Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Chronic obstructive pulmonary disease prediction using Internet of things-spiro system and fuzzy-based quantum neural network classifier

G.S. Karthick [a,b,*], P.B. Pankajavalli [b,*]

[a] Department of Software Systems, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India
[b] Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 13 April 2022
Received in revised form 25 July 2022
Accepted 19 August 2022
Available online xxxx

Keywords:
Keywords volatile organic compounds
Exhaled breath
IoT Spiro-system
COPD diagnosis
Feature selection
Machine learning classifiers

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a multifarious progressive disease that increases the mortality and morbidity ratio as well as becoming a life-threatening issue of an individual. Accurate and cost-effective diagnosis of diseases plays a primary role in the medical domain and a wide range of research has been carried out on disease prediction using sensory approaches along with the assistance of machine learning techniques. The traditional disease diagnosis procedures are invasive, costlier and the decision support systems were unreliable most of the time. The human exhaled breath discharged from the body is composed of various Volatile Organic Compounds (VOCs) which can be influenced by metabolic and disease activities. Hence, the analysis of VOCs in exhaled breath has an incredible potentiality for COPD diagnosis and can rapidly decrease the mortality rate. In this research, IoT-Spiro System is designed and an intelligent machine learning forecasting framework (IMLFF) has been proposed. IoT-Spiro System perceives the various VOCs patterns available in exhaled breath and that real-time parameter has been analyzed using IMLFF. The proposed framework incorporates a hybrid Genetic Big Bang-Big Crunch (GBB-BC) algorithm for selecting the optimal features from the real-time dataset and a Fuzzy-based Quantum Neural Network (F-QNN) classifier for diagnosing COPD. The experimental results illustrate that IMLFF outperforms when compared to recent existing approaches concerning various statistical parameters and performance metrics. From the result analysis, it has been determined that IoT-Spiro System and IMLFF framework can serve as an efficient assisting model to the medical practitioner for diagnosing COPD.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous and slowly progressive disease characterized by irreversible airway inflammation, which is a fourth-leading disease that causes mortality worldwide. Due to the inadequacy of early chronic obstructive pulmonary disease (COPD) diagnosis, the death ratio is constantly increasing and thereby it's becoming a life-threatening challenge to the healthcare service providers [1,2]. Accurate and well-timed disease diagnosis is obligatory for providing effective treatment and to reduce the cost involved in personalized treatment. The disease diagnosis and treatment planning are currently based on the invasive disease diagnosis procedures and knowledge of the healthcare

* Corresponding authors.
E-mail addresses: karthick@psgcas.ac.in (G.S. Karthick), pankajavalli@buc.edu.in (P.B. Pankajavalli).

https://doi.org/10.1016/j.tcs.2022.08.021
0304-3975/

provider rather than intelligent tools. On accounting for the major challenges posing towards the early disease diagnosis, there is a never-ending search for appropriate techniques which can potentially diagnose the disease at an early stage and improve the patient's experience. From the study, a non-invasive method of analyzing the volatile organic compounds (VOCs) in exhaled breath can mentor as promising biomarkers in early disease diagnosis [3]. The origination of VOCs in exhaled breath can be either exogenous or endogenous, in which endogenous group of VOCs is formed by lipid peroxidation or pathophysiological process or bacteria in the liver, kidney, and lungs [4].

Most of the invasive methods used in COPD diagnosis are depending upon the patient's medical record, symptom analysis, and clinical examination report. These methods may tend to provide a wrong diagnosis and also it may create a huge time delay due to human errors; besides, it is more expensive and computationally complex. The patient record is a file that holds the clinical history which includes identity, clinical trials, examinations reports, treatments, medications, and other healthcare services rendered to the patient [5,6]. Moreover, the medical datasets are in the form of digitized translation of patient records which upholds various attributes or features. The critical feature selection [7] has become an important arena where many healthcare researchers were focusing on the achievement of accurate disease diagnosis [8]. In most healthcare datasets, the number of attributes or features may range from tens to thousands on which feature selection can be applied for finding the significant subset of features that can extensively improve the accuracy of the classifier models.

Internet of Things (IoT) is a ground-breaking technology adapted to medical devices which facilitates remote monitoring of patients, discharging the potential to save patient's life and empowering the healthcare providers to deliver ultimate care [9]. The patient commitment and satisfaction towards interaction with physicians have also increased and become more efficient. IoT is undoubtedly creating an impact by redefining the space between the patients and healthcare providers by removing the geographical barriers and improving treatment outcomes. The key beneficiary aspect of remote patient monitoring is supported by cloud technology, where the data collected via sensor are stored and which can be accessed remotely without any physical constraints [10]. In addition, IoT-based medical devices with cloud assistance will serve as a time-critical application for diagnosing the disease. Machine learning owes a major role in the decision-making process by handling a huge amount of data. Disease diagnosis using IoT and machine learning approaches might deal with an enormous amount of data which are of different forms and also data privacy may arise when the data belongs to healthcare applications [11].

From the literature, it has been witnessed that the researchers have not consented to a generalized machine learning framework for the diagnosis of COPD. Most of the invasive methods used in COPD diagnosis are depending upon the patient's medical record, symptom analysis, and clinical examination report. These methods may tend to provide a wrong diagnosis and also it may create a huge time delay due to human errors; besides, it is more computationally expensive and complex. Therefore, IoT-based exhaled VOC analysis can be a capable non-invasive method for personalized monitoring and disease diagnosis. The main objective of the present work is to minimize the hurdles and barriers in disease diagnosis by developing an IoT-Spiro System for gathering the VOC patterns in human exhaled breath. Then the gathered VOC patterns are analyzed using the proposed machine learning framework, which offers the various impending benefits such as early disease diagnosis, cost reduction, personalized health monitoring, better-quality care, and patient experience. The major research contributions rendered in this article are compiled and summarized below:

- A hybrid Genetic Big Bang-Big Crunch (GBB-BC) algorithm is presented for optimal feature selection, which comprises of Genetic Algorithm (GA), Vantage Point-Tree, and Big Bang Big Crunch (BB-BC) Algorithm. This forms the root cause objective of the research work that optimizes the feature selection operation for enhancing the disease prediction accuracy.
- The output of the hybrid GBB-BC optimal feature selection algorithm has been analyzed using the proposed Fuzzy-based QNN classifier which is comprised of the Fuzzy Analytic Hierarchy Process (Fuzzy AHP) and Quantum Neural Network (QNN).
- The result of the Fuzzy-based Quantum Neural Network (F-QNN) classifier is compared with the existing techniques to substantiate the supremacy of the proposed classifier.
- Finally, this research article recommends the GBB-BC optimal feature selection algorithm and F-QNN classifier which yields a high level of accuracy in COPD diagnosis.

The remaining part of this research article is organized as follows: Section 2 briefly discusses data collection, preprocessing, feature selection, and classification models used in the machine learning framework of this research. Section 4 elaborately presents the experimental results with statistical analysis. Finally, Section 5 of the article is concerned with the conclusion and the future recommendation of this research work is summarized.

2. Related works

2.1. VOC analysis for COPD diagnosis

Pulmonary disease and diversified clinical challenges are the major cause of morbidity in children and adults [12,13]. The non-invasive diagnosis of COPD would be extremely advantageous in developing a personalized healthcare system for

the future. Most of the VOC patterns witnessed in the exhaled breath may be formed or modified by the pathophysiological process in human body parts. Various researchers investigated the distinctive patterns of VOCs were found in COPD subjects and its significant subjects. The research works [14] and [15] have suggested that the VOC patterns in exhaled breath may be useful in predicting COPD in patients. In the past, several techniques have been used for collecting and analyzing the exhaled VOCs [16,17] and in which gas chromatography (GC) and eNose are commonly used techniques. GC techniques initially collect the exhaled breath samples and store it in inert bags and then the individual VOCs are analyzed by mass spectrometry or flame ionization detection mechanism [16]. The eNose can also be used for analyzing the breath samples with the assistance of nanosensors incorporated within the device [17]. The major drawback of eNose is that the analysis of individual VOC patterns from exhaled breath is impossible and none of the techniques are cost-effective and the accessibility ratio is also too low. Enormous researchers illustrated that VOC profiles [18–20] can accurately discriminate COPD subjects from healthy subjects. Fens et al. reported that the VOC profiles of COPD patients were associated with eosinophilic and neutrophilic cell counts abnormality [21].

2.2. Feature selection

The existence of irrelevant data inside the feature set must be reduced for extracting the subset of features is commonly referred to as feature selection in machine learning and optimization problems [8]. In this scope, feature selection is significant to enhance the quality and speed up the learning process of the data. Feature selection is to choose a subset of attributes from the input data which can proficiently define the input data while the impact of irrelevant and noise is minimized without compromising the disease prediction. Especially, it plays an imperative role in shrinking the progression scale of data whereas irrelevant and replicated features were eliminated. Feature selection is an effective practice in preprocessing the data with high dimensions and also the effectiveness of the learning is improved. The necessity of obtaining optimal features works well with an optimization algorithm that reduces the error factor in the disease prediction [22]. Generally, wrapper or filter techniques are used in feature selection [23–25] whereas in this research work, GA and BB-BC algorithms were employed together with a vantage point tree for optimizing the critical feature selection.

Genetic Algorithm (GA) shows significance in optimization-based issues and it is also a robust search method in the multimodal landscape [26]. GA is inspired by natural genetics and evolution through which an optimal global solution can be attained with high probability and it has high exploration capability towards searching for optimal features in a search space [27]. The execution of GA is initiated with the random distribution of a chromosome. The occurrence of crossover and mutation generates the subsequent generation of items which poses diversifying characteristics. The schema of GA was assessed and the context of better production opportunities was allocated to the chromosome that directs to the best solution. Due to its high exploration capability, GA has been applied on various domains particularly for solving the feature or attribute selection problems such as Das et al. [28], Ratta et al. [29], Garcia et al. [30], and Triguero et al. [31] but GA does not have exploitation ability to identify optimal feature around the search space.

Big Bang–Big Crunch (BB-BC) algorithm was formulated from cosmological science in which big bang describes the expansion of the universe from an explosion of the particles in the space, whereas the big crunch theory signifies the gravitational force and it inversed the process of expansion which ultimately pull back everything into its original state [32]. From the wider perception, the universal evolution was an eternal process of big bang expansion and big crunch constriction phase. In the big bang phase, candidate items are generated for optimization problems and the items were dispersed over the exploration space [33]. In the big crunch phase, randomly dispersed candidates were drawn into distinct regions which are assigned based on the centric point of the population. The output of the big bang phase was fed as an input to the big crunch phase. Convergence of the expansion phase is reinstated with fitness value and current position of the candidate item to generate weighted average point which is center of mass. The expansion and contraction are reiterated until the termination criterion is reached. Even though BB-BC endows good exploitation capability but it suffers from a lack of exploration capability [34].

The vantage-point tree (VP Tree) is a metric tree that splits the information in a metric space by choosing a needed position in a space [35]. The VP Tree partitions the data points into two levels that are data points closer to the vantage point than the assigned threshold value and the points that are not close to the data point. The partitioning procedure is applied recursively into data and a tree structure is generated. Every node in the tree holds the input and radius value. In common, the VP tree serves better for searching the desired number of nearest neighbors and it can scale up with any size of the dataset.

2.3. Machine learning classifiers

The invasive-based methods for disease diagnosis depend only on the patient's medical records, physical investigation report, and examination of concerned symptoms by medical practitioners. The existing methods typically cause inaccurate diagnosis and also consume more time for diagnosing due to its error-prone characteristics. Besides, its computational complexity is also high and more expansive. To overcome these drawbacks in invasive-based disease diagnosis, a non-invasive system has been developed by many researchers with the support of machine learning techniques. Machine learning predictive models like k-Nearest Neighbor (k-NN), AdaBoost (AB), Decision Tree (DT), Support Vector Machine (SVM), Artificial

Neural Network (ANN), Logistic Regression (LR), Naïve Bayes (NB), and Fuzzy Logic (FL) has been highly considered by many researchers for earlier and accurate diagnosis of diseases [36].

