IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 12, 2020, accepted March 29, 2020, date of publication April 20, 2020, date of current version May 6, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2987057*

# Filter-Based Multi-Objective Feature Selection Using NSGA III and Cuckoo Optimization Algorithm

**ALI MUHAMMAD USMAN**[1,2], **(Member, IEEE), UMI KALSOM YUSOF**[1], **AND SYIBRAH NAIM**[3], **(Member, IEEE)**

[1]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[2]Department of Computer Sciences, Federal College of Education (Technical), Gombe 760212, Nigeria
[3]Technology Department, Endicott College of International Studies (ECIS), Woosong University, Daejeon 300718, South Korea

Corresponding author: Umi Kalsom Yusof (umiyusof@usm.my)

This work was supported by the Universiti Sains Malaysia together with the Endicott College of International Studies (ECIS), Woosong University, South Korea, under Grant 1001/PKOMP/8014084.

**ABSTRACT** Feature selection aims to confiscate inappropriate features and yet improve classification performance. These aims are conflicting with one another, and a choice must be made in the presence of the trade-off between them. Numerous researches deal with feature selection problem but, they are mostly single-objective based. Nowadays, multi-objective optimisation approaches are becoming the most suitable approaches to deal with feature selection problems. They can easily create a balance between selected features and classification accuracy or error rate. Evolutionary computation techniques have been applied for multi-objective feature selection. Cuckoo optimisation algorithm is among the most popular technique that is exceptional in solving the problems of feature selection. Based on the binary cuckoo optimisation algorithm, two different multi-objective filter-based feature selection frameworks are presented with the idea of nondominated sorting genetic algorithms NSGAIII (BCNSG3) along with NSGAII (BCNSG2). Thus, four multi-objective filter-based feature selection approaches are proposed by employing mutual information along with gain ratio based-entropy as the respective filter evaluation measures in all the proposed frameworks. The results obtained are examined and analysed against the existing methods and single objective scheme on fourteen (14) datasets of varying degree of difficulties. The outcome of the experiments displays that the proposed multi-objective algorithms successfully derive a set of nondominated solutions that used the least feature size and attained the best error rate than using full-length features. In general, BCNSG2 obtained the best results compared to the existing methods and single-objective algorithm, whereas BCNSG3 outdoes all other approaches.

**INDEX TERMS** Cuckoo optimization algorithm, multi-objective feature selection, NSGA III, NSGA II, gain ratio based-entropy, mutual information, machine learning and classification.

## I. INTRODUCTION

We are nowadays in the epoch of big data where data has become ubiquitous in various domain ranging from bioinformatics, social media, healthcare, manufacturing industries and online education. The rapid expansion of data is a severe challenge in handling the information effectively. Thus, the necessity to put on data mining together with machine learning approaches to determine unseen knowledge arising out of the large stored data [1], [53]. Classification is one of the data mining technique that is employed to categorise each row in a dataset into a set of groups according to their class label. It is a known fact that feature size is the key problem that deters the work of all classifiers [2]. However, if a piece of previous information about the useful and most relevant features is available the task is not challenging, else it will be hard to discover the most valuable and relevant features primarily when the number of features is considerable [3]. The term feature selection is introduced to select the ultimate and appropriate features from these enormous volumes of data.

FS is also one of the data mining processes that are used to pick the appropriate features from a dataset. The main issue of FS is in what way one can explore for

The associate editor coordinating the review of this manuscript and approving it for publication was Shangce Gao.

the perfect subsets and then, assess the perfectly generated subsets [4].

The current algorithms used as a search technique cannot efficaciously explore the large search space of FS without been stuck into the local optima [5]. Currently, evolutionary computations (EC) have been employed as search techniques to explore the large search space in an FS problem. However, most of them go through early convergence. Cuckoo optimisation algorithm (COA) is amongst the EC techniques that are testified in [6] to have resourceful exploration operatives that determine the whole promising area in the exploration space and converges earlier than many other EC based techniques. Based on that, binary COA (BCOA) is employ as a search method to explore the most appropriate subsets of feature automatically.

Evaluating the subset of the features produced depends on the type of FS. This can be either filter or wrapper. Wrappers use a classifier to measure the accuracy for each of the selected subgroup of features. Nevertheless, this procedure is computationally cost more especially on datasets the large number of features [7]. Alternatively, Filter-based approaches are computationally cheap and performed well on big datasets. The main drawback of the filter-based FS is the absence of feature dependency or connection among the carefully chosen features [1], [5]. Information theory is amongst the whole theories used to estimate both the relevance and redundancy amongst two or more features along with their target class [8].

Using the concepts of information theory, to discover the redundancy, as well as relevancy of nominated features using different EC techniques, is becoming popular nowadays. For example, Cervantes *et al.* in [9] and [10] both used the ideas of information theory, especially mutual information (MI) along with entropy as a fitness function in a binary particle swarm optimisation algorithm (BPSO). Different weights values are employed to enhance the relevancy and reduce the redundancy on the datasets, and a better result was achieved. Recently, in the work of Hancer *et al.*, in [11] differential evolution (DE) was used for feature grading with the assistance of information theory ideas including relief f, MI and Fisher scores. The outcome got supersede single and multi-objective methods offered. The literature, showed that EC techniques are gaining popularity for filter-based FS, predominantly with the idea of information theory [11]. However, there are other EC techniques like COA that showed an encouraging outcome and yet not been considered for FS specifically using the idea of information theory.

FS aimed at minimising error rate and consequently reduce the size of the features, thus, considered as multi-objective optimisation problem [12], [40]. These aims are conflicting to one another, and the optimal choice needs to be carried out in the company of the compromise between them, however, very little work is conducted on multi-objective FS. References [10], [12] use the idea of nondominated sorting genetic algorithm II (NSGAII) BPSO and PSO respectively. Generally, NSGAII was commonly used as multi-objective optimisation to find the optimal solution of the various objective functions [13]. Although, NSGAII is reported to be slightly computationally expensive and outdated but can successfully evolve the set of nondominated solutions [14]. Recently, in the work of [15], the idea of NSGAIII was proposed. It is less computationally expensive and can successfully evolve the set of nondominated solutions for many-objective functions. Since the inception of NSGAIII, its neither use for FS nor enhance with other EC techniques to solve a feature selection problem. To our knowledge, no work used BCOA specifically, as a multi-objective FS to date.

In a nutshell, COA can solve only the continuous optimisation problem, and FS can be best solved as a binary discrete optimisation problem; thus, BCOA is proposed in this study. A “1” means a feature is selected while “0” means otherwise. Moreover, Feature selection is now considered as a multi-objective optimisation problem that aims at reducing the number of selected features and consequently improving the classification performance — hence considered as two objective optimisation problem. In this study, two concepts of multi-objective optimisation algorithms, particularly NSGAII and NSGAIII, optimisation are employed to tackle the issues of multi-objective feature selection and obtained the set of nondominated solutions. NSGAII can solve two-objective optimisation problems like feature selection. However, NSGAIII is strictly meant to address many-objectives optimisation problems. In this study, both NSGAIII and NSGAII are used for the first time to solve the feature selection problem along with BCOA.

