy* 19(6):1–17
- Alazab M, Awajan A, Mesleh A, Abraham A et al (2020) COVID-19 prediction and detection using deep learning. *Int J Comput Inf Syst Ind Manag Appl* 12:168–181
- Ali ZH, Ali HA (2020) QoS provisioning framework for service-oriented internet of things (IoT). *Clust Comput* 23:575–591
- Arunadevi J, Ganeshamoorthi K, Rampriya R (2019) Application of feature weighting for the intensification of data classification. *IITEE* 9:879–887
- Ayed A, Halima M, Alimi A (2015) Survey on clustering methods: towards fuzzy clustering for big data. In: *Proceedings of the 2014 6th international conference of soft computing and pattern recognition (SoCPar)*. IEEE, Tunis, Tunisia, pp 331–336
- Ayyad S, Saleh AI, Labib L (2019) Gene expression cancer classification using modified K-Nearest Neighbors technique. *Bio-Systems* 176:41–51
- Bano S, Khan N (2018) A survey of data clustering methods. *Int J Adv Sci Technol* 113(2018):133–142
- Barstugan M, Ozkaya U, Ozturk S (2020) Coronavirus (COVID-19) Classification using CT images by machine learning methods. [arXiv:2003.09424](https://arxiv.org/abs/2003.09424)
- Benabdellah A, Bengahrit A, Bouhaddou I (2019) A survey of clustering algorithms for an industrial context. *Procedia Comput Sci* 148:291–302
- Berrar D (2018) Bayes' theorem and naive bayes classifier. *Encycl Bioinform Comput Biol* 1:403–412
- Brinati D, Campagner A, Ferrari D, Locatelli M et al (2020) Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *J Med Syst* 44(135):1–12
- Cabitzza F, Campagner A, Ferrari D, Di Resta C et al (2020) Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. <https://doi.org/10.1515/cclm-2020-1294>
- Chen H, Guo J, Wang C, Luo F et al (2020a) Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. *Lancet* 395(10226):809–815
- Chen X, Tang Y, Mo Y, Li S (2020b) A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological

- semantic and clinical features: a multi-center study. *Eur Radiol* 30:4893–4902
- Dada E, Bassi J, Chiroma H, Abdulhamid S et al (2019) Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5(6):1–23
- Das H, Naik B, Behera H (2020) A Jaya algorithm based wrapper method for optimal feature selection in supervised classification. *J King Saud Univ Comput Inf Sci*. <https://doi.org/10.1016/j.jksuci.2020.05.002>
- Dokmanic I, Parhizkar R, Ranieri J, Vetterli M (2015) Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Process Mag* 32(6):12–30
- Ferrari D, Motta A, Strollo M, Banfi G et al (2020) Routine blood tests as a potential diagnostic tool for COVID-19. *CCLM* 58(7):1095–1099
- Fletcher S, Slam M (2018) Comparing sets of patterns with the Jaccard index. *Austral J Inf Syst* 22:1–17
- Gietema H, Zelis N, Nobel J, Lambriks L et al (2020) CT in relation to RT-PCR in diagnosing COVID-19 in The Netherlands: a prospective study. *medRxiv*. <https://doi.org/10.1101/2020.04.22.20070441>
- Hewage P, Trovati M, Pereira E, Behera A (2020) Deep learning-based effective fine-grained weather forecasting model. *Pattern Anal Appl*. <https://doi.org/10.1007/s10044-020-00898-1>
- Huang C, Wang Y, Li X, Ren L et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395(10233):497–506
- Jabeen F, Maqsood M, Ghazanfar M, Aadil F et al (2019) An IoT based efficient hybrid recommender system for cardiovascular disease. *Peer-to-Peer Netw Appl* 12(5):1263–1276
- Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z et al (2020) Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *IEEE Access* 8:109581–109595
- Ji H, Huang S, Wu Y, Hui Z, Zheng C (2019) A new weighted naive Bayes method based on information diffusion for software defect prediction. *Softw Qual J* 27(3):923–968
- Jiang L, Zhang L, Yu L, Wang D (2019) Class-specific attribute weighted naive Bayes. *Pattern Recogn* 88:321–330
- Kang H, Xia L, Yan F, Wan Z et al (2020) Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. *IEEE Trans Med Imaging* 39(8):2606–2614
- Kasteren P, Veer B, Brink S, Wijnsman L et al (2020) Comparison of seven commercial RT-PCR diagnostic kits for COVID-19. *J Clin Virol* 128:1–5