K-NN is a simple supervised learning algorithm, which learns all the possible cases and then predicts the class labels for new cases with the help of similarity measures [37]. There are numerous methods have been used for calculating the similarities between two instances with n number of attributes. Every method has the following three properties. Consider dist(X,Y) be the distance between two points X, Y then,

(i) dist(X,Y)$\geq$0 and dist(X,Y)=0 iff X=Y
(ii) dist(X,Y)=dist(Y,X)
(iii) dist(X,Z)$\leq$dist(X,Y)+dist(Y,Z)

Property 3 describes the triangle in equality which is a straight line and also the shortest distance between two points [38]. From the studies, it has been identified that the performance of K-NN is not good and research is essential to enhance the accuracy and efficiency.

The SVM is a machine learning classification model which uses a margin strategy for solving complex problems and it results by offering high performance when it has been applied for classification applications [39,40]. According to a binary classification problem, the attributes are divided with a hyper plane $v^T d+m=0$, where v is a dimensional coefficient vector, m is the offset value from the origin, d is dataset values. The results of SVM are v and m, in which v can be obtained by using linear case lagrangian multipliers. The data points plotted on borders are support vectors and therefore, the solution v can be written as $v = \sum_{i=1}^{n} \propto_i c_i d_i$, where n denotes the number of support vectors, c_i are target labels to d_i.

The NB is a classification algorithm that depends on the conditional probability theorem which identifies the class labels of a new feature vector [41]. For each given class, the NB calculates the conditional probability of a training dataset. Further, the probability is computed for a given new vector class by considering each vector independently and this algorithm is well-concerned for the text-based classification problems.

The ANN is an iterative supervised machine learning algorithm that solves the problems without identifying and explaining the method of a particular problem to be solved, without constructing algorithms or developing programs, and also without knowing the problem to be solved [42]. Moreover, the ANN comprises three major constituents such as inputs, functions, and outputs. The input component utilizes the attribute values and weights, in which the weights of each attribute are modified during the training process of the neural network. The output component produces the known class by altering the weights to stabilize the error rate between the output obtained and the actual class. Most of the applications of ANN indirectly identify the multifaceted relationships between the variables [43].

Multi-Criterion Decision Making (MCDM) issues were handled with the Analytical Hierarchy Process (AHP) in addition to fuzzy set theory [44]. The implementation of fuzzy numbers achieved the sensible judgments of humans. The decision-maker solved the most complex criterion with the assist of the Fuzzy Analytic Hierarchy Process (FAHP), in which pair-wise comparison is done and numerous alternate solutions were generated based on the preferences of the decision-makers. FAHP tends to yield a high degree of accuracy by considering various fuzzy metrics and better results were achieved through multiple preferences.

Quantum Neural Network (QNN) computation is evolved from the mathematical model and the nature of quantum mechanics which is a certainty of physics. QNN is an explicit depiction of the process of computation that is the correlation of behavior with other systems. QNN had achieved rapid parallel solutions by considering poly-logarithmic size and depth. In QNN the process of computation is attained in a few steps with $(O(\log^k n), k \geq 1)$ [45]. The significance of complexity-theory channels numerous aspects of computation that is depth, size, and precision. In the QNN, depth denotes the iteration; size denotes the size of the dataset, and precision is a problem-solving technique. The computational complexity is minimized with the QNN approach [46].

3. Materials and methods

3.1. Intelligent machine learning forecasting framework for COPD diagnosis

The main objective of developing an intelligent machine learning forecasting framework for COPD diagnosis is to support the IoT-Spiro system in the process of decision making with a high degree of accuracy and early diagnosis. The intention of this framework may lead to serving as an efficient assisting model to the medical practitioner for automated diagnosing of COPD. The developed IMLFF aims to acquire the real-time dataset using IoT-Spiro System and to pre-process the highly variant real-time clinical dataset, thus pre-processing is a practicable step. Then the intelligent framework consists of optimal feature selection along with examining the impact of machine learning classification models. To facilitate the understanding of the developed framework IMLFF, the workflow is illustrated schematically in Fig. 1. In this section, the process of acquiring real-time data using the IoT-Spiro System is explained. Second, the preprocessing method used for standardizing the real-time data is briefed. Finally, the mechanism used for optimal feature selection and classification models used for diagnosing COPD patients and healthy subjects is explained.

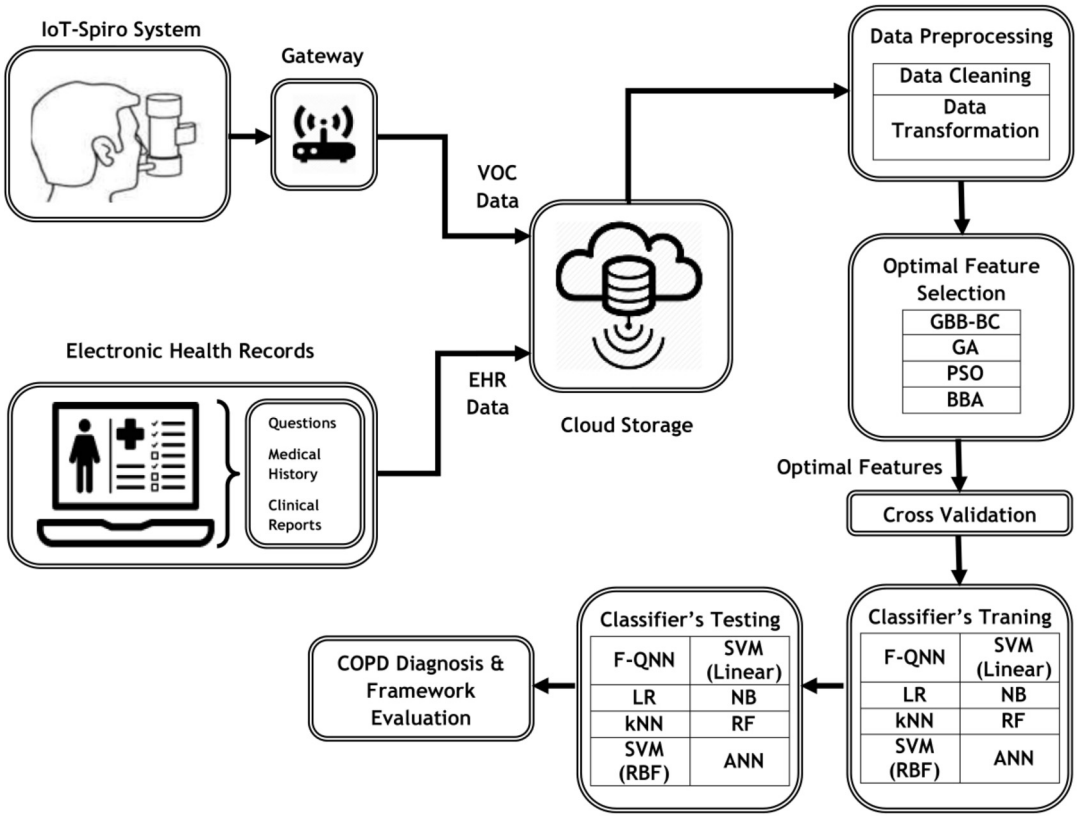

Fig. 1. IMLFF: intelligent machine learning forecasting framework.

3.1.1. Real-time data acquisition using IoT-Spiro system

The improvements in computational power and real-time data analytics influenced the healthcare industry with accurate disease diagnoses. There has been a huge demand for cost-effective and accurate disease diagnosis, which in turn results in the development of the IoT-Spiro System for acquiring the data from healthy and non-healthy subjects. This device incorporates an array of gas sensors, microcontrollers, a Wi-Fi module, and other supporting components. The developed device recognizes twelve VOCs which include isobutene, ethanol, acetone, benzene, formaldehyde, ammonia, nitrogen dioxide, carbon monoxide, hydrogen, propane, toluene, and acetaldehyde. Then, the VOC data were uploaded to the cloud storage using the pushing box, which is an application programming interface (API) that acquires data from a microcontroller to cloud storage. From the cloud storage data were extracted and analyzed using the proposed machine learning framework.

In this research experiment, 150 healthy subjects and 150 COPD subjects (160 males and 140 females, age ranges from 21 to 70 years) have participated. The subjects included in this study were accepted and signed the informed consent before the experiment. The work was scrutinized and approved by the Institutional Human Ethical Committee of Bharathiar University (BUHEC). The data collection was conducted at Kongunadu Hospital Campus and the subjects were seated comfortably during the process. For each subject, vital clinical factors were considered and acquired that consist of numerical and nominal features. The factors which are more prominent for COPD diagnosis have been identified from the relevant medical literature and also based on the opinions of the senior medical experts. The acquired real-time dataset contains 30 features, in which 12 different VOC features and 18 prominent medical factors related to COPD diagnosis have been considered.

3.1.2. Preprocessing

The characteristics of medical data are originated via the clinical process and wearable devices tend to overcome various issues like availability of data and the complex representation of data turns the decision-making applications challenging. Preprocessing is mandatory for transforming the data into well-formed and efficient data, so the machine learning algorithms can be utilized efficiently. As a part of the development of the machine learning framework for disease diagnosis using the IoT-Spiro system, preprocessing methods like the elimination of missing values, MinMaxScaler, and Standard Scaler has been used to standardize the dataset [32]. The missing values are identified and eliminated from the dataset. MinMaxScaler strategy can be useful for machine learning algorithms when the dataset contains the input values with differing scales and therefore, the application of the MinMaxSxaler technique to the dataset improves the efficacy of classifier training. The

MinMaxScaler is a typical technique that transforms given features or attributes individually by scaling it within a given range (between 0 and 1), can be written as,

$$\frac{f_i - \min(f)}{\max(f) - \min(f)} \tag{1}$$

Similarly, Standard Scaler ensures whether the value of each attribute or feature is normally distributed and scales them to mean zero and variance one, can be written as,

$$\frac{f_i - mean(f)}{stdev(f)} \tag{2}$$

where f is the feature or attribute of a given dataset. Therefore, effective preprocessing techniques were used in this research work to avoid the negative outcome due to noisy and un-normalized data.

3.1.3. Feature selection using hybrid GBB-BC algorithm

Feature selection is an efficient technique in a machine learning framework that identifies the significant features related to the disease. The identification of significant features helps in eliminating unnecessary, redundant features from the dataset which directly impose the machine learning classifiers to provide quick and better results. Besides, this technique improves the comprehensibility of healthcare data, reduces the complexity and training time of machine learning classifiers. In this research work, GA has been employed to select the optimal features because this technique explores the search space of all probable subsets to obtain the optimal features which minimize the irrelevant features and maximize the predictive accuracy of the disease diagnosis system. The disadvantage of GA is that suffers from a lack of exploitation and to overcome this downside, the present work developed a hybrid metaheuristic feature selection model that fuses GA, BB-BC, and vantage point tree.

The potentiality of the metaheuristic techniques is identified by its ability to balance the diversification and intensification throughout the process of searching the optimal features. Intensification is an integral factor that deepens the process of searching and identifying the neighborhood candidates of the optimal solution, whereas diversification is another co-integral factor that enables the algorithm to explore the search space efficiently. The extreme intensification of the algorithm increases the risk of searching the candidates in a local optimal space by visiting only a portion of the local search space. However, extreme diversification may tend to have slower convergence, where solutions are residing around the optimal solution.

The BB-BC algorithm reveals the superior intensification capability, but it suffers from the pitfall of less diversification. At the big bang phase, most of the candidates are to be found in a small portion of the search space but few candidates may be locked in the subdomain space. To overcome this pitfall, a GBB-BC has been developed which blends BB-BC, GA, and VP Tree, the procedure flow is shown in Fig. 2 as well as the steps were presented in Algorithm 1. The benefits of each of these algorithms are combined to conquer the shortcomings of its other constituent algorithms. The BB-BC algorithm holds the ability of exploitation (intensification) and obtains the optimal attributes with the help of the VP Tree by searching the candidates around the center of mass. The GA facilitated with the exploration (diversification) ability which avoids the searching of candidates only within a local optimal space and upholds the exploration widely among the search space.