The generic aim of this study, is to adopt BCOA [16] with entropy (gain ratio based entropy) and MI as the evaluation measures together with the idea of nondominated sorting genetic algorithms NSGAII (BCNSG2MI and BCNSG2E), and NSGAIII (BCNSG3MI and BCNSG3E) to find the set of nondominated solutions with fewer number of features and comparable or better classification performance than using the full-length features and within a short period.

The proposed FS algorithms were investigated and scrutinised on the UCI standard benchmark datasets of varying degree of difficulties. Precisely, this study will scrutinise whether

- the filter-based single objective approach with gain ratio based-entropy (BCOA-E ) and MI (BCOA-MI) as the evaluation measures might select fewer features and enhance classification accuracy than using the full-length features.
- the proposed multi-objective BCNSG2MI and BCNSG2E FS algorithms can evolve a set of nondominated solutions that might perform better than the filter-based single objective and other existing methods; and,
- the multi-objective BCNSG3MI and BCNSG3E can evolve a set of nondominated solutions that might perform better than the approaches above as well as other existing methods.

Apart from the introduction, the rest is organised as: Section 2 illustrates the contextual information including

COA, BCOA, multi-objective optimisation, information theory concepts as well as related works. Section 3 presents the proposed filter-based multi-objective BCOA, each using the information theory concepts along with NSGAII and NSGAIII respectively. Section 4 displays the experimental design while Section 5 demonstrates the outcomes and discussions. To end, in Section 6, the conclusions were examined along with more research directions.

## II. BACKGROUND

### A. TRADITIONAL FILTER-BASED FEATURE SELECTION

Majority of the filter-based approaches are employed to rank a feature to its target class based on some suitable evaluation measures. The target is to confiscate irrelevant and redundant features, and the challenge is how to hunt for the best subset of features with the standard evaluation measures. For example, Kira and Rendell in [17] presented a classical filter-based FS algorithm known as relief algorithm. It uses some statistical methods and hence avoids the heuristic search. It allocates weight to all feature to symbolise how statistically importance a feature is to its target class. Nevertheless, the relief algorithm does not consider irrelevant features since it concentrates on finding all statistically relevant features irrespective of the redundancy amid them. Also, a decision tree (DT) algorithm was proposed by [18] to enhance the classification performance of case-based learning. The results achieved indicate that the features produced by the DT can automatically aid to diminish the error rate of the DT classifier.

Another filter-based algorithm named FOCUS is presented by Almuallim and Dietterich in [19], FOCUS is an exhaustive search algorithm that explores all the possible feature subset, then, later on, select the least subset. However, this makes it computationally expensive due to the exhaustive search, especially on large dimensional datasets. In another perspective, [20] developed an MI feature selector (MIFS) method in a supervised neural network and categorised the features as relevant and redundant. Redundant features are features with low information content or high redundancy. A heuristic function was employed to control and balance between the relevance and redundant features. Lastly, features are selected greedily as it is in the greedy algorithm apart from the fourth step. Then, [21] enhanced the limitation of MIFS mentioned by introducing another greedy search and uniformly improved MI feature selector (MIFS-U) is used to choose the useful features and halts as soon as it has reached the required number of features. One of the algorithms considers using MI along with input features and output classes compared to the MIFS that cannot perform well on nonlinear problems.

Bishop and Bishop in [22] proposed a supervised filter-based FS algorithm called Fisher score. It works by ranking features based on discriminant ability agreement, which evaluates the features individually. The limitation of this algorithm is that there is still redundancy on the chosen features since there is no correlation among the chosen features. Similarly, [23] developed a fast correlation-based feature selection (CFS) that can work for continuous and discrete data. The results obtained showed that it outperformed naïve Bayes, instance-based learning, relief F and DT. The CFS algorithm used heuristic techniques for FS. As such, it finds features that are extremely correlated to the target class but not correlated with each other. Even though, systematic uncertainty was applied to measure the level of the correlation; nevertheless, the relationship among the features cannot work well on several features. Reference [24] presented a relief F a variant of relief algorithm for feature ranking which also ranks a score for each feature separately based on the KNN algorithm. Despite being amongst the best filter-based FS, its, however, have some redundant subset of features. On the other hand, [25] proposed an alternative way of selecting features that have maximum relevance to the target class. In that case, the selected features will individually have the largest mutual information with the target class. The proposed technique works in two stages, at the initial stage a two-stage FS by merging minimal redundancy maximal relevance (mRMR) and other wrapper-based FS techniques. After selecting the best features at a little cost, the outcomes exposed that mRMR achieved better results on both accuracy and the nominated feature size.

Later Ling and Tang in [26] introduced class relevance and redundancy framework based on information theory. A novel algorithm named conditional informative feature extraction that improves the info carried by the entire set of features by clearly minimising the class redundancies. Besides, the computational cost as one of the major issues of information theory drastically reduced by coupling discrete approximation along with 1D Parzen window method and the local active region method. In order to, ranked features in descending order of mean and standard deviation along with their class label. A Pearson correlation coefficient was introduced in [27]. The algorithm chooses the subsets with the smallest validation error. Also, a predictor was used on the M nested subsets. Although it is computationally inexpensive and simple to implement, it leads to feature independent since it can only recognise a linear relationship between a feature and its target class.

Also, Huawen *et al.*, in [28] developed a dynamic MI feature selection, where the MI of particular features were recomputed on unlabelled instances, compared to the entire sampling space. The results obtained performed well on 16 UCI datasets with four standard classifiers.

Furthermore, Estevez *et al.*, in [29] presented a normalised MI feature selection (NMIFS) a development over MIFS, MIFS-U, and mRMR approaches. The mean of the NMIFS was applied to estimate redundancy among the selected features. The experimental outcomes showed that it outperformed the three other MI methods on several benchmark datasets without demanding any user-defined parameter.

On the other hand, [30] proposed an extension of the Shannon MI amid feature and class label along with the use of this extension to the naturally derived space of possible filter criteria. This was achieved by adding a class-conditional correlation to the main equation of mutual information denoted as the first- order utility. Other solid mathematical backgrounds and theoretical concepts of mutual information are presented.

Laplace Score (LS) is among the favourite filter-based ranking technique that is used for both supervised and unsupervised FS. It works with useful features and rejects the features with high variance. Reference [31] proposes LS along with entropy measure, the idea is to select successful features by substituting the standard k-means in LS with an information distance measure. A better result was achieved compared to the LS in terms of efficiency, stability and scalability. An iterative LS based neighbourhood graph was proposed in [32], and the results showed that better features were chosen according to the structure of the graph.

Still Foithong, Pinngern, and Attachoo in [33] developed another FS approach through MI measure deprived of demanding a user-defined parameter for the choice of the candidate feature set. Despite [34] established a comprehensive library for FS which presents other measures, like MI, and Fisher Score to compute correlations amongst features. Recently, [11] introduced a new filter criterion encouraged by the concepts of MI, Relief F, as well as Fisher Score. As an alternative of using shared redundancy, the expected norm attempts to select the peak ranked features regulate by Relief F and Fisher Score while specifying the mutual relevance within features as well as the target class labels.