- Kaur G, Oberoi A (2020) Novel approach for brain tumor detection based on Naïve Bayes classification. In: Sharma N, Chakrabarti A, Balas V (eds) *Data management, analytics and innovation. Advances in intelligent systems and computing* (1042). Springer, Singapore, pp 451–462. https://doi.org/10.1007/978-981-32-9949-8_31
- Kaviani P, Dhotre S (2017) Short survey on naive bayes algorithm. *Int J Adv Eng Res Dev* 4(11):607–611
- Khanday A, Rabani S, Khan Q, Rouf N et al (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. *Int J Inf Technol* 12:731–739
- Khotimah B, Miswanto M, Suprajitno H (2020) Optimization of feature selection using genetic algorithm in Naïve Bayes classification for incomplete data. *Int J Intell Eng Syst* 13(1):334–343
- Kovács A, Palásti P, Veréb D, Bozsik B et al (2020) The sensitivity and specificity of chest CT in the diagnosis of COVID-19. *Eur Radiol*. <https://doi.org/10.1007/s00330-020-07347-x>
- Kukar M, Gunčar G, Vovko T, Podnar S et al (2020) COVID-19 diagnosis by routine blood tests using machine learning. *arXiv preprint arXiv:2006.03476*
- Kumar D, Amgoth T, Annavarapu CH (2019) Machine learning algorithms for wireless sensor networks: a survey. *Inf Fus* 49:1–25
- Lee C, Gutierrez F, Dou D (2011) Calculating feature weights in naive bayes with Kullback-Leibler measure. In: *Proceedings of the 11th IEEE international conference on data mining, IEEE*, pp 1146–1151
- Lei Y, Yang B, Jiang X, Jia F et al (2020) Applications of machine learning to machine fault diagnosis: a review and roadmap. *Mech Syst Signal Process* 138:1–39
- Li K, Ping H, Zhou X, Li S (2016) Feature selection based on multiple correlation measures for medical examination dataset. In: *Proceedings of the advanced information management, communication, electronic and automation control conference (IMCEC)*. IEEE, Xi'an, China, pp 845–849
- Li Y, Yao L, Li J, Chen L, Song Y et al (2020a) Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. *J Med Virol*. <https://doi.org/10.1002/jmv.25786>
- Li C, Zhao C, Baoa J, Tang B et al (2020b) c) Laboratory diagnosis of coronavirus disease-2019 (COVID-19). *Clin Chim Acta* 510:35–46
- Li L, Qin L, Xu Z, Yin Y et al (2020c) Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* 296(2):65–71
- Li Q, Feng W, HuiQuan Y (2020d) Trend and forecasting of the COVID-19 outbreak in China. *J Infect* 80(4):469–496
- Li Y, Cao J, Zhang X, Liu G et al (2020e) Chest CT imaging characteristics of COVID-19 pneumonia in preschool children: a retrospective study. *BMC Pediatr* 20(227):1–8
- Liu Q, Du S, Wyk B, Sun Y (2020) Niching particle swarm optimization based on Euclidean distance and hierarchical clustering for multimodal optimization. *Nonlinear Dyn* 99:2459–2477
- Maghdid H, Asaad A, Ghafoor K, Sadiq A et al (2020) Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. *arXiv preprint arXiv:2004.00038*
- Mishra A, Das S, Roy P, Bandyopadhyay S (2020) Identifying COVID19 from chest CT images: a deep convolutional neural networks based approach. *J Healthc Eng* 2020:1–7
- Oluleye B, Leisa A, Leng J, Dean D (2014) A genetic algorithm-based feature selection. *Int J Electron Commun Comput Eng* 5(4):899–905
- Ozturk T, Talo M, Yildirim E, Baloglu U et al (2020a) COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Comput Biol Med* 121:1–12
- Ozturk T, Talo M, Yildirim E, Baloglu U et al (2020b) Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med* 121:1–11
- Pandit S, Gupta S (2011) A comparative study on distance measuring approaches for clustering. *Int J Res Comput Sci* 2(1):29–31
- Qiu P, Zhou Y, Wang F et al (2020) Clinical characteristics, laboratory outcome characteristics, comorbidities, and complications of related COVID-19 deceased: a systematic review and meta-analysis. *Aging Clin Exp Res* 32:1869–1878
- Rabie AH, Saleh AI, Abo-Al-Ez K (2015) A new strategy of load forecasting technique for smart grids. *IJMTER* 2(12):332–341
- Rabie AH, Ali SH, Ali HA, Saleh AI (2019a) A fog based load forecasting strategy for smart grids using big electrical data. *Clust Comput* 22(1):241–270