The population is initiated randomly and the fitness values are computed for each candidate in the population. The population is divided into two different groups after initialization and the candidates are allocated to the groups using linear distribution mechanisms. Then one group will pass into the GA phase and another group will pass into the BC phase, as follows:

$$sGA = \left(1 - \frac{k}{Ti}\right) \mathrm{x} nP \tag{3}$$

$$sBC = \frac{k}{Ti} \quad \mathrm{x} nP \tag{4}$$

where sGA is the candidate size that passes into the GA phase, sBC is the candidate size that passes into the BC phase, k refers to the current iteration, Ti refers to the total iterations and nP refers to the total candidates in the population.

The size of the candidate's passes into the GA phase is relatively high at the beginning and then the size of the candidate's passes into the GA phase will linearly decrease whereas the size of the candidate's passes into the BC phase will linearly increase. The implementation of this rule avoids the poor diversification (exploration) ability of the BB-BC algorithm and however, the poor intensification (exploitation) ability of the GA algorithm is rectified by the big bang operator at the end.

The GA and BC phases are executed in parallel to obtain the solution set which contains the optimal attributes and also to overcome the disadvantages of the constituent algorithms. In the GA phase, the optimization process starts by applying genetic operators such as selection, crossover, and mutation on the selected solutions from the population. The crossover and mutation can be considered as important evolutionary operators of GA. The solutions set is generated by selecting a two-parent candidate solution using tournament selection, followed by a single or double or uniform crossover operation is applied to the selected parent candidates to generate two offspring solutions. During mutation operation, random local

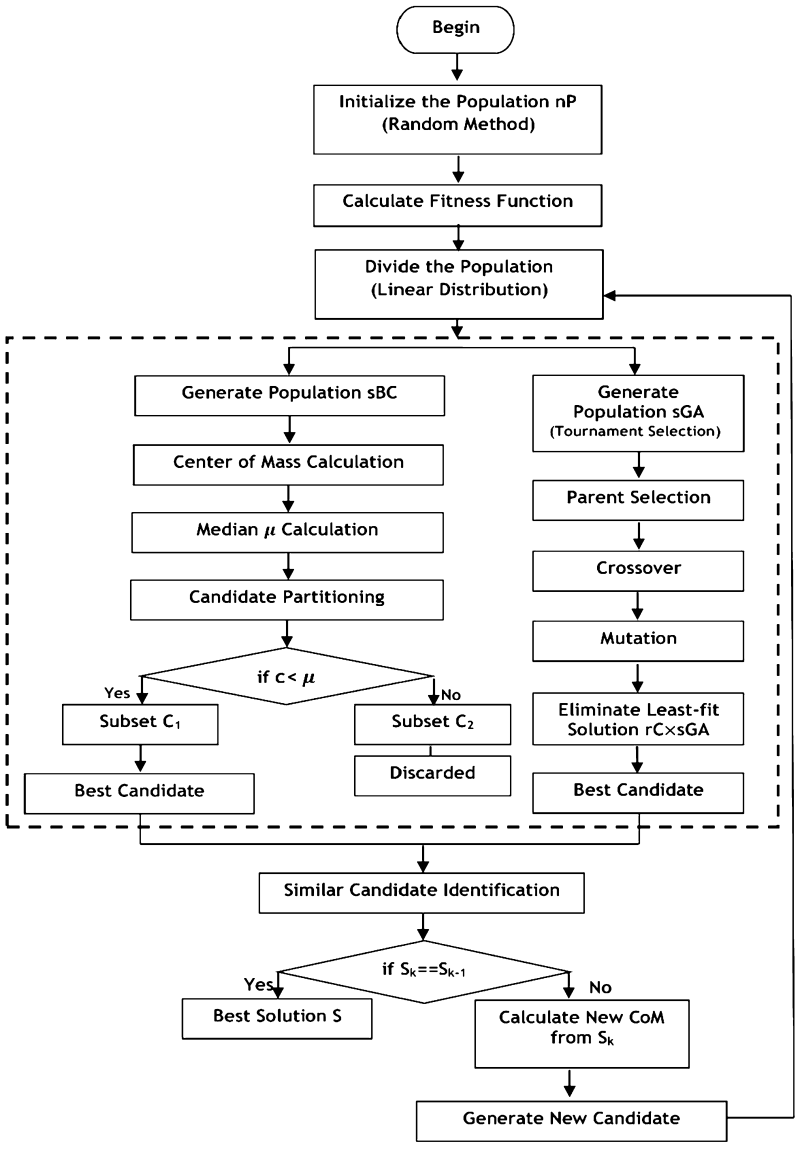

Fig. 2. Flowchart of GBB-BC optimal feature selection algorithm for COPD diagnosis.

changes are applied to the generated offspring which updates the population by swapping a few candidate solutions in the population using the elitism strategy. The updated population is taken into account for consideration in the next iteration and at each iteration; the solution set is evaluated using a fitness function. Finally, the best candidate solution set $S_{x,k}$ (best candidate solution of the GA phase at current iteration) is returned by the GA phase. Table 1 illustrates the parameters setting of GA.

The initial process of the BC phase tends to calculate the center of mass (CoM) for the candidates passed into the phase using the equation,

$$CoM_{pos} = \frac{\sum_{i=1}^{n} \frac{1}{fit_k^i} pos_k^i}{\sum_{i=1}^{n} \frac{1}{fit_k^i}} \tag{5}$$

The VP tree mechanism is used for improving the exploitation process of the BC phase and the VP tree mechanism is applied to the CoM_{pos} to intensify the searching process [35]. Every CoM_{pos} is considered as a vantage point that acts as a root node and from CoM_{pos} point distance values of all other candidate points are determined. Then, μbe the median of obtained distance values is computed based on,

Algorithm 1: Hybrid GBB-BC Algorithm for Optimal Feature Selection

Input: D: Dataset, P: P: Population, nP: Population Size, Ti: Total Number of Iterations, rC: Crossover Rate, rM: Mutation Rate

Output: S: Set of Optimal features

Procedure:

1. Begin
2. Initialize the Population P of size nP using Random Method
3. Calculate the fitness function of each candidate in population P
4. Divide the population P into two groups by using Linear Distribution:

$sGA = (1 - \frac{k}{Ti}) \times nP$

$sBC = \frac{k}{Ti} \times nP$

GA Phase:

5. Select two parent candidates from population *sGA* using tournament selection
6. Then perform crossover operation on selected parent candidates to produce offspring solutions
7. If (rand(0.0,0.1)≤rM)
8. Perform mutation on offspring's solutions
9. End If
10. Add the offspring's to the population
11. Eliminate the identified least-fit solutions from the population by applying: rC x sGA
12. Return best candidate solutions, $S_{x,k}$

Big Crunch Phase:

13. Compute the CoM for each set of candidates

$CoM_{pos} = \frac{\sum_{i=1}^{n} \frac{1}{fit_k^i} pos_k^i}{\sum_{i=1}^{n} \frac{1}{fit_k^i}}$

14. Determine the distance $d(CoM_{pos}, c_i)$; $i = 1, 2, \ldots..n$.
15. Calculate the median μ of the distance values

$\mu = \frac{(\frac{nD}{2}) + (\frac{nD}{2} + 1)}{2}$

16. Partition the candidates set into two subsets C_1 and C_2,

$C_1 = \{c \in C_1 \mid d(CoM_{pos}, c_i)_{<} \mu\}$

$C_2 = \{c \in C_2 \mid d(CoM_{pos}, c_i) \geq \mu\}$

17. Accept the subset C_1 as a solution set $S_{y,k}$ of BC phase and discard subset C_2.
18. Identify the similar candidates from the solutions $S_{x,k}$ and $S_{y,k}$ using $S_k = S_{x,k} \cup S_{y,k}$.
19. If ($S_k == S_{k-1}$) then $S = S_k$
20. Return S
21. Else

Big Bang Phase:

22. Calculate the new CoM from the candidate solution S_k.
23. Generate the new candidates around the new CoM using the equation:

$P_{k+1} = CoM_{pos} + \frac{rand_k \times \omega \times (P_{\max} - P_{\min})}{k+1}$

24. End If
25. Repeat from Step 2 to Step 19 until the condition specified at Step 19 is satisfied.
26. End.

Table 1
Parameters setting for GA phase.

Parameters	Values
Population Size	30
Selection Mechanism	Tournament Selection
Crossover Mechanism	Single, Double or Uniform
Crossover Rate	0.7
Mutation Rate	0.1

$$\mu = \frac{(\frac{nD}{2}) + (\frac{nD}{2} + 1)}{2} \tag{6}$$

where μbe the median, nD is the number of candidates in the given population. Then consider μ value as a threshold for partitioning the candidate solution into two candidate subsets namely C_1 and C_2, which has been defined as,

$$C_1 = \left\{c \in C_1 \mid d(CoM_{pos}, c_i)_{<} \mu\right\}$$

$$C_2 = \left\{ c \in C_2 \mid d(CoM_{pos}, c_i) \geq \mu \right\}$$

where $d(CoM_{pos}, c_i)$ is the distance between the points CoM_{pos} and c_i, c_i denotes the candidates. The candidate subset C_1 can be considered as a candidate solution set $S_{y,k}$ (best candidate solution of the BC phase at current iteration) which satisfies the condition, and C_2 can be discarded. From the solutions obtained from GA and BC phases, similar candidates at current iteration S_k are identified which can be defined as $S_k = S_{x,k} \cup S_{y,k}$. Then the candidate solution set S_k is compared with the candidate solution set generated by the previous iteration to identify the chance of obtaining significant changes in the forthcoming iterations. In case the condition ($S_k == S_{k-1}$) is satisfied then the current candidate solution set is returned as optimal feature solution set S and in the state of unsatisfied condition, a candidate is elected as a new CoM. Then the new candidates are generated around the new CoM which can be defined as,

$$P_{k+1} = CoM_{pos} + \frac{rand_k \times \omega \times (P_{\max} - P_{\min})}{k+1} \tag{7}$$

where $P_{k+1}$is the new candidates, $rand_k$ is the random number generated by the normal distribution, ω is limitation parameter of search space and $P_{\max}$ and $P_{\min}$ are upper and lower limits respectively. Then the steps of GA and BC phases will be repeated until the specified termination condition is met.

3.1.4. Diagnosis of COPD using fuzzy-based quantum neural network classification algorithm

Algorithm 2: Fuzzy-Based Quantum Neural Network Classification Algorithm

Input: Optimal features Set S, Target Value TarV

Output: Class Labels

Procedure:

1. Begin
2. Initialize $\partial = 0.01$, $\theta^l = 0$, $\delta\theta^l = 0.25$, nS=13
3. Perform pair-wise comparison among feature alternatives using triangular membership function
4. Compute the relative weight of optimal feature using eigenvector
   $CFW_i = [\prod_{i=1}^{nS} r_{ij}]^{1/nS}$
5. Standardize the Eigenvector Solution
   $nEV = CFW_i / \sum_{i=1}^{nS} CFW_i$
6. Calculate the net weight
   $NetW_j = \sum_{i}^{nS} r_i \times GloW_{ij} + b$
7. Calculate the value of linear activation for the hidden layer
   $nt_j = \sum LW_{ij} NetW_j$
8. Compute the output of the hidden layer using the simple sigmoid function
   $O_j = \sum_{l=1}^{NetW_j} (\frac{1}{1+\exp(-nt_j + \theta^l)})$
9. Calculate the value of linear activation for the output layer
   $nt_k = \sum LW_{jk} O_j$
10. Compute the output of the output layer by using the logarithmic sigmoid function
    $O_k = (\frac{1}{1+\exp(nt_k)})$
11. Estimate the error rate of the output layer
    $\Delta_k = (TarV_k - O_k) - O_k(1 - O_k)$
12. If $\Delta_k <$ $TarV_k$
13. Else If
14. Else
15. Update the quantum interval
    $\theta^l = \theta^l + \delta\theta$
16. End If
17. Update the output layer weight
    $LW_{kj}^{New} = LW_{kj}^{Old} + \partial \Delta_k O_j$
18. Estimate the error rate of the hidden layer
    $\Delta_j = O_j(1 - O_j) \sum_k LW_{kj}^{New} \Delta_k$
19. Update the hidden layer weight
    $LW_{ij}^{New} = LW_{ij}^{Old} + \partial \Delta_j O_i$
20. End

To diagnose COPD, the F-QNN classification algorithm is proposed (given in Algorithm 2), where the optimal features are fed into the FAHP technique for determining the relative weights of features and feature alternatives, and then the weights are fed into a quantum neural network for the diagnosis of COPD. The proposed model is explained in this section and the workflow of the proposed classification algorithm is depicted in Fig. 3.