### B. EVOLUTIONARY COMPUTATION FOR FILTER-BASED FEATURE SELECTION

Moghadasian and Hosseini in [36] developed a filter-based cuckoo search algorithm (CSA) along with MI and entropy are used as evaluation criteria on some high dimensional datasets. The results of the classification accuracy using ANN showed that almost 90% of the real features are minimised considerably. CSA with entropy performed well on classification performance whereas CSA with MI on the selected features than using the full complete features.

Similarly, Cervante *et al.*, in [9] presented a BPSO algorithm together with entropy and MI as an evaluation measure. The results obtained on the four data sets showed that BPSO with mutual information could develop a set of features along with fewer features. Whereas, BPSO with entropy has more classification accuracy using a DT compared to BPSO with MI. Similarly, the work is extended in [37], whereby, a multi-objective filter-based FS using BPSO and nondominated sorting genetic algorithm with information measures as the evaluation criteria are presented. The results obtained was tested on six data sets where DT was used to measure the classification error rate. Moreover, [38] developed another multi-objective filter-based FS. GA fitness function with both MI and entropy as evaluation measures are embedded as a single-objective based FS. While GA+MI chose the least but an appropriate number of features, GA with entropy performed well in terms of classification performance. Furthermore, strength Pareto evolutionary algorithm (SPEA2) and NSGAII are enhanced with MI and entropy. The results showed that both SPEA2 and NSGAII outperformed the single-objective algorithm and NSGAII outperformed SPEA2 specifically on the features that are carefully chosen.

Xue, Zhang and Browne in [12] developed a crowding, dominance, and mutation PSO (CMDPSOFS) for multi-objective FS by improving the performance and defining suitable operators. Similarly, a cost-based multi-objective PSO for FS named hybrid mutation PSO (HMPSOFS) was presented in [39]. The proposed HMPSOFS used a hybrid mutation and updated the speeding up coefficients together with an adaptive mechanism. Whereas, the CMDPSOFS enhances the variety of search by applying both the regular and irregular mutation operators together with the anticipated mutation mechanism. However, the planned approaches can be used only to solve feature selection problems, whereas other approaches might yield better results.

Nguyen *et al.*, in [40] introduced insert, swap and remove PSO feature selection (ISRPSOFS) a local search based on sequential, a forward or backward search is performed by inserting removing and swapping operators. However, the proposed method is computationally expensive, particularly on more extensive data where the redundant and irrelevant features are many.

A filter-based FS based on differential evolution (DE) was developed in [11]. MI of the highest rank features by Relief F and Fisher score are selected. Based on that, two filter-based DE are proposed. The first one has just one objective in a weighted way. Whereas, the second one is on multi-objective optimisation. The proposed method was compared with mutual information feature selection (MIFS) adopted also using DE single-objective as well as multi-objective approaches.

Moreover, it performed better than MIFS and DE for the pair of single-objective along with multi-objective on all the data sets with reduced feature size and better classification accuracy. In the same vein, [41] presented another multi-objective filter-based FS using artificial bee colony (ABC). Both the numerical ABC, as well as its binary counterparts, are examined using nondominated sorting method and genetic operators. The binary ABC outperformed its numerical counterparts both on accuracy and as well as the selected features.

Applied data envelopment analysis (DEA) method along with COA for dealing with multi-objective optimisation problems, are presented in [42]. The profit function of the COA is substituted by the efficiency value that is obtained from DEA. Later on, COA is hybridised with simple additive weighting (SAW) [43]; the proposed COAW algorithm has high speed in finding the Pareto frontiers and can find the starting and stop points of Pareto frontiers appropriately. However, all the COA-based multi-objective presented are a hybrid based not multi-objective based.

On the other hand, there has been no COA based multi-objective optimisation proposed in the literature like other EC based techniques mentioned earlier. Recently [4] developed a filter-based COA using the general filter algorithm as the fitness function of the COA. Some heart disease data sets were applied to validate the efficacy of the proposed method. However, the results obtained are in favour of filter-based CSA, especially on the small size datasets. Just because most of the data sets have few numbers of features. Perhaps, if its demonstrated on high dimensional data or enhance to avoid redundancy among selected subsets, it may provide a better result as argued by [6].

Most of the existing studies show that COA has limited application, especially in FS compared to other evolutionary computation based techniques likes PSO, GA, ACO and ABC among others.

On the other hand, NSGA is the most common multi-objective optimisation algorithm. Since it has shown promising results in solving different kinds of multi-objective optimisation problems in various domain.

With the introduction of NSGAII in [13] it becomes more potent in handling multi-objective issues. Hamdani *et al.*, in [49] proposed the first multi-objective FS framework using the NSGAII. Based on that, [10] applied the framework and developed a multi-objective filter-based FS using BPSO. In addition to that, PSO along with MI and entropy were used as evaluation criteria within the NSGAII in [38]. However, these methods are limited to the application of NSGAII along with PSO and BPSO alone. Whereas, there are other EC techniques such as COA with proven records and yet not use in that regards.

In an attempt to reduce the computational cost of wrapper-based FS without jeopardising the results of the FS, [55] presented a faster multi-objective FS by incorporating an improved ABC based on particle update model into the framework. In the framework, k-means clustering, along with ladder-like sample utilisation, are employed to minimise the cost of the evolutionary process. The experimental results showed that it has promising results and performed better than NSGAII-FS, among others.

To achieve local a trade-off between both local exploitation and global exploration [56] proposed binary DE with self-learning strategy to solve the multi-objective FS problems. Based on that, three operators are employed to achieve better and promising results. New binary mutation operator that will aid and fasten in locating the most promising regions. And new one-bit purifying search operator that can aid the self-learning strategy of elite individuals and (3. A nondominated sorting operator with crowding distance that can reduce the time consumption of selection operators. The proposed MOFS-BDE performed well on public data sets and competitive in comparison with DEMOFS, NSGAFS, MOPSOFS, and B-MOABCFS) and a new MOEA/D method (MOEA/D-2TMFI. However, the results obtained are not compared and analysed with NSGAII and NSGAIII.

On the other hand, [60] introduced a novel swarm intelligence algorithm, known as Rc-BBFA, and effectively used it to solve FS problems. The proposed algorithm extends the idea of FFA by presenting binary variables. Three new strategies, i.e. the return-cost attractiveness, the Pareto dominance-based selection, and the binary movement with the adaptive jump, are employed in the novel algorithm, which is effective in handling the FS problems. Experiments on ten well-known datasets were conducted, and promising results were obtained compared to others mentioned in the paper. However, the results are not compared with the most recent multi-objective evolutionary algorithms such as NSGAIII and MOEA/D, among others.

Recently [57] proposed a new PSO-based unsupervised FS approach, known as filter-based bare-bone particle swarm optimisation algorithm (FBPSO). Local filter-based search strategy based on feature redundancy is employed to enhance the exploitation ability of the swarm, on the other hand, space reduction strategy using the mean of mutual information is employed to eliminate the irrelevant and redundant features faster.

Since most of the existing multi-objective evolutionary algorithms experience difficulties in resolving many-objective optimisation problems owing to the inability to balance the convergence and diversity in high-dimensional space. Reference [58] propose a new many-objective evolutionary algorithm using a one-by-one selection strategy. It works like this; once an individual is selected, its neighbours are de-emphasise using a niche technique to guarantee the diversity of the population, in which the similarity between individuals is examined and evaluated using a distribution indicator. The comparative results show the goodness of the proposed method. However, this method is not examined on multi-objective FS problems.