- Rabie AH, Ali SH, Saleh AI, Ali HA (2019b) A new outlier rejection methodology for supporting load forecasting in smart grids based on big data. *Clust Comput* 23(2):509–535
- Rabie AH, Ali SH, Saleh AI, Ali HA (2020) A fog based load forecasting strategy based on multi-ensemble classification for smart grids. *J Ambient Intell Hum Comput* 11(1):209–236

- Rustam F, Reshi A, Mehmood A, Ullah S et al (2020) COVID-19 future forecasting using supervised machine learning models. *IEEE Access* 8:101489–101499
- Saleh AI, Rabie AH, Abo-Al-Ezb K (2016) A data mining based load forecasting strategy for smart electrical grids. *Adv Eng Inform* 30(3):422–448
- Shegal S, Agarwal M, Gupta D, Sundaram S et al (2020) Optimized grass hopper algorithm for diagnosis of Parkinson's disease. *SN Appl Sci* 2(6):1–18
- Shaban W, Rabie AH, Saleh AI, Abo-Elsoud M (2020) A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowl-Based Syst* 205:1–8
- Shanthi S, Rajkumar N (2020) Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods. *Neural Process Lett*. <https://doi.org/10.1007/s11063-020-10192-0>
- Shinde G, Kalamkar A, Mahalle P, Dey N et al (2020) Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. *SN Comput Sci* 1(197):1–15
- Shirkhorshidi A, Aghabozorgi S, Wah T (2015) comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS One* 10(12):1–20
- Sivanandam S, Deepa S (2008) *Introduction to genetic algorithms*. Springer, Berlin
- Sowmiya C, Sumitra P (2020) A hybrid approach for mortality prediction for heart patients using ACO-HKNN. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-020-02027-6>
- Subramanian R, Prabha D (2020) Customer behavior analysis using Naïve Bayes with bagging homogeneous feature selection approach. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-020-01961-9>
- Sun L, Mo Z, Yan F, Xia L et al (2020) Adaptive feature selection guided deep forest for COVID-19 classification with chest CT. *IEEE J Biomed Health Inform* 24(10):2798–2805
- Taha A, Mustapha A, Chen S (2013) Naive Bayes-guided bat algorithm for feature selection. *Sci World J* 2013:1–10
- Tahamtan A, Ardebili A (2020) Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev Mol Diagn* 20(5):453–454
- Taheri S, Yearwood J, Mammadov M, Seifollahi S (2014) Attribute weighted Naive Bayes classifier using a local optimization. *Neural Comput Appl* 24(5):995–1002
- Visa S, Ramsay B, Ralescu A, Knaap E (2011) Confusion matrix-based feature selection. In: *Proceedings of the twenty-second midwest artificial intelligence and cognitive science conference (MAICS)*. Cincinnati, USA, pp 120–127
- Waheed A, Goyal M, Gupta D, Khanna A et al (2020) CovidGAN: data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access* 8:91916–91923
- Waller J, Kaur P, Tucker A, Lin K et al (2020) Diagnostic tools for coronavirus disease (COVID-19): comparing CT and RT-PCR viral nucleic acid testing. *Am J Roentgenol* 215(4):1–5
- Wang S, Zha Y, Li W, Wu Q et al (2020) A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *Eur Respir J* 39(8):1–44
- Wosiak A, Zakrzewska D (2018) Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. *Complexity* 2018:1–12
- Yearwood J, Taheri S, Mammadov M, Seifollahi S (2014) Attribute weighted Naive Bayes classifier using a local optimization. *Neural Comput Appl* 24(5):995–1002
- Yu L, Jiang L, Wang D, Zhang L (2019) Toward naive Bayes with attribute value weighting. *Neural Comput Appl* 31(10):5699–5713
- Zhang H, Jiang L, Yu L (2021) Attribute and instance weighted naive Bayes. *Pattern Recogn* 11:1–11
- Zhong L, Mu L, Li J, Wang J et al (2020) Early prediction of the 2019 novel coronavirus outbreak in the Mainland China based on simple mathematical model. *IEEE Access* 8:51761–51769
- Zhu X, Wang Y, Li Y, Tan Y et al (2019) A new unsupervised feature selection algorithm using similarity-based feature clustering. *Comput Intell* 35(1):2–22
- Zu Z, Jiang M, Xu P, Chen W et al (2020) Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology* 296(2):15–25

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.