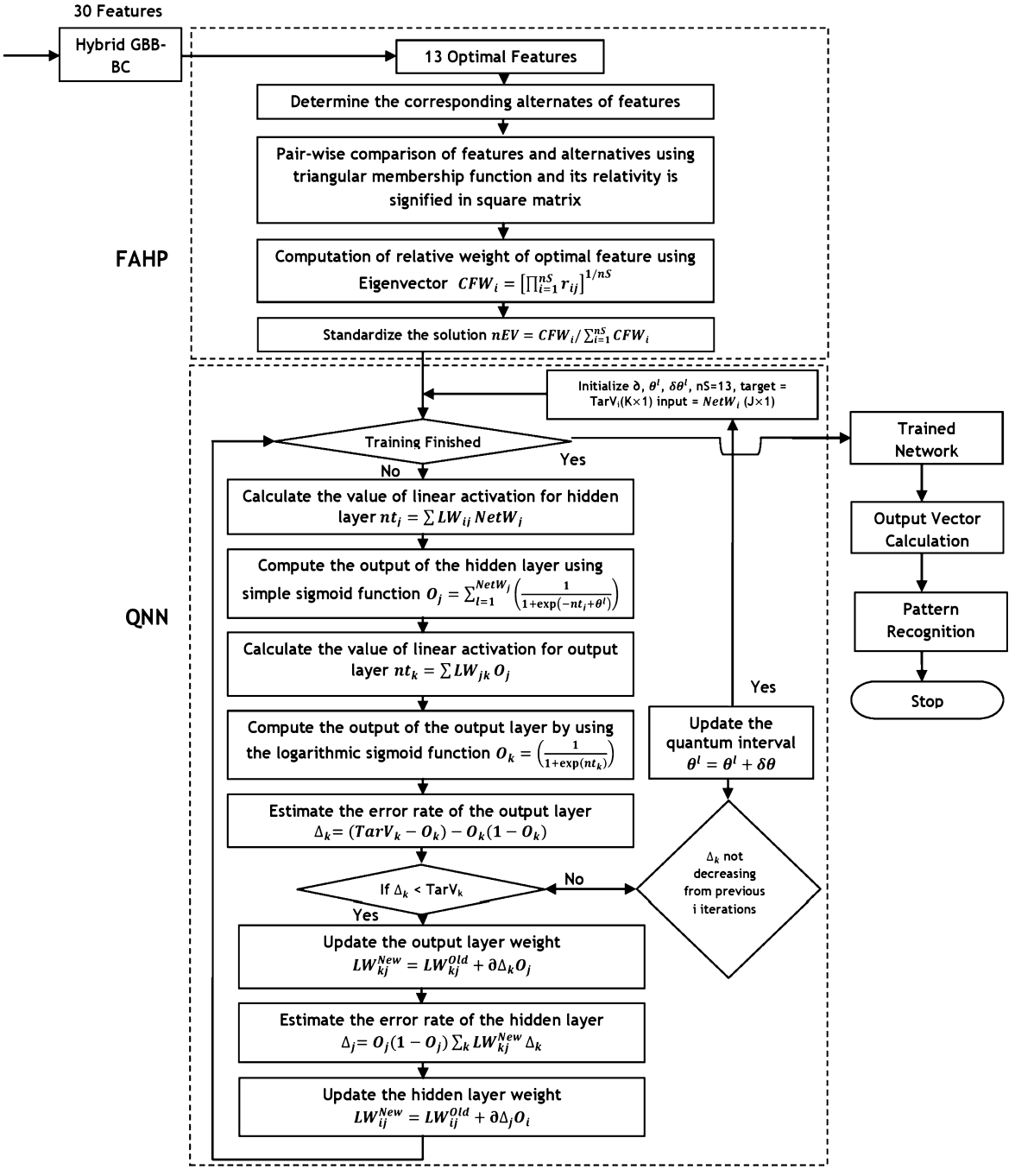

Fig. 3. Flowchart of fuzzy-based QNN classification algorithm for COPD diagnosis.

The problem of diagnosing COPD is expressed using a tree-like structure that illustrates the correlation between the aim of disease diagnosis, optimal features, and related alternatives. The problem is discriminated into aim, its optimal features, and respective feature alternatives. The pair-wise comparison among feature alternatives is made with the assistance of the fuzzy triangular membership function. The resultant value may mislead when the uncertainty of manual decision-making is not considered. The pair-wise comparison is made for every optimal feature with the other features and articulated as semantic judgments, which are transformed to numerical values using the satty fundamental preference scale.

The significance of the comparison of a particular optimal feature s_i to another optimal feature s_j is signified as a square matrix, as illustrated below:

Table 2
Optimal features global weights and feature alternatives local weights.

No	Optimal features	Optimal features global weight	Feature alternatives local weight			
			1	2	3	4
1	Age	0.1675	0.0169	0.0365	0.0478	0.0663
2	Body Mass Index	0.1016	0.0102	0.0203	0.0305	0.0406
3	Status of Smoking	0.1511	0.0231	0.0496	0.0784	–
4	History of Continuous Cough	0.1052	0.0384	0.0668	-	–
5	Neutrophils Value	0.0632	0.0107	0.0209	0.0316	–
6	Eosinophils Value	0.0738	0.0153	0.0238	0.0347	–
7	History of Chronic Lung Disease	0.1260	0.0504	0.0756	-	–
8	Type of Chest Pain	0.2096	0.0246	0.0483	0.0611	0.0756
9	NH_3 in Exhaled Breath	0.0246	0.0078	0.0168	–	–
10	C_3H_6O in Exhaled Breath	0.0985	0.0374	0.0611	-	–
11	H_2 in Exhaled Breath	0.0719	0.0246	0.0473	–	–
12	NO in Exhaled Breath	0.0753	0.0251	0.0502	–	–
13	CO in Exhaled Breath	0.0513	0.0201	0.0312	–	–

$$R = [r_{ij}] = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{ij} \\ 1/r_{12} & 1 & r_{23} & \dots & r_{2j} \\ 1/r_{13} & 1/r_{23} & 1 & \dots & r_{3j} \\ \dots & \dots & \dots & 1 & \dots \\ 1/r_{ij} & 1/r_{2j} & 1/r_{3j} & \dots & 1 \end{bmatrix} \tag{8}$$

where r_{ij} specifies the relative significance of the element's square matrix elements. The assessment of r_{ij} is done based on the condition given below:

$$\begin{cases} \text{if } r_{ij} = r, \text{ then } r_{i,j} = \frac{1}{r}, & r \neq 0; \\ \text{if } s_i \text{is equal to } s_j \text{ then } r_{ij} = 1, & r_{i,j} = 1, \ \forall i \end{cases} \tag{9}$$

The eigenvector is used for evaluating the relative weights of each optimal feature with respect to its feature alternatives transversely across the hierarchical levels.

$$CFW_i = \left[\prod_{i=1}^{nS} r_{ij} \right]^{1/nS} \tag{10}$$

where CFW_i is an optimal feature (i^{th}) weight and nS is the size or count of optimal features; therefore, nS=13.

The eigenvector solution is standardized based on the ratio of every optimal feature with respect to the summation of the entire optimal features is represented below.

$$nEV = CFW_i / \sum_{i=1}^{nS} CFW_i \tag{11}$$

where nEV is the standardized eigenvector and the process of standardization is recursively estimated until the divergence between the resultant value of current and previous computation is less than 0.0001. Every optimal feature weight in predicting COPD is computed as given in Table 2.

The local weights of feature alternatives are specified in a form of a matrix in which every column value is multiplied with the $GloW_{ij}$ (global weight) of corresponding optimal feature and the resulting values are summited with an appropriate bias 'b' as shown below.

$$NetW_j = \sum_{i}^{nS} r_i \times GloW_{ij} + b \tag{12}$$

where r_i is the incoming value, $GloW_{ij}$ is the weights of the link among the nodes i.e. global weight of every optimal feature.

The local weights of feature alternatives and the global weights of optimal features are passed to the subsequent phase of the classifier, namely, the quantum neural network. The subsequent level of the classification model is composed of hidden, input, and output layers. The QNN is equipped with only one hidden layer with three sub-states with varied quantum interval θ^l of the quantum level 'l'.

The non-linear activation function which comprises of superposition of the multi-sigmoid function is employed at the hidden layer of QNN. Hence, more states can be represented in a hidden layer whereas traditional sigmoid function can able to express only two states.

Let us consider that ∂ represents the learning rate of the network, Δ_k denotes the error rate of the output layer, Δ_j denotes the rate of the hidden layer, O_j is the output of the hidden layer, O_k is the output of the output layer, correspondingly. LW_{ij} is the link weight between input and hidden layers and LW_{kj} is the link weight between hidden and output layers.

Initially, weights are assigned with small random numbers and *TarV* represents the target value of the training network. The training pair T is $\{NetW_1, \mathrm{TarV}_1; NetW_2, \mathrm{TarV}_2; NetW_T, \mathrm{TarV}_\mathrm{T}\}$; where $NetW_i(\mathrm{J}\times 1)$ is the input and $\mathrm{TarV_i}$ $(\mathrm{K}\times 1)$ is the target value for the given inputs.

The error rate of the output (Δ_k) and hidden layers (Δ_j) are expressed as,

$$\Delta_k = (TarV_k - O_k) - O_k(1 - O_k) \quad \{k = 1, 2, 3, \ldots..K\} \tag{13}$$

$$\Delta_j = O_j(1 - O_j)\sum_k LW_{kj}^{New}\Delta_k \quad \{j = 1, 2, 3, \ldots J\} \text{ and } \{k = 1, 2, 3, \ldots K\} \tag{14}$$

Accordingly, the weights of output layer (LW_{kj}^{New}) and hidden layers (LW_{ij}^{New}) are updated using the following equation,

$$LW_{kj}^{New} = LW_{kj}^{Old} + \partial\Delta_k O_j \tag{15}$$

$$LW_{ij}^{New} = LW_{ij}^{Old} + \partial\Delta_j O_i \tag{16}$$

From the recent literature, the sigmoid function is considered as a predictability function and outperforms prediction-oriented problems. Hence, the sigmoid activation function has been employed for determining the output of the hidden layer and output layer, where different graded levels have been used for every hidden neuron between the hidden and output layer. The linear activation function and output estimation at a hidden layer can be expressed as,

$$nt_j = \sum LW_{ij}NetW_j \tag{17}$$

$$O_j = \sum_{l=1}^{NetW_j}\left(\frac{1}{1 + \exp(-nt_j + \theta^l)}\right) \tag{18}$$

where nt_j is the linear activation function for the hidden layer, $NetW_j$ is the (input) weight of optimal feature, O_j is the output of the hidden layer and LW_{ij} is the weight among the input and hidden layers.

The linear activation function and the output of the output layer is obtained by the applying logarithmic sigmoid function can be expressed as,

$$nt_k = \sum LW_{jk}O_j \tag{19}$$

$$O_k = \left(\frac{1}{1 + \exp(nt_k)}\right) \tag{20}$$

where nt_k is the linear activation function at the output layer and LW_{jk} is the weight among the hidden and output layer, and O_k is the output of the output layer.

The quantum interval will be increased by a small interval $\delta\theta^l$, only when the error rate is not less than the target value and when there is no change in error rate from the previous iteration, the quantum interval updation can be expressed as,

$$\theta^l = \theta^l + \delta\theta^l \tag{21}$$

If the gained error rate is less than the target value, the weights of the hidden and output layer are updated and then the training of the quantum network is terminated.

4. Results and discussion

This section deals with the discussion on the outcomes of optimal feature selection algorithms and classification models from different perspectives.