Similarly, [59] proposed another multi-objective evolutionary optimisation based on reference points (RPEA). It exploited the potential of the reference points in handling many-objective optimisation problems. The proposed RPEA can primarily be categorised as: (1) adaptively generating a series of reference points with good convergence and distribution based on the evolution of a population; (2) greatly increasing the selection pressure toward the Pareto front by calculating the distances between the reference points and the individuals in the environment selection process. The proposed method was applied to seven benchmarks many-objective optimisation problems and compared with the other four state-of-the-art methods to evaluate its performance. The results reveal that RPEA is very competitive to the others in terms of seeking for a solution set with good approximation and distribution in many-objective optimisation. Also, this work is not tested on multi-objective FS.

Moreover, [60] introduced a novel swarm intelligence algorithm, known as Rc-BBFA, and effectively used it to solve FS problems. The proposed algorithm extends the idea of FFA by presenting binary variables. Three new strategies, i.e. the return-cost attractiveness, the Pareto dominance-based selection, and the binary movement with the adaptive jump, are employed in the novel algorithm, which is effective in handling the FS problems. Experiments

on ten well-known datasets were conducted and promising results were obtained compared to others mentioned in the paper. However, the results are not compared with the most recent multi-objective evolutionary algorithms such as NSGAIII and MOEA/D among others.

In the same vein, [61] proposed an improved MOPSO, termed as BMOPSOFS to solve FS problems with unreliable data. To achieve that, the probability-based encoding strategy, the reinforced memory and the hybrid mutation, together with several established techniques, such as the external archive and the crowding distance are proposed. It makes BMOPSOFS more effective in dealing with the multi-objective FS problems and performs well on various benchmark datasets.

Recently, [62] presented an unsupervised FS approach by combining the discriminative information of class labels with subspace learning. The nonnegative Laplacian embedding was initially employed to produce pseudo labels, to enhance the classification accuracy. Then, an optimal feature subset is chosen by the subspace learning guiding by the discriminative information of class labels, on the premise of maintaining the local structure of data. Based on that, an iterative strategy for updating similarity matrix and pseudo labels was developed, which bring more accurate pseudo labels that provide the convergence of the proposed strategy. The results on six real-world datasets show the goodness of the proposed method over other seven state-of-the-art methods.

To enhance convergence and exploitation ability of ABC, [54] presented a two archived guided multi-objective ABC called TMABC-FS. The first archives comprise of the external archive and the leader archive that are employed to improve the searchability of various kinds of bees. And two new operators; convergence-guiding search for employed bees and diversity-guiding search for onlooker bees, are proposed for gaining a group of non-dominated subsets of the feature with better distribution and convergence. The proposed TMABC-FS is validated on different UCI benchmark datasets and is compared with two traditional algorithms and three multi-objective approaches. The results have shown that TMABC-FS is an effective and vigorous optimisation method for solving cost-sensitive FS problems.

The concept of NSGAIII is introduced in [50], and it has since recorded numerous achievement since its introduction [14]. However, the used of NSGAIII, particularly for filter-based FS, is limited in the literature. Therefore, in this study, the frameworks of both NSGAII and NSGAIII are adopted with BCOA along with MI and the combined entropy as an evaluation measure.

### C. CUCKOO OPTIMISATION ALGORITHM

An innovative EC-based technique called the cuckoo optimisation algorithm (COA) was developed by [16]. COA has its rules as follows:

1) The variables should be in an array named “habitat” of $N_{pop} \times N_{var}$.

$$habitat = [x_1, x_2, \ldots, x_{Nvar}] \tag{1}$$

2) The upper and lower limit iterations use 5-20 eggs respectively.
3) The maximum range distance for egg laying is

$$ELR = \alpha \times \frac{number\ of\ current\ cuckoos}{total\ number\ of\ eggs} \times v_{hi} - v_{low} \tag{2}$$

where, $\alpha$ is an integer, and $v_{hi}$,$v_{low}$ are respective limits in step 2 above. In the Eq.2, $\alpha$ is set to 1. The search space is in the interval of (-55, 55) and twenty cuckoos in the population. By nature, a cuckoo can only lay 5-20 eggs. In this study, the same concept was used that five cuckoos with less profit lay five eggs and also other fifteen cuckoos lay an egg in the interval [6, 20] proportional to their profit. Thus, the total number of eggs will be 220. To compute the ELR of a cuckoo whose profit is in the 5th order we used:

$$ELR = 1 \times \frac{16}{220} \times (55 - (-55)) = 8$$

It signifies that a cuckoo with profit of 16 degrees can lay egg within a circle of 8 radius.
4) Just a p% of the eggs i.e. 10% with a smaller amount of profit value and more cost will be killed.
5) A k-means of 3-5 is sufficient in most simulations.
6) Every single cuckoo flies only λ % distance towards goal line habitat with a deviation of $\omega$ radians as shown below:

$$\lambda \sim U(0, 1) \qquad \varphi \sim (-\omega, \omega) \tag{3}$$

The labelled, $\lambda \sim U(0, 1)$ shows that λ is a constantly distributed arbitrary number within the range of 0 and 1. $\omega$ is a parameter that limits a nonconformity from goal line habitat. An $\omega$ of $\pi/6$ rad is mostly okay and suitable.

The detailed algorithm of the typical COA is shown in Algorithm 1.

**Algorithm 1** The Typical COA Pseudocode

```
Begin
    Set cuckoo locations through some arbitrary ideas on the global function
    Dedicate some eggs roughly to respectively cuckoos
    Compute ELR for every single cuckoo
    Allow the cuckoos to lay their eggs in their matching ELR
    Destroy those cuckoos familiar by the multitude birds
    Allow egg to hatch and baby chicken raise
    Estimate the location of every newly mature cuckoo
    Restricts cuckoos' highest number in location and destroy those that exist in substandard locations
    Group cuckoos and discover the best cluster and choose goal line environment
    Allow the new cuckoo populace to settle at the goal line environment
    If stop criteria are fulfilled stop, else go to 3
End
```

Later after the development of the COA, since its meant to solve only continuous optimisation problems. Then, Mahmoudi and Rajabioun in [45] introduced the BCOA that is capable of dealing with binary discrete optimisation problems. To compute the $X_{goal}$ and $X_{Curpos}$ of the habitat the following equation is used below:

$$X_{Nhabitat} = X_{Curpos} + rand(X_{goal} - X_{Curpos}) \tag{4}$$

To offer a new habitat $X_{Nhabitat}$ appropriate for discrete binary difficulties, a sigmoid function in the Eq.(5) was employed. The reason is to map $X_{Nhabitat}$ into the range [0,1]. Lastly, Eq.(6) will modify the values in the habitat as 0 or 1. Whereby rand in Eq.(6) is an arbitrary number, that is produced randomly.

$$S = \frac{1}{(1 + e^{-X_{Nhabitat}})} \tag{5}$$

$$IF \quad S > rand \qquad \text{THEN} \qquad X_{Nhabitat} = 1 \qquad \text{AND}$$
$$IF \quad S < rand \qquad \text{THEN} \qquad X_{Nhabitat} = 0 \tag{6}$$

### D. INFORMATION THEORY

Entropy H(X) is the degree of ambiguity of an arbitrarily variable relative to the possibility of manifestation of an event. The detailed definition of entropy is shown in Eq.(7). The possibility of the manifestation of an event happens only if the entropy is high else not.

$$H(X) = -\sum_{i=1} P(x_i) log_2 P(x_i) \tag{7}$$

The termed, $X$ is an Where both the joint and conditional entropy of X and Y are:

$$H(X, Y) = -\sum_{i,j} P(x_i, y_j) log_2 P(x_i, y_j) \tag{8}$$

$$H(X|Y) = -\sum_{i,j} (P(x_i, y_j) log_2 P(x_i|y_j) \tag{9}$$

where $X = x_1, x_2, \ldots, x_i \ldots, x_n$ and
$Y = y_1, y_2, \ldots, y_j \ldots, y_m$
Mutual information (MI) is employed to measure the relationship amongst two arbitrary variables and evaluate the relevance of the feature subset [46]. The MI between X and Y features can be defined as

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$
$$I(X; Y) = -\sum_{i,j} P(x_i, y_j)\ log_2 P \frac{P(x_i, y_j)}{P(x_i).P(y_j)} \tag{10}$$

Eq.(10) means that the $I(X; Y)$ is larger if $X$ and $Y$ are interconnected. Otherwise, they are not connected whatsoever.

### E. MULTI-OBJECTIVE OPTIMISATION

By nature feature selection is considered as multi-objective optimisation problems (MOP). MOP usually occur when optimum decisions are required to be made in the company of the trade-offs or agreement amid the different objectives [47]. It comprises minimising or maximising the various disagreeing objective functions. The solution to the problem is normally a set of solutions that define the best trade-off between competing objectives. Mathematically, it can be written as follows:

$$Minimise\, f_m(x), \qquad m = 1, 2, \ldots, M \tag{11}$$
$$subject\ \text{to}: \ g_j(x) \leq 0, \qquad j = 1, 2, \ldots J,$$
$$and \qquad h_k(x) = 0, \qquad k = 1, 2, \ldots K \tag{12}$$

$x_i^{(L)} \leq x_x \geq x_j^{(U)}$ $i = 1, 2, \ldots, n\, x$ is the candidate solution vector, $f_m(x)$ is the *mth* objective function to be *minimize* or *maximise*, $f(x)$ is the objective function, $h_k(x)$ and $g_j(x)$ are the constraint functions, $J$ and $K$ are integer numbers $x_i$ and $x_j$ are lower and upper bound respectively.

In the single-objective optimisation problem, the superiority of a solution over other solutions is readily determined by comparing their objective function values. In the case of the multi-objective optimisation problem, enhancing one objective may worsen another. As such balance in trade-off solutions is accomplished if a solution cannot enhance any objective deprived of degrading one or more of the other objectives and this is called Pareto improvement [48].

The dominance determines the goodness of a solution. For instance, let $y$ and $z$ be two candidate solution vectors of the $f_m(x)$ to be maximize or minimize. If the criteria in Eq.(13) are satisfied, then $y$ dominates $z$ or $y$ is good compared to $z$ or $z$ is dominated by $y$

$$\forall i: f_i(y) \leq f_i(z) \qquad AND \qquad \exists j: f_j(y) \leq f_j(z)$$
$$i, j \in 1, 2, \ldots, M \tag{13}$$

When a solution is nondominated by any other solutions or no further Pareto improvement can be made, it is referred to as a Pareto-optimal solution or nondominated solutions. The set of the complete Pareto-optimal solutions forms the agreement outward in the search space and is known as the Pareto front [12], [47].

FS has two opposing objectives; these are reducing feature size along with the error rate of a classifier. Thus, considered a multi-objective minimisation problem.

## III. THE PROPOSED BCOA FILTER-BASED APPROACHES

This section presents the proposed filter-based approaches. The first one is the single objective filter-based using gain ratio based-entropy together with MI as the fitness evaluation measures. Whereas, the second one is according to MOP, especially the NSGAII and NSGAIII frameworks in addition to the single objective.

### A. BCOA FILTER-BASED SINGLE-OBJECTIVE APPROACH

Two filter-based BCOA algorithms BCOA-MI and BCOA-E, each with MI and gain ratio based-entropy as the respective evaluation criteria, are proposed in this section. The details of both BCOA-MI, as well as BCOA-E, is depicted in Algorithm 2

### 1) BCOA-MI

The essence of MI is to measure the relationship between two pair of features together with their target class. The target is to choose highly relevant features and eliminate the most redundant features. Majority of the researches that address the issue of feature interaction between the pair of features used the MI in Eq.(14). The details is as shown below:

$$\begin{aligned} Fit_{mi} &= \beta(Rel_{mi} + Red_{mi}) - Red_{mi} \\ \text{where,} \quad Rel_{mi}(X; C) &= max \sum_{i} I(x; c) \\ Red_{mi}(X; Y) &= min\left(\frac{1}{|M|} \sum_{i,j} I(x_i; y_j)\right) \end{aligned} \tag{14}$$

$X$ and $Y$ stands for the distinct binary feature subsets, $M$ is the feature size, $C$ is the target class label, $Rel_{mi}$ applies a pairwise method to compute the MI relevance amongst every feature together with its class label, and finally, $Rel_{mi}$ remove the redundancy that remains in each pair of the chosen features. As such, in Eq.(14) $Fit_{mi}$ is a maximisation function that makes the best use of the relevancy $Rel_{mi}$ and synchronously decreases the redundancy $Red_{mi}$ of the selected features.

### 2) BCOA-E

In contrasts to the $Fit_{mi}$, the $Fit_E$ is employ to compute the relevance along with the redundancy among a group of features not necessarily between two pair of features alone. Eq.(15) displays the fitness function as:

$$\begin{aligned} Fit_E &= \beta(Rel_E + Red_E) - Red_E \\ \text{where,} \quad Rel_E(X; C) &= max(GR \sum_{i} I(x; c)) \\ GR(x) &= Gain(x)/splitinfo\, x(c) \\ Red_E(X; Y) &= min\left(\frac{1}{|M|} \sum_{i,j} GR(x\, X/x)\right) \end{aligned} \tag{15}$$

$Rel_E$ estimates the gain ratio of the features in $X$, using (15). $Fit_E$ is also consider as a maximisation function that makes the most used of relevancy $Rel_E$ and synchronously reduces the redundancy $Red_E$ of the selected subset of features.

### 3) WEIGHT FOR BOTH BCOA-MI AND BCOA-E

Weighted values of 0.9, 0.8, 0.75, 0.7, 0.6 and 0.5 are used for both $\beta 1$ and $\beta 2$. The two $\beta$ ($\beta 1$ and $\beta 2$) values are used as MI and gain ratio based-entropy to examined the most relevant and redundant weighted values respectively. The algorithm is shown in in Algorithm 2 whereby, Eq.(14) and Eq.(15) are been used as the fitness function.