4.1. Performance analysis of feature selection algorithms

The real-time dataset was used in this experiment for determining the suitable machine learning models for the diagnosis of COPD. The dataset has a size of 300 records, 30 features, and the target class label has two classes that represent the healthy subjects and COPD subjects. The data size was constantly increased while executing the algorithms and the univariate ANOVA is utilized to analyze the performance of feature selection algorithms based on the three values such as average execution time, the number of reduced features, and the number of generations.

Table 3
Average execution time vs data sizes.

	Algorithm	Data size						
		20	50	100	150	200	250	300
Avg.	**GBB-BC**	0.00055	0.001147	0.08434	0.12705	0.169965	0.21139	0.254305
Exe.	**GA**	0.001437	0.005089	0.096684	1.132612	1.278256	2.410868	2.65349
Time	**PSO**	0.001254	0.004954	0.094534	0.14234	1.25437	1.39756	2.63579
(Mins)	**BBA**	0.127232	1.23534	1.32546	1.49565	2.628262	2.798227	3.052532

Table 4
Summary of ANOVA for average execution time.

Model	Degree of freedom	Sum of squares	Mean square	F ratio	Significance of F (P-value)
Regression	4.000	64258.522	16064.631	380.973	0.003
Residual	2.000	84.335	42.167		
Total	6.000	64342.857			

Table 5
Summary of correlation coefficients with respect to average execution time.

Algorithm	Coefficients	Standard error	t Stat	Sign (P-value)
(Constant)	20.546	6.595	3.116	0.089
GBB-BC	576.474	110.922	5.197	0.035
GA	18.243	8.468	2.154	0.164
PSO	6.642	6.524	1.018	0.416
BBA	20.684	7.147	2.894	0.102

Table 6
Summary of regression statistics for average execution time.

Model	Multiple R	R square	Adjusted R square	Standard error
Regression Statistics	0.999	0.999	0.996	6.494

4.1.1. Analysis of average execution time

From the experimental results, it has been observed that the GBB-BC algorithm determines the optimal features with minimum execution time when compared to the commonly used meta-heuristic algorithm [45]. In the literature, it has been shown that the traditional metaheuristic feature selection techniques perform well in terms of the number of optimal features selected [45]. The significant difference between these algorithms is evidenced concerning average execution time through statistical analysis is shown in Table 3 and Fig. 4a.

The test of formulated hypothetical statements becomes:

Hypothesis H_0 – No significant difference between GA, PSO, BBA, and GBB-BC algorithms with respect to average execution time for optimal feature selection. ($\mathbf{H_0}$: $ET_1 = ET_2 = ET_3 = ET_4$)

Hypothesis H_1 – Significant difference between GA, PSO, BBA, and GBB-BC algorithms with respect to average execution time for optimal feature selection. ($\mathbf{H_1}$: $ET_1 \neq ET_2 \neq ET_3 \neq ET_4$)

As can be seen in Table 4, the average execution times (independent variable) of the algorithms are not all the same with respect to the data sizes (dependent variable); the p-value concludes that the regression model is a good fit for the data. Thus, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted.

Table 5 depicts that the algorithms are positively correlated as average execution time and data size are moving in the same direction. From the regression statistics summary shown in Table 6, the observed value of 'R' is 0.999 denotes the goodness of fit, and $R^2 = 128517.045/128685.714 = 0.999$ which denotes that 99% of the variance in average execution time can be determined by this model. Then based on the regression statistics summary, a significant difference between GA, PSO, BBA, and GBB-BC algorithms is identified concerning average execution time.

4.1.2. Analysis of the number of generations

The GBB-BC algorithm identifies the optimal features with the minimum number of generations (iterations), in contrast with the GA, PSO, and BBA algorithms. The total number of generations (iterations) needed for identifying the optimal features by these algorithms on various sizes of data is illustrated in Table 7 and Fig. 4 b.

The test of formulated hypothetical statements becomes:

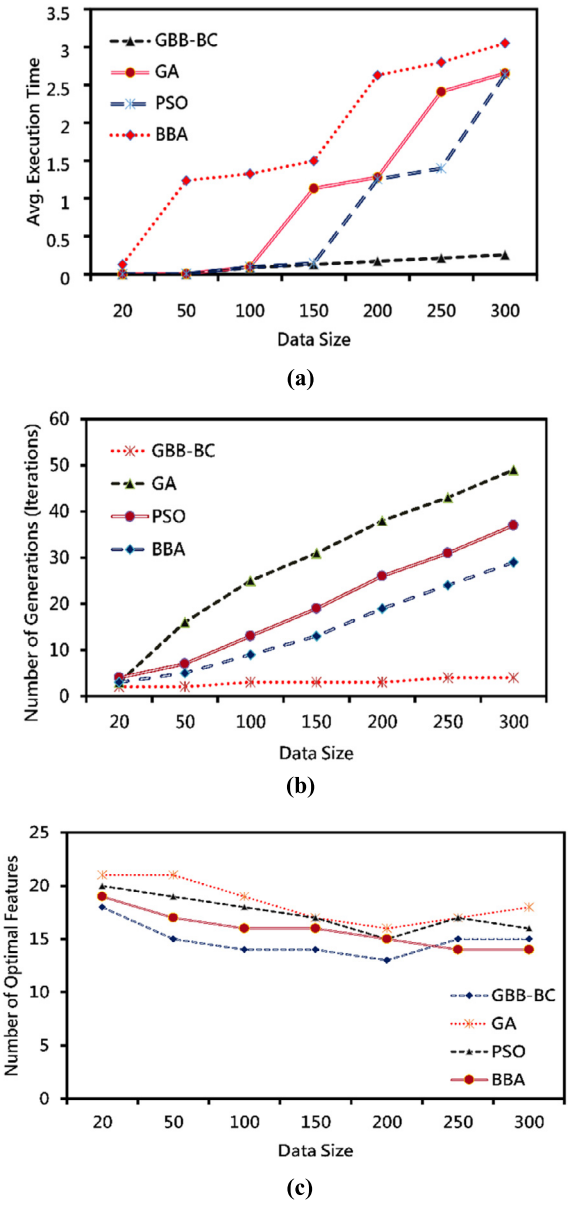

Fig. 4. Analysis of feature selection algorithms. (a) Average execution time vs data sizes. (b) Number of generations vs data sizes. (c) Number of optimal feature vs data sizes.

Hypothesis H_0 – No significant difference between GA, PSO, BBA, and GBB-BC algorithms with respect to the number of generations (iterations) needed for optimal feature selection. ($\mathbf{H_0}$: $nG_1 = nG_2 = nG_3 = nG_4$)

Hypothesis H_1 – Significant difference between GA, PSO, BBA, and GBB-BC algorithms with respect to the number of generations (iterations) needed for optimal feature selection. ($\mathbf{H_1}$: $nG_1 \neq nG_2 \neq nG_3 \neq nG_4$)

According to Table 8, the number of generations (independent variable) taken by these algorithms for producing optimal features is significantly different with respect to the data sizes (dependent variable); the p-value concludes that the regression model is a good fit for the data. Thus, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted.

The summary of the correlation coefficient shown in Table 9 indicates that the number of generations needed to produce the optimal features is positively correlated with the changing rate of data size. From the regression statistics summary shown in Table 10, the observed value of 'R' is 0.999 which denotes the goodness of fit, and $R^2 = 64317.068/64342.857 = 0.999$ which denotes that 99.9% of the variance in the number of generations has been determined by this model. Then

Table 7
Number of generations vs data sizes.

	Algorithm	Data size						
		20	50	100	150	200	250	300
No. of generations (iterations)	**GBB-BC**	2	2	3	3	3	4	4
	GA	3	16	25	31	38	43	49
	PSO	4	7	13	19	26	31	37
	BBA	3	5	9	13	19	24	29

Table 8
Summary of ANOVA for the number of generations needed to produce optimal features.

Model	Degree of Freedom	Sum of squares	Mean square	F ratio	Significance of F (P-value)
Regression	4	64317.068	16079.267	1246.963	0.001
Residual	2	25.789	12.895		
Total	6	64342.857			

Table 9
Summary of the correlation coefficient for the number of generations needed to produce optimal features.

Algorithm	Coefficients	Standard error	t Stat	Sign (P-value)
Intercept	−21.812	10.093	−2.161	0.163
GBB-BC	5.870	5.446	1.078	0.394
GA	0.618	0.561	1.101	0.386
PSO	5.153	2.681	1.922	0.195
BBA	2.615	2.791	0.937	0.448

Table 10
Summary of regression statistics for the number of generations needed to produce optimal features.

Model	Multiple R	R square	Adjusted R square	Standard error
Regression Statistics	0.999	0.999	0.999	3.591

Table 11
Number of optimal feature vs data sizes.

	Algorithm	Data size						
		20	50	100	150	200	250	300
No. of optimal features	**GBB-BC**	18	15	14	14	13	15	15
	GA	21	21	19	17	16	17	18
	PSO	27	27	24	22	20	18	18
	BBA	15	18	19	21	21	23	23

based on the regression statistics summary, the considerable difference between GA, PSO, BBA, and GBB-BC algorithms is identified with respect to the number of generations required for determining the optimal features.

4.1.3. Analysis of the number of optimal features

In specific, the number of features reduced by GA, PSO, and BBA algorithms is considerably greater than the features reduced by the GBB-BC algorithm. When there is an increase in data size, remarkable changes in the number of optimal features for diagnosing the presence of the disease are experimentally observed as illustrated in Table 11 and Fig. 4 c.

In this research work, the feature selection algorithms like GA, PSO, and BBA are compared with the GBB-BC algorithm, and from the above three analyses, it is concluded that the GBB-BC algorithm outperforms the other algorithms.

The test of formulated hypothetical statements becomes:

Hypothesis H_0 – No significant difference between GA, PSO, BBA, and GBB-BC algorithms in terms of feature reduction. ($\mathbf{H_0}$: $nOF_1 = nOF_2 = nOF_3 = nOF_4$)

Hypothesis H_1 – Significant difference between GA, PSO, BBA, and GBB-BC algorithms in terms of feature reduction. ($\mathbf{H_1}$: $nOF_1 \neq nOF_2 \neq nOF_3 \neq nOF_4$)

According to Table 12, a significant difference is observed between GA, PSO, BBA, and GBB-BC algorithms in terms of the number of reduced optimal features; the p-value concludes that the regression model is a good fit for the data. Thus, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted.

Table 12
Summary of ANOVA for the reduction in the number of optimal features.

Model	Degree of freedom	Sum of squares	Mean square	F ratio	Significance of F
Regression	4	63593.279	15898.320	42.419	0.023
Residual	2	749.579	374.789		
Total	6	64342.857			

Table 13
Summary of the correlation coefficient for the reduction in the number of optimal features.

Algorithm	Coefficients	Standard error	t Stat	Sign (P-value)
(Constant)	550.965	480.108	1.148	0.370
GBB-BC	−6.851	11.651	−0.588	0.616
GA	18.819	12.365	1.522	0.267
PSO	−31.321	10.905	−2.872	0.103
BBA	2.745	12.362	0.222	0.845

Table 14
Summary of regression statistics for the reduction in the number of optimal features.

Model	Multiple R	R square	Adjusted R square	Standard error
Regression Statistics	0.994	0.988	0.988	19.359

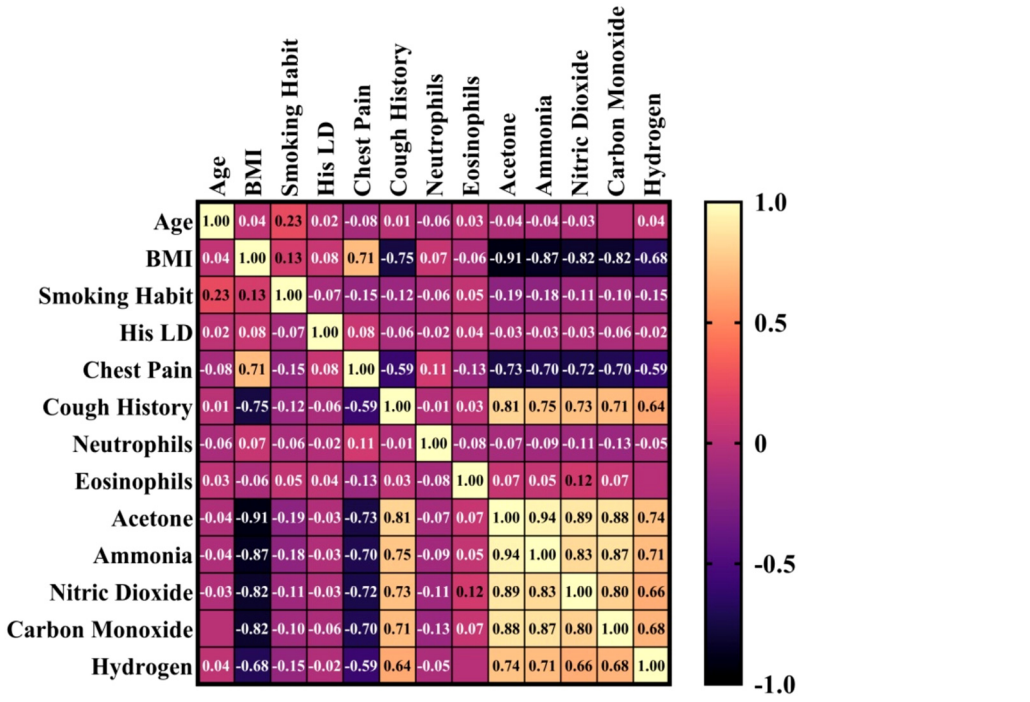

Fig. 5. Heatmap derived using pearson coefficient of selected optimal features after applying GBB-BC on real-time dataset derived. The heatmap scale corresponds to the negatively and positively correlated optimal features.

Therefore, the summary of the correlation coefficient (shown in Table 13) signifies the positive and negative values correspondingly, in which data size is increasing constantly and the numbers of optimal features are decreasing at a nominal rate except the BBA algorithm. According to the regression statistics summary presented in Table 14, the obtained value of 'R' is 0.994 which denotes the goodness of fit, and $R^2 = 63593.279/64342.857 = 0.988$ represents 98.8% of the variance in the feature reduction has been determined by this model. Therefore, the model proves that there is a considerable difference between the algorithms in terms of reduction in the number of optimal features.

In sum, a strong significant difference has been observed between the existing metaheuristic feature selection algorithms and hybrid GBB-BC algorithm based on three different independent variables such as average execution time, number of generations, and number of optimal features. The heatmap presented in Fig. 5 shows the Pearson correlation coefficient

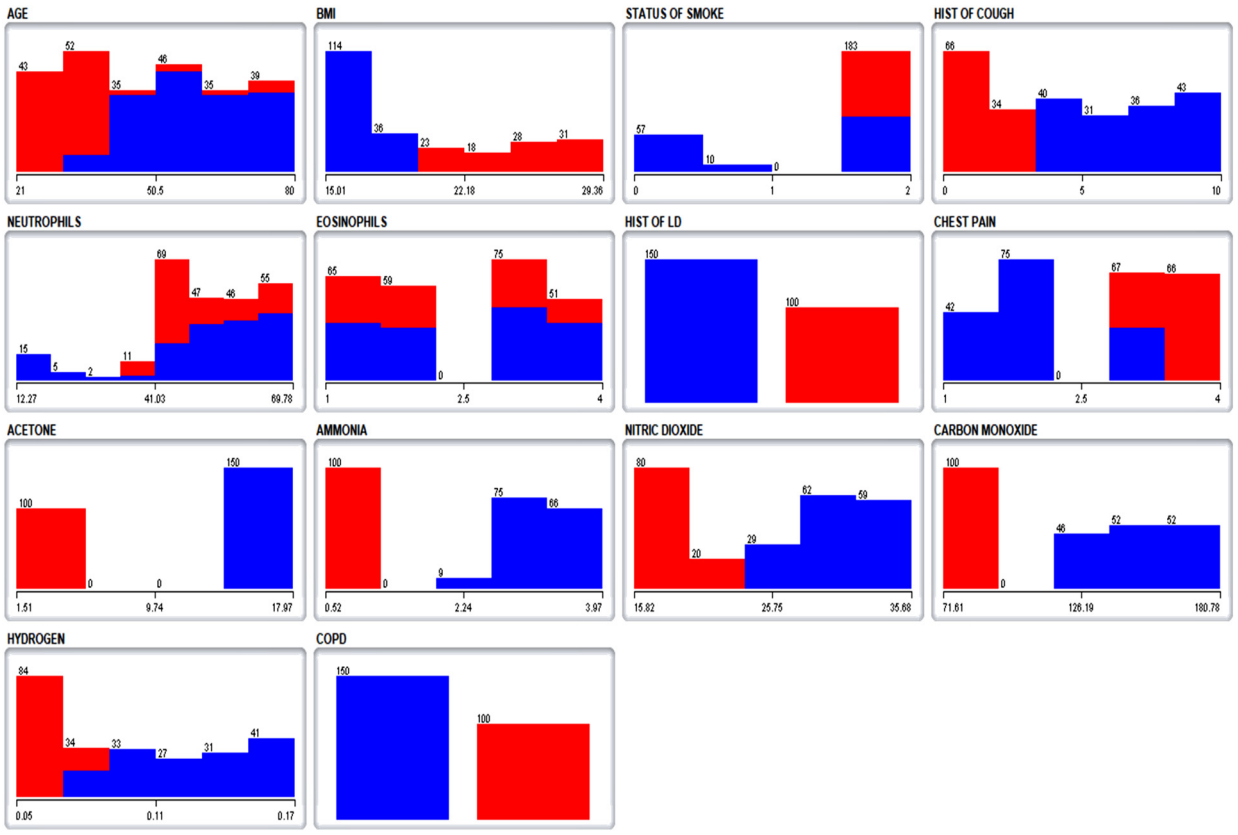

Fig. 6. Crosstab visualization illustrates the distribution of the dependent variable upon each independent variable of the real-time dataset. Blue tabs indicate COPD subjects and red tabs indicate the healthy subjects. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 15
Confusion matrix.

	Diagnosed COPD patients	Diagnosed healthy subjects
Actual COPD patients	TP	FN
Actual healthy subjects	FP	TN

of 13 optimal features obtained by GBB-BC. From the heatmap, it has been observed that the VOC patterns tend to have significantly strong evidence, which facilitates the accurate prediction of COPD through VOC analysis. The cough history has a strong positive correlation with the optimal VOC features whereas BMI and chest pain has a strong negative correlation with optimal VOC features. Crosstab visualization is shown in Fig. 6, which illustrates the distribution of the dependent variable upon each independent variable of the real-time dataset. From the crosstab analysis, it has been observed that COPD subjects exhibit the symptom of continuous cough, history of lung disease, and typical chest pain. Further, the optimal VOC features {acetone, ammonia, nitric oxide, carbon monoxide, and hydrogen} are observed to be abnormal in COPD subjects.

4.2. Performance analysis of machine learning classifiers

The synthetic training dataset is given into the machine learning classifiers which include logistic regression (LR), k-nearest neighbors (k-NN), support vector machine (SVM), Naïve Bayes (NB), random forest (RF), artificial neural network (ANN) and proposed F-QNN approaches. These machine learning classifiers are trained and tested with respect to different applicable tuning parameters. In the section, the performance of classifiers was analyzed based on five performance evaluation metrics and the performance of the classifier is evaluated using 10-fold cross-validation. The confusion matrix determines the ability of classifiers for the accurate diagnosis of COPD disease, which is illustrated in Table 15.

Based on the confusion matrix, the following parameters were computed: True Positive (TP) - The total number of records labeled as the presence of COPD while they are COPD subjects. True Negative (TN) - The total number of records labeled as the absence of COPD while they are healthy subjects. False Positive (FP) – The total number of records labeled as the presence of COPD while they are healthy subjects. False Negative (FN) - The total number of records labeled as the

Table 16
Classifiers performance evaluation on optimal features.

Classifiers	Tuning hyperparameters	Performance evaluation metrics of classifiers					
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Processing time (s)	MCC	AUC
LR	$C = 1$	74	66	82	2.313	74	74
	$C = 10$	75	67	82	2.352	74	75
	$C = 100$	78	67	88	2.159	78	79
k-NN	K=1	83	75	90	8.001	84	84
	K=3	85	74	94	8.046	84	85
	K=7	84	72	94	8.399	85	83
	K=9	81	73	88	8.875	84	80
SVM Kernel=RBF	C=100, g=0.0001	88	75	96	6.019	85	84
	C=10, g=0.001	85	70	94	5.009	84	84
SVM Kernel=Linear	C=100, g=0.0001	84	74	96	7.023	85	84
	C=10, g=0.001	82	75	96	7.005	84	84
Random Forest	ntree=50	83	70	90	4.774	83	83
	ntree=100	84	73	92	4.606	83	84
ANN	Hidden Neurons- 16	86	77	94	7.430	85	85
	Hidden Neurons- 20	82	70	94	7.750	82	81
	Hidden Neurons- 40	71	38	88	8.400	69	69
F-QNN	$\partial = 0.01$, $\delta\theta^l = 0.25$	95	95	97	1.784	89	89
	$\partial = 0.001$, $\delta\theta^l = 0.25$	96	95	98	1.742	92	91

absence of COPD while they are COPD subjects. The various performance metrics are calculated using the confusion matrix and the performance metric involved in this research work is described as follows: Accuracy: The accuracy is the number of testing samples for which the presence of COPD or absence of COPD is accurately diagnosed and represented mathematically using the elements of confusion matrix as shown in Equation (22). Sensitivity: The proportion of positive records that are appropriately recognized to all positive records is the sensitivity or recall, which is also termed as a true positive rate and represented mathematically using the elements of the confusion matrix as shown in Equation (23). Specificity: The proportion of negative records that are appropriately recognized to all negative records is known to be specificity (true negative rate) and represented mathematically using the elements of the confusion matrix as shown in Equation (24). Processing Time: The amount of time taken by the classifier for diagnosing COPD is known as processing time. ROC and AUC: The receiver optimistic curve (ROC) is a graphical representation that analyzes the prediction ability of the machine learning classifiers by comparing the true positive rate and false-positive rate. The area under the curve (AUC) defines the characterization of classifiers ROC and the effectiveness of the classifier is denoted by the highest value of AUC.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{No. of Patients}} \tag{22}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{23}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{24}$$

In this experiment, the IMLFF framework tends to learn from the optimal features selected by GBB-BC using different classification models with various turning parameters, and the results were averagely computed based on 10-fold cross-validation. In return, the best possible model is identified by evaluating performance metrics. The performance of machine learning classifiers with the optimal features is evaluated with various metrics as depicted in Table 16.

From the result analysis, the LR classifier achieves as {78%, 67%, 88%, 78%, 79%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} with the value of tuning hyperparameters C is100. The processing time of LR classifiers is minimum with the value of C=100 when compared to other values of tuning hyperparameters. The performance of LR can also be perceived from Fig. 7a and the variations in the performance under various hyperparameters can be observed from Fig. 9a under tuning of hyperparameters. The performance of the k-NN classifier is good {85%, 74%, 94%, 84%, 85%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} with the hyperparameter k=3 but the processing time is not good at k=3. The consistent performance of the k-NN classifier is visualized in Fig. 7b and the minor deviation in the performance under various hyperparameters, which can be noted from Fig. 9b.