**Algorithm 2** Proposed BCOA-MI and BCOA-E

```
Start
    Initialise each habitat with some features from a dataset
    Collect the features in their respective habitats
    Explain ELR for every single cuckoo using Eqs 5 and 6
    Allow the cuckoos to lay their eggs in their matching ELR
    Destroy those cuckoos familiar by the multitude birds
    Allow egg to incubate and baby chicken raise
    Estimate the environment of every recently grownup cuckoo
    Limits cuckoos' highest number in location and abolish those that exist in poorer environments
    Group cuckoos and discover finest cluster and choose goal line environment
    Allow the new cuckoo populace to settle at the goal line environment
    Return the optimum solution (selected features)
    Evaluate the fitness function according to 14 and 15
    If the stop condition is satisfied stop, else go to 3
Stop
```

From Algorithm 2, one can observe that Eq.(1) is used to initialise each dataset. Unwanted features that are recognised based on the computation of the fitness function in Eq.(14) and Eq.(15) are detached. It happens mostly if the population in the worst area is killed because it's less than the maximum value or else it gets some profit values. The nest with the best survival rate (feature subsets) can then move to the best environment using Eq.(5) and Eq.(6). The ELR is calculated using Eq.(2). The steps mentioned above will repeat until the best solution with the highest-ranked features is returned. Then a classifier is employed to compute the error rate.

The time complexity of the relevance and redundancy seen in Eq.14 are $O(m)$ respectively. Where $m$ is the number of selected features. Thus, the computational complexity of BCOA-MI is $O(m) + O(m) = O(m)$. On the other hand, the time complexity of the relevance and redundancy of gain ratio based entropy in 15 is $O(m)$ and $O(m^2)$ respectively. As such, the time complexity of the BCOA-E is $O(m) + O(m^2) = O(m^2)$. On the other hand, the binary search (BCOA) for both BCO-MI and BCOA-E runs in $O(n)$ time, where n is the population size. Therefore, BCOA-MI can complete its process within a shorter time in most cases compared to its BCOA-E counterpart. Hence BCOA-MI is computationally more efficient. It is merely because BCOA-MI interacts with the only pair of features in contrast to the BCOA-E that interacts with a group of features.

## B. BCOA FILTER-BASED MULTI-OBJECTIVE APPROACH

COA is developed originally to solve a single objective optimisation problem. So far, the use of COA for the multi-objective optimisation problem is quite scarce if at all exist in the literature, especially in feature selection. To solve MOP, developers may be attracted towards a set of Pareto optimum points as an alternative to a sole point. Meanwhile, GA works together with a populace of points, and it appears normal to apply GAs in a MOP to catch a few solutions simultaneously.

This section presents two different but related multi-objective optimisations algorithms using the idea of NSGAII and NSGAIII frameworks in BCOA. Which leads to BCNSG2 as well as BCNSG3 methods. In each of the proposed methods, MI and gain ratio based-entropy are added as the evaluation measures to have a total of four multi-objective filter-based algorithms (BCNSG2MI, BCNSG3MI, BCNSG2E and BCNSG3E. The detail of the algorithms is presented in the subsequent sections below.

### 1) BCNSG2MI AND BCNSG2E

The experiments on BCOA-MI and BCOA-E, clearly showed that both MI along with gain ratio based-entropy is an effective evaluation measure for filter-based FS. However, the weights employed in their fitness functions want to be pre-defined. Therefore, according to BCOA, we developed filter-based multi-objective FS using NSGAII along with MI (BCNSG2MI)and entropy (BCNSG2E) with the target of minimising the feature size and improving the greatest significant features with their class label to discover the Pareto front of the FS issue. The pseudocode for the BCNSG2MI and BCNSG2E is depicted in Algorithm 3.

COA, as well as its binary version, are initially meant to deal with the single-objective optimisation problem. The most significant task in spreading COA to multi-objective optimisation is to determine an outstanding environment of cuckoo for all habitat from the group of possible non-dominated solutions. Reference [13] introduced a popular multi-objective optimisation technique known as the NSGAII. Since then, researchers are driven to use it and solve problems related to multi-objective optimisation approaches. For instance, [51] used the concept of NSGAII with PSO to develop a multi-objective PSO based on the NSGAII. Then, [10] and [12] used that idea to solve filter-based multi-objective FS problems using BPSO. Similarly, GA is employed for a single-objective and multi-objective using PSO in the work of [38]. However, other EC-based techniques such as COA was reported to have faster convergence and performed better than many other ECs, yet its potential for multi-objective optimisation as well as feature selection is not fully investigated.

Therefore, in this study, a BCOA multi-objective framework for FS, according to NSGAII, was presented. Thus, two pairs of filter-based multi-objective FS algorithms are advanced, and that is BCNSG2MI and BCNSG2E. While BCNSG2MI use $Rel_{mi}$, BCNSG2E use $Rel_E$ to assess the significance or relevance between a pair of features with their target class.

**Algorithm 3** Proposed BCNSG2MI and BCNSG2E

```
Begin
    Divide the dataset into training set and test set;
    Initialise the habitat;
    Allow the cuckoos to lay their eggs in their matching ELR;
    Recognise the cuckoos in the nondominated solutions;

    Compute reference point of the cuckoos and generate initial population;
    Use nondominated population sorting mechanism;
    WHILE maximum iteration is not reached DO
    Evaluate the two fitness values of each cuckoo  feature size together with their relevance in $Rel_{mi}$ BCNSG2MI of Eq. 14 and $Rel_E$ in BCNSG2E of Eq. 15 on the training set*
        Identify the habitats (nonDomCOAList);
        Compute the crowding distance in nonDomCOAList and sort them;
for i=1 to PopulationSize DO
        Update the ith habitat;
        Select the best $habitat_i$ from the highest ranked solutions in nonDomCOAList;
        Update the ELR;
end
        Update the ith habitat;
Add, the original habitats and update the habitats in the unioun;
identify the different levels of nondominated fronts $Fit = (Fit_1, Fit_2, Fit_3, \ldots)$;
Empty the cuckoos for the next iteration;
 i=1;)
 WHILE $|habitat| < PopulationSize$
        if $(|habitat| + |Fit_i| <= PopulationSize)$ THEN
        Add $Fit_i$ to habitat;
        $i = i + 1$;
    END
    if $(|habitat| + |Fit_i| > PopulationSize)$ Then
        compute the crowding distance for each habitat in $F_i$;
        Sort the habitats in $Fit_i$;
        Add the $(PopulationSize - |habitat|)$ least crowded habitat to cuckoo;
        END WHILE
    Compute and store the error rate of the feature subsets (solutions)on the test set;
    Compute the error rate of the solutions or feature subsets in $Fit_i$ on the test set;
    Return the feature subsets along with their classifier error rates in $Fit_i$;
END
```

The detailed of how the multi-objective filter-based algorithm (BCNSG2MI and BCNSG2E) works is depicted in Fig. 1. The core target is to used nondominated sorting of Phase VII to choose the best cuckoo for all habitat and amend the nonDomCOAList in the evolutionary process. As a