The classifier SVM (Kernel=RBF) shows better performance as {88%, 75%, 96%, 85%, 84%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} with the hyperparameter of C=100, g=0.0001. The performance of (Kernel=RBF) can also be perceived from Fig. 7c and the disparities in the performance can be observed from Fig. 9c under tuning of hyperparameters. The

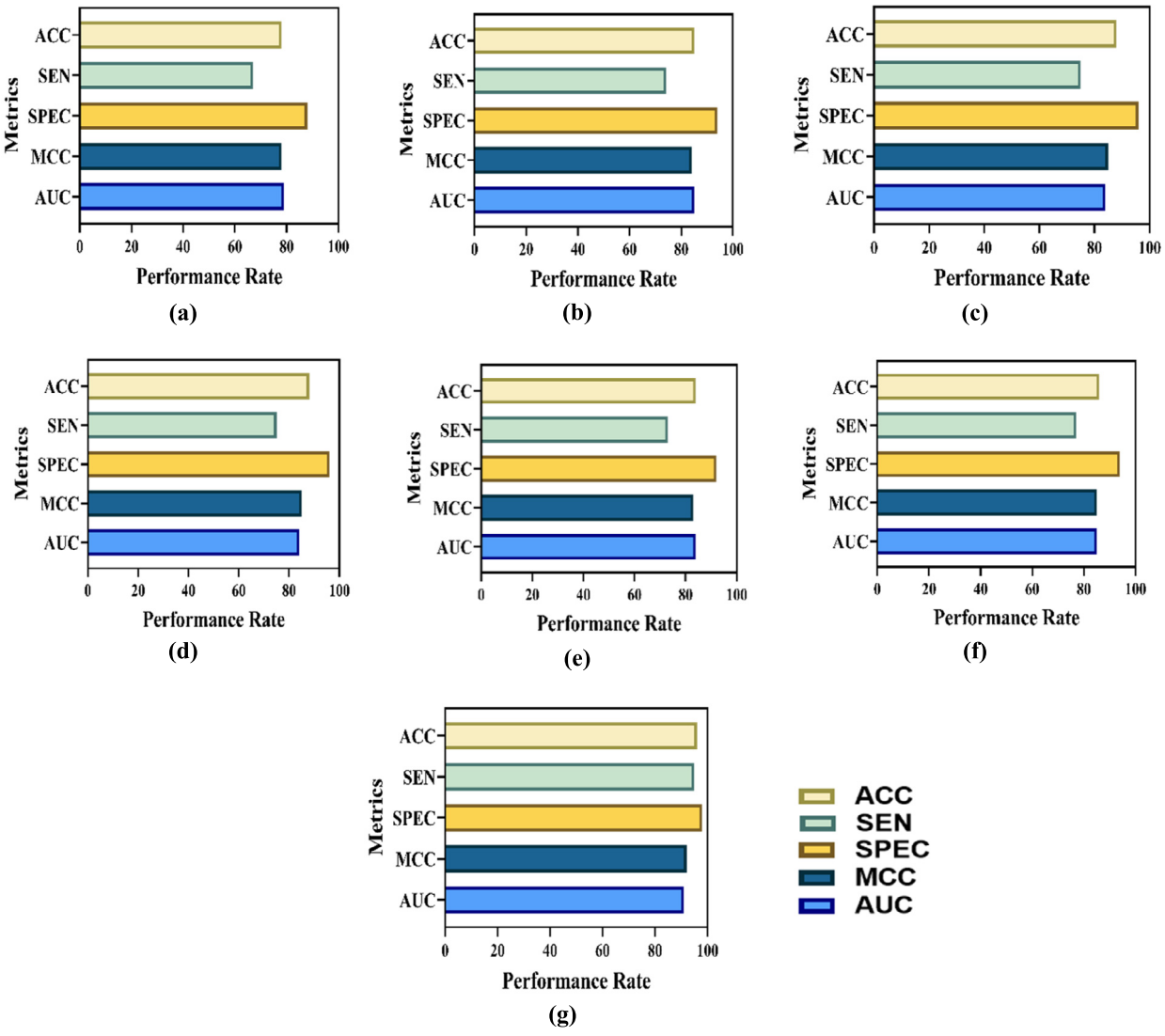

Fig. 7. Performance of corresponding classifiers with respect to the accuracy, sensitivity, specificity, MCC, and AUC under certain tuning hyperparameters yields best results. (a) LR, (b) k-NN, (c) SVM-RBF, (d) SVM-Linear, (e) RF, (f) ANN (g) F-QNN.

performance of SVM (Kernel=Linear) classifier achieves best results as {84%, 74%, 96%, 85%, 84%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} with the hyperparameter of C=100, g=0.0001. The processing times of SVM (RBF) and SVM (Linear) at C=100, g=0.0001 is greater than the processing times at C=10, g=0.001, and SVM (RBF) shows a better accuracy rate than the SVM (Linear). The performance of SVM (Kernel=Linear) can also be visualized in Fig. 7d and the slight deviation in the performance under various hyperparameters can be noted from Fig. 9d.

The RF classifier shows the better performance as {84%, 73%, 92%, 83%, 84%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} with the hyperparameter of tree size 100. The significant performance of the RF classifier is visualized in Fig. 7e and there is no major deviation in the performance under various hyperparameters, which can be noted from Fig. 9e. The performance of ANN classifier with 16 hidden neurons provides the best results as {86%, 77%, 94%, 85%, 85%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} than the ANN classifier with 20 and 40 hidden neurons. The performance of ANN can also be perceived from Fig. 7f and the high-pitched variations in the performance under various hyperparameters can be observed from Fig. 9f under tuning of hyperparameters.

The F-QNN classifier achieves {96%, 95%, 98%, 92%, 91%} for {Accuracy, Sensitivity, Specificity, MCC, AUC} and requires lesser processing time at learning rate $\partial = 0.001$. Although, the F-QNN classifier is substantially consistent than other classifiers which are observed from Fig. 7g, and the minor deviation in the performance under different learning rates, which can be visualized from Fig. 9g. The classifier's performance in terms of minimum processing time under certain hyperparameters is shown in Fig. 8a and the deviation in the processing time of classifiers under various hyperparameters can be visualized in Fig. 8b. In sum, by using the thirteen optimal features endowed by GBB-BC, the F-QNN classifier achieved an accuracy

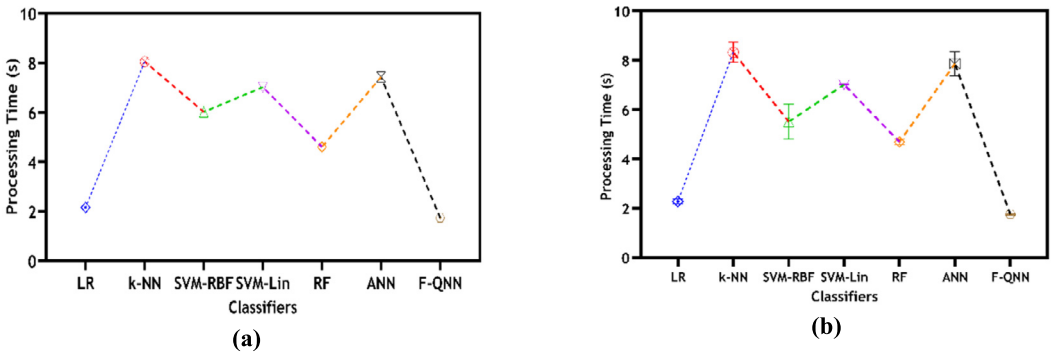

Fig. 8. Performance analysis of classifiers in terms of processing time under certain tuning hyperparameters yields best results. (a) Minimum processing time of classifiers under certain hyperparameter. (b) Deviation in the processing time of classifiers under various hyperparameters.

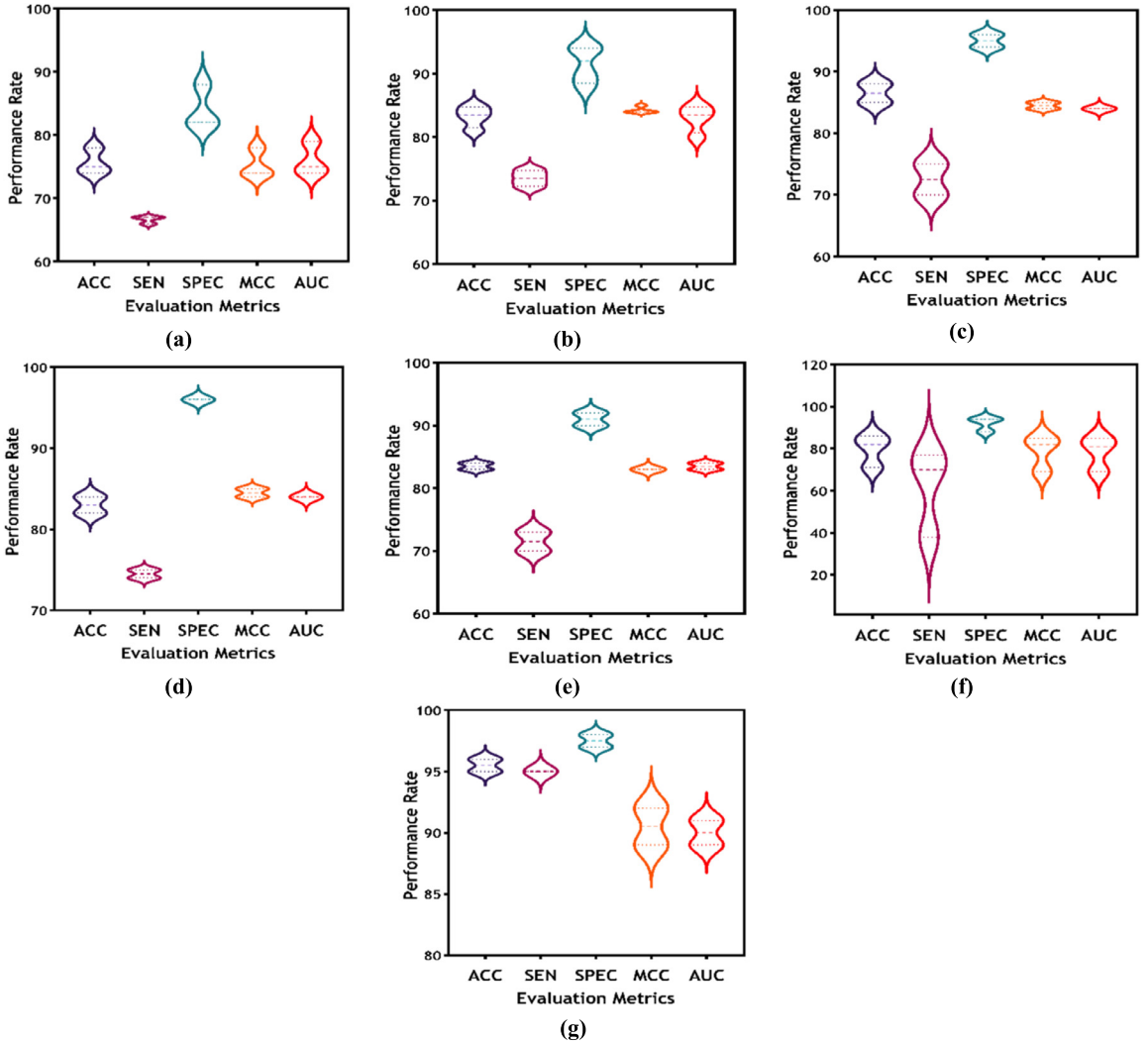

Fig. 9. Violin plot illustrates the distribution of the corresponding classifier. (a) LR, (b) k-NN, (c) SVM-RBF, (d) SVM-Linear, (e) RF, (f) ANN, (g) F-QNN, where the violin plot depicts the maximum and minimum value with respect to the accuracy, sensitivity, specificity, MCC and AUC under various tuning hyperparameters.

rate of 96% than the other classifiers. The processing time for the diagnosis of COPD using the F-QNN classifier is lesser than the processing time required by other classifiers. Therefore, it can be concluded that the F-QNN model outperforms on using the optimal features selected by GBB-BC for the diagnosis of COPD.

5. Conclusion

In this research work, an IoT-Spiro System has been developed and a complete machine learning framework has been designed for the diagnosis of COPD. The framework is evaluated on real-time data in which VOC patterns have been acquired via the IoT-Spiro system and some additional medical factors are also considered. Among the algorithms adapted to IMLFF to carry out the feature selection operation, hybrid GBB-BC performs better by identifying thirteen optimal features that are highly prominent to discriminate COPD subjects from healthy subjects. From the experimental and statistical analysis, the results depict that the hybrid GBB-BC outperforms the proven existing feature selection algorithms with respect to execution time, the number of generations needed for selecting the optimal features, and the number of reduced features. The significant difference between the optimal feature selection techniques has been proved and verified statistically.

Besides, the optimal features have been given as input to the well-known existing classifiers and a proposed F-QNN classifier for diagnosing COPD. The investigational results illustrate that the time required for processing is considerably reduced while using the optimal features selected by the GBB-BC algorithm for diagnosing COPD. The prediction accuracy of the F-QNN classifier is comparatively better than the other classifiers and the AUC rate proved the effectiveness of the proposed classifier. Hence, from the statistical and mathematical analysis, it is concluded that the IMLFF offers a promising diagnosis of COPD. Based on the analysis, the study confirmed that five VOCs (ammonia, acetone, hydrogen, nitric oxide, and carbon monoxide) in exhaled breath are significantly supporting the diagnosis of COPD and also the effectiveness of the developed IoT-Spiro system is proven. In the future perspective, studies will be conducted by analyzing the various VOC patterns that are related to other diseases using the IoT-Spiro System, and IMLFF will be tested against the diagnosis of other diseases. Also, fog computing concepts will be introduced and the irrelevant features will be removed at the edge of the cloud storage, and optimization techniques also be tested for improving prediction accuracy.

Declaration:

Ethics Approval and Consent to Participate:

Ethical approval The clinical data used in this research study was conducted retrospectively from data obtained for clinical purposes. In addition, exhaled breath samples were collected in line with the principles of the Declaration of Helsinki. Approval was granted by the Bharathiar University Human Ethical Committee (BUHEC) (Dated 30.01.2020/No.BUHEC-036).

Human and Animal Rights:
No violation of Human and Animal Rights is involved.

Funding: This work has been supported by the Department of Science and Technology—Interdisciplinary Cyber-Physical System (DST-ICPS), New Delhi, India, under Grant DST/ICPS/CPS-Individual/2018/193.

Authorship contributions:
There is no authorship contribution

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors thank the Department of Science and Technology—Interdisciplinary Cyber-Physical System (DST-ICPS), New Delhi, India, under Grant DST/ICPS/CPS-Individual/2018/193 for providing financial support during the period of this research work.

References

[1] D.M.G. Halpin, G.J. Criner, A. Papi, D. Singh, A. Anzueto, F.J. Martinez, A.A. Agusti, C.F. Vogelmeier, Global initiative for the diagnosis, management, and prevention of chronic obstructive lung disease. The 2020 GOLD science committee report on COVID-19 and chronic obstructive pulmonary disease, Am. J. Respir. Crit. Care Med. 203 (1) (2021 Jan 1) 24–36.
[2] L.P. Boulet, J.M. FitzGerald, M.L. Levy, A.A. Cruz, S. Pedersen, T. Haahtela, A guide to the translation of the global initiative for asthma (GINA) strategy into improved care, Eur. Respir. J. 39 (2012) 1220–1229.
[3] R. Schnabel, R. Fijten, A. Smolinska, Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia, Sci. Rep. 5 (2015) 17179.
[4] L. Xiang, S. Wu, Q. Hua, C. Bao, H. Liu, Volatile organic compounds in human exhaled breath to diagnose gastrointestinal cancer: a meta-analysis, Front. Oncol. 11 (2021) 606915.
[5] N. Muni Kumar, R. Manjula, Role of Big data analytics in rural health care – a step towards svasth bharath, 2014.

[6] T. Eswari, P. Sampath, S. Lavanya, Predictive methodology for diabetic data analysis in big data, Proc. Comput. Sci. 50 (2015) 203–208.
[7] S. D'Souza, K.V. Prema, S. Balaji, Feature selection and modeling using statistical and machine learning methods, in: 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2020, pp. 18–22.
[8] Z. Wang, Z. Lin, Optimal feature selection for learning-based algorithms for sentiment classification, Cogn. Comput. 12 (2020) 238–248.
[9] G. Yang, et al., A health-IoT platform based on the integration of intelligent packaging, unobtrusive bio-sensor, and intelligent medicine box, IEEE Trans. Ind. Inform. 10 (4) (2014) 2180–2191, https://doi.org/10.1109/TII.2014.2307795.
[10] Y. Yan, Q. Li, H. Li, X. Zhang, L. Wang, Open Access: a home-based health information acquisition system, Health Inf. Sci. Syst. 1 (2013) 12, https://doi.org/10.1186/2047-2501-1-12.
[11] F. Firouzi, et al., Internet-of-things and big data for smarter healthcare: from device to architecture, applications and analytics, Future Gener. Comput. Syst. 78 (2018) 583–586, https://doi.org/10.1016/j.future.2017.09.016.
[12] S. Muro, M. Ishida, Y. Horie, et al., Machine learning methods for the diagnosis of chronic obstructive pulmonary disease in healthy subjects: retrospective observational cohort study, JMIR Med. Inform. 9 (7) (2021) e24796, Published 2021 Jul 6.
[13] E.D. Bateman, S.S. Hurd, P.J. Barnes, J. Bousquet, J.M. Drazen, M. FitzGerald, P. Gibson, K. Ohta, P. O'Byrne, S.E. Pedersen, et al., Global strategy for asthma management and prevention: GINA executive summary, Eur. Respir. J. 31 (1) (2008) 143–178, https://doi.org/10.1183/09031936.00138707.
[14] M. Phillips, R.N. Cataneo, A.R. Cummin, A.J. Gagliardi, K. Gleeson, J. Greenberg, R.A. Maxfield, W.N. Rom, Detection of lung cancer with volatile markers in the breath, Chest 123 (6) (2013) 2115–2123, https://doi.org/10.1378/chest.123.6.2115.
[15] A. Schieweck, J. Gunschera, D. Varol, et al., Analytical procedure for the determination of very volatile organic compounds (C_3–C_6) in indoor air, Anal. Bioanal. Chem. 410 (2018) 3171–3183.
[16] Dina Hashoul, Hossam Haick, Sensors for detecting pulmonary diseases from exhaled breath, Eur. Respir. Rev. 28 (2019) 152.
[17] Pavlou Aktapf, Sniffing out the truth: clinical diagnosis using the electronic nose, Clin. Chem. Lab. Med. 38 (2) (2000) 99–112.
[18] M. Basanta, R.M. Jarvis, Y. Xu, G. Blackburn, R. Tal-Singer, A. Woodcock, D. Singh, R. Goodacre, C.L. Thomas, S.J. Fowler, Non-invasive metabolomic analysis of breath using differential mobility spectrometry in patients with chronic obstructive pulmonary disease and healthy smokers, Analyst 135 (2) (2010) 315–320, https://doi.org/10.1039/b916374c.
[19] G.S. Karthick, P.B. Pankajavalli, Effective usage of exhaled volatile organic compounds in disease diagnosis: a comprehensive review, Indian J. Public Health Res. Dev. 10 (4) (2019).
[20] J.J. Van Berkel, J.W. Dallinga, G.M. Moller, R.W. Godschalk, E.J. Moonen, E.F. Wouters, F.J. van Schooten, A profile of volatile organic compounds in breath discriminates COPD patients from controls, Respir. Med. 104 (4) (2020) 557–563, https://doi.org/10.1016/j.rmed.2009.10.018.
[21] N. Fens, S.B de Nijs, S. Peters, T. Dekker, H.H. Knobel, T.J. Vink, N.P. Willard, A.H. Zwinderman, F.H. Krouwels, H.G. Janssen, et al., Exhaled air molecular profiling in relation to inflammatory subtype and activity in COPD, Eur. Respir. J. 38 (6) (2011) 1301–1309, https://doi.org/10.1183/09031936.00032911.
[22] M. Shardlow, An Analysis of Feature Selection Techniques, The University of Manchester, 2016.
[23] I.H. Sarker, Machine learning: algorithms, real-world applications and research directions, SN Comput. Sci. 2 (2021) 160.
[24] M.A. Tawhid, A.M. Ibrahim, Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm, Int. J. Mach. Learn. Cybern. 11 (2020) 573–602.
[25] A. Shahrjooihaghighi, H. Frigui, Local feature selection for multiple instance learning, J. Intell. Inf. Syst. (2021).
[26] P.L. Braga, A.L.I. Oliveira, S.R.L. Meira, A GA-based feature selection and parameters optimization for support vector regression applied to software effort estimation, in: Proceedings of ACM-SAC, ACM, 2018, pp. 1788–1792.
[27] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recognit. 33 (2000) 25–41.
[28] N. Das, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, D.K. Basu, A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application, Appl. Soft Comput. 12 (5) (2012) 1592–1606.
[29] G.A. Ratta, J. Vega, A. Murari, JET-EFDA Contributors, Improved feature selection based on genetic algorithms for real time disruption prediction on JET, Fusion Eng. Des. 87 (9) (2012) 1670–1678.
[30] S. Garcia, J. Derrac, J.R. Cano, F. Herrera, Prototype selection for nearest neighbor classification: taxonomy and empirical study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 417–435.
[31] I. Triguero, S. Garcia, F. Herrera, Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification, Pattern Recognit. 44 (4) (2011) 901–916.
[32] M. Sedighizadeh, S. Ahmadi, M. Sarvi, An efficient hybrid Big Bang–big crunch algorithm for multi-objective reconfiguration of balanced and unbalanced distribution systems in fuzzy framework, Electr. Power Compon. Syst. 41 (1) (Jan. 2013) 75–99.
[33] O. Hasançebi, S. Kazemzadeh Azad, An exponential big bang-big crunch algorithm for discrete design optimization of steel frames, Comput. Struct. 110–111 (2012) 167–179.
[34] A. Kaveh, S. Talatahari, Optimal design of Schwedler and ribbed domes via hybrid Big Bang–Big Crunch algorithm, J. Constr. Steel Res. 66 (3) (2010) 412–419, https://doi.org/10.1016/j.jcsr.2009.10.013.
[35] T.A.W.-C. Fu, P.M.-S. Chan, Y.-L. Cheung, Y.S. Moon, Dynamic vp-tree indexing for n-nearest neighbor search given pair-wise distances, VLDB J. 9 (2) (2000) 154–173.
[36] X. Wu, V. Kumar, Top 10 Algorithms in Data Mining, Springer, Berlin, Germany, 2007.
[37] S. Uddin, A. Khan, M. Hossain, et al., Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inform. Decis. Mak. 19 (2019) 281, https://doi.org/10.1186/s12911-019-1004-8.
[38] K. Mittal, G. Aggarwal, P. Mahajan, Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy, Int. J. Inf. Technol. 11 (2019) 535–540.
[39] H.-L. Chen, B. Yang, J. Liu, D.Y. Liu, A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Syst. Appl. 38 (7) (2011) 9014–9022.
[40] V. David Sánchez, Advanced support vector machines and kernel methods, Neurocomputing 55 (1–2) (2003) 5–20.
[41] J.D. Rodriguez, A. Perez, D. Arteta, D. Tejedor, J.A. Lozano, Using multidimensional Bayesian network classifiers to assist the treatment of multiple sclerosis, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 42 (6) (Nov. 2012) 1705–1715, https://doi.org/10.1109/TSMCC.2012.2217326.
[42] P.B. Pankajavalli, G.S. Karthick, R. Sakthivel, An efficient machine learning framework for stress prediction via sensor integrated keyboard data, IEEE Access 9 (2021) 95023–95035, https://doi.org/10.1109/ACCESS.2021.3094334.
[43] M.M. Mehdy, P.Y. Ng, E.F. Shair, N.I. Saleh, C. Gomes, Artificial neural networks in image processing for early detection of breast cancer, Comput. Math. Methods Med. (2017).
[44] Somayeh Moghaddam, Mohammad Fallah, Hamed Kazemipoor, Amir Salehipour, A fuzzy inference-fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases, Expert Syst. Appl. 95 (2017), https://doi.org/10.1016/j.eswa.2017.11.001.
[45] K. Beer, D. Bondarenko, T. Farrelly, et al., Training deep quantum neural networks, Nat. Commun. 11 (2020) 808, https://doi.org/10.1038/s41467-020-14454-2.
[46] Renu Narain, Sanjai Saxena, Achal Goyal, A novel heart disease prediction system based on quantum neural network using clinical parameters, Annu. Res. Rev. Biol. 14 (2017) 1–10, https://doi.org/10.9734/ARRB/2017/10456.