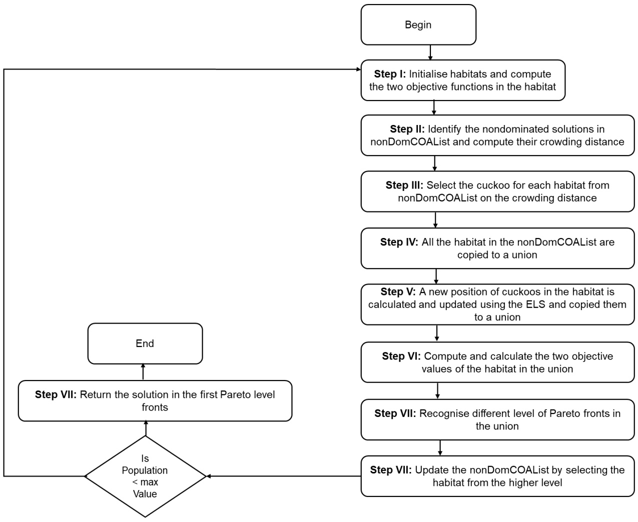

Begin
Step I: Initialise habitats and compute the two objective functions in the habitat
Step II: Identify the nondominated solutions in nonDomCOAList and compute their crowding distance
Step III: Select the cuckoo for each habitat from nonDomCOAList on the crowding distance
Step IV: All the habitat in the nonDomCOAList are copied to a union
Step V: A new position of cuckoos in the habitat is calculated and updated using the ELS and copied them to a union
Step VI: Compute and calculate the two objective values of the habitat in the union
Step VII: Recognise different level of Pareto fronts in the union
Step VII: Update the nonDomCOAList by selecting the habitat from the higher level
Is Population < max Value
Step VII: Return the solution in the first Pareto level fronts
End

**FIGURE 1.** Flowchart of the Multi-objective BCNSG2MI and BCNSG2E.

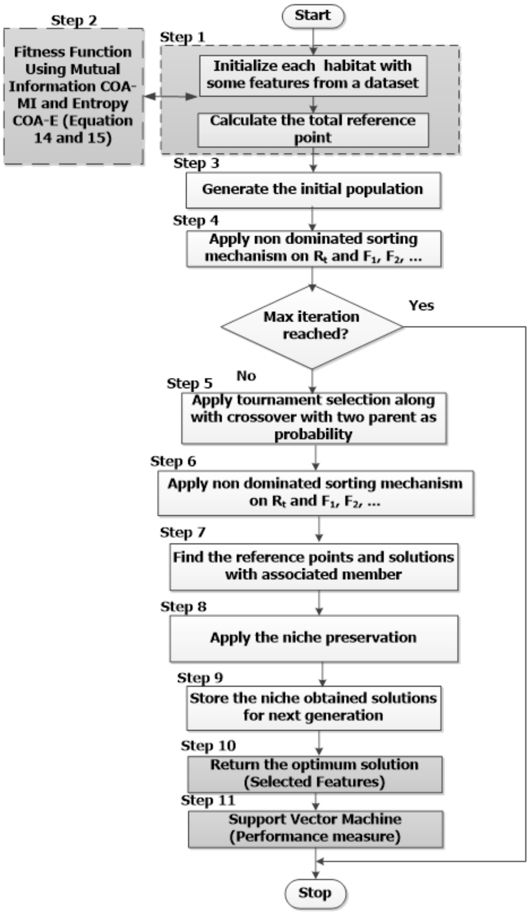

Start
Step 1
Initialize each habitat with some features from a dataset
Calculate the total reference point
Step 2
Fitness Function Using Mutual Information COA-MI and Entropy COA-E (Equation 14 and 15)
Step 3
Generate the initial population
Step 4
Apply non dominated sorting mechanism on R_t and F_1, F_2, ...
Max iteration reached?
Yes
No
Step 5
Apply tournament selection along with crossover with two parent as probability
Step 6
Apply non dominated sorting mechanism on R_t and F_1, F_2, ...
Step 7
Find the reference points and solutions with associated member
Step 8
Apply the niche preservation
Step 9
Store the niche obtained solutions for next generation
Step 10
Return the optimum solution (Selected Features)
Step 11
Support Vector Machine (Performance measure)
Stop

**FIGURE 2.** Flowchart of the Multi-objective BCNSG3MI and BCNSG3E.

display in Fig. 1, during every repetition, the algorithms start by identifying the nondominated features in the nonDomCOAList and compute the crowding distance, and all the nondominated feature subsets are arranged based on the crowding distance in Phase II. While in Phase III, a random cuckoo is chosen from the smallest crowded solutions, which is the uppermost graded part of the sorted nondominated solutions. All the habitats in the nonDomCOAList are copied to a union in Phase IV. After determining the best habitat where cuckoo lives, a new position for the next cuckoos' habitat is calculated according to Phases in Eq.(5) and Eq.(6) moreover, is added into the union in Phase V. In Phase VI, the two objective functions of the habitat are assessed where the relevance is assessed by $Rel_{mi}$ in BCNSG2MI and $Rel_E$ in BCNSG2E.

The nondominated sorting procedure is shown in Phase VII. Precisely, the nondominated solutions in the union are named the initial nondominated front and are afterwards removed out of the union. Next, the nondominated features in the remaining union are termed the second nondominated front, and it continues like that. The subsequent stages of the nondominated fronts are recognised by reiterating this process. Finally, Phase VIII displays the procedure of altering nonDomCOAList for the resulting repetition. Precisely, habitats are chosen from the top points of the nondominated fronts, beginning with the initial front and so on. If the solutions required is more than the features or solutions that remain in the present nondominated front, the complete solutions are joined into the next repetition. Phase II, until Phase VIII is repetitive until the end condition, is satisfied. Then, the proposed algorithm recovers the initial nondominated Pareto front in the union.

### 2) BCNSG3MI AND BCNSG3E

In the previous subsection, NSGAII was used along with filter-based BCOA for multi-objective FS. Although NSGAII performed well with both PSO, GA, and even the BCOA, However, it lacks some reference point; instead, it used the crowding distance and mutation operators for its computation. Moreover, a crowded comparison can restrict the convergence of NSGAII. Considering these limitations [50] proposed a more robust NSGAIII.

In contrast to the NSGAII, the maintenance of diversity among population members in NSGAIII is supported by providing an adaptively amending several well-spread reference points. As such, another multi-objective BCOA for filter-based FS using the concepts of nondominated sorting in NSGAIII is also presented. Based on these, other pairs of filter-based multi-objective FS algorithms are developed BCNSG3MI and BCNSG3E. Both BCNSG3MI and BCNSG3E used (14) and (15) respectively.

From Fig. 2, the proposed multi-objective BCNSG3MI and BCNSG3E is made up of eleven related steps. The focal impression is to use the nondominated sorting of NSGAIII in BCOA to select the best cuckoo environment for feature selection. At the end of each iteration, the proposed algorithm performs Step I to IV. Step I initialise each habitat with some features from a dataset, and the total reference point of the features are computed. In Step II, the fitness evaluation function of the proposed approach is calculated for both MI and gain ratio based-entropy using (14) and (15), respectively. The relevancy is evaluated by using the $Rel_{mi}$ and $Rel_E$. Step III generates the initial population using the idea of the COA and BCOA. After that, in Step IV, the nondominated population sorting mechanism is employed. This identifies the various levels of the Pareto fronts in the union. If the maximum iteration is not reached, it continues to the next stage. The initial iteration is always set to zero. Thus, it must proceed to the next stage at the beginning. Step V used the tournament selection and crossover with two parents as a probability. Then another Step IV is repeated in Step VI, while Step VII find the reference points and solutions with the associated member. Step VIII apply the niche preservation, and Step IX stores the niche obtained solutions for the next generation using the BCOA concept in (5) and (6). Step X returns the optimum solution of the selected features. Finally, in Step XI a classifier is employed to measure the error rate of chosen features. Step V- VIII is repetitive until the highest number of repetitions is gotten. The detailed of the pseudocode is shown in Algorithm 4.

**Algorithm 4** Proposed BCNSG3MI and BCNSG3E

```
Begin
    Divide the dataset into training set and test set;
    Initialise the habitat;
    Evaluate the two fitness values of each cuckoo feature size together with their relevance in $Red_{mi}$ BCNSG3MI of Eq. 14 and $Red_E$ in BCNSG3E of Eq. 15 on the training set*
    Allow the cuckoos to lay their eggs in their matching ELR;
    Recognise the cuckoos in the nondominated solutions;

    Compute reference point of the cuckoos and generate initial population;
    Use nondominated population sorting mechanism;
    WHILE maximum iteration is not reached DO
        Apply tournament selection and crossover with two parents as probability;
        Again, apply nondominated population sorting mechanism on the cuckoos;
        Apply normalization on the population;
        Find out reference points and solution with associated member based on associate procedure;
        Apply the niche preservation (niche procedure);
        Keep the niche obtained solutions for the next generations;
    END WHILE
    Compute and store the error rate of the feature subsets (solutions)on the test set;
    Return the feature subsets along with their classifier error rates;
End
```

Both BCNSG2MI, along with BCNSG3MI, can complete the FS in a shorter time than BCNSG2E and BCNSG3E in all the datasets. The time complexity of the fitness function seen in Eq. 14 for the BCNSG2MI, as well as BCNSG3MI, is $O(m)$ where $m$ is the number of selected of features. Alternatively, the time complexity of the fitness function seen in Eq. 15 for both BCNSG2E and BCNSG3E are $O(m) + O(m^2) = O(m^2)$. Besides, the binary search (BCOA) runs in $O(n)$ time, where $n$ is the population size. NSGAII and NSGAIII results in a similar computational complexity of $O(log_2 n)$. Thus, the total computational complexity for the algorithms BCNSG3MI and BCNSG2MI are $O(m) + O(n) + O(log_2 n)$. Whereas, that of BCNSG3E and BCNSG2E is $O(m^2) + O(n) + O(log_2 n)$. Also, the use of NSGAII and NSGAIII make the computation complex due to the nondominated sorting and external archive. However, NSGAIII is computationally fair than the NSGAII since it has a more concise way to renew or select individuals as well as the use of the reference points.

## IV. EXPERIMENTS

### A. DATASETS

The standard datasets employed in this experiment is display in Table 1. The datasets are obtained from the popular repository in [52]. It contains a different feature size, instances and classes of varying degree of difficulties. For example, Lymphography dataset takes the smallest size of both features and instances, whereas, Madelon takes the maximum feature size and Coil2000 with the maximum number of instances.

While conducting the experiments, the instances of all the datasets are separated randomly into training and testing test. While the training test takes 70% of the instances whereas tests take 30%. The planned algorithms run on the training test first to choose the subsets of features and later, the error rate of the chosen features is computed on the test set using the classification algorithm. There are quite varieties of classification algorithms such as SVM, KNN, GNB and DT, among others. In this paper, SVM is chosen because of its popularity and proven records in computing classification accuracies in different researches.

The SVM computes the error rate of the nominated features in the multi-objective approach using the Eq. 16 below.

$$Error \quad rate = \frac{(FP + FN)}{(TP + TN + FP + FN)} \tag{16}$$

The termed $TP$, $TN$, $FP$, $FN$ represents true positives, true negatives, false positives and false negatives correspondingly.

#### 1) EXPERIMENTAL PARAMETER SETTINGS

The parameter settings used for the proposed BCOA-MI, BCOA-E, BCNSG2MI, BCNSG2E, BCNSG3MI and BCNSG3E algorithms are chosen based on the work of [45]; [16] where both the initial and upper population are set to five and twenty respectively. Besides all the proposed algorithms are run 40 separated times on all the dataset.

In the single objective filter-based approaches, both BCOA-MI and BCOA-E used five different values of $\beta_1$

**TABLE 1. Datasets analysis.**

| S/N | Datasets | No. of Features | No. of Instances |
|---|---|---|---|
| 1 | Lymphography | 18 | 148 |
| 2 | SpectEW | 22 | 267 |
| 3 | Leddisplay | 24 | 1000 |
| 4 | Dermatology | 34 | 366 |
| 5 | Soyabeans Large | 35 | 307 |
| 6 | Chess | 36 | 3196 |
| 7 | Connect4 | 42 | 3196 |
| 8 | Promoter | 57 | 106 |
| 9 | Splice | 60 | 3190 |
| 10 | Optic | 64 | 5620 |
| 11 | Audiology | 68 | 226 |
| 12 | Coil2000 | 85 | 9000 |
| 13 | DNA | 180 | 3186 |
| 14 | Madelon | 500 | 2600 |

and $\beta_2$ (0.9, 0.8, 0.75, 0.6 and 0.5) in the experiments for each dataset. Where $\beta_1$ is for the BCOA-MI and $\beta_2$ for the gain ratio based entropy. In addition to that, the Wilcoxon Rank Sum test was conducted on the BCOA-MI and BCOA-E, whereby 0.05 was employed as the level of significance, to confirm the significant change between the methods on different values of $\beta$ compared to the full-length features. If the p-value $>=$ 0.05, then our proposed method significantly outperformed the full-length features at 95% of the level of guarantee.

Based on the work [10], [49] and [12], both BCNSG2 (BCNSG2MI and BCNSG2E) and BCNSG3 (BCNSG3MI and BCNSG3E) used $1/n$ mutation rate. Where $n$ is the maximum feature size in each of the datasets. Also, cross over probability is set to 0.5. The reference point for the BCNSG3 was set to 15 based on the work of [50].

The multi-objective algorithms (BCNSG2 and BCNSG3) obtain a set of nondominated solutions in all runs. The 40 sets of solutions attained by all the multi-objective algorithms are united into a single union set. The union set contains the subsets of features such as the feature size and their respective error rate. Thus, the set of average solution (named Pareto front) is gotten through the mean of the classification error and the matching number of features. Apart from the average Pareto front, the nondominated solutions inside the union set too are offered in the subsequent segment.

## V. RESULTS AND DISCUSSION

This segment presents the outcomes of the experiments conducted. Tables 2 and 3, displays the results of the filter-based single-objective (BCOA-MI and BCOA-E) with changing weights in their respective fitness functions. Similarly, Figures 3 and 4 displayed the results of the multi-objective filter-based methods as well as the comparison between NSGAII (BCNSG2MI and BCNSG2E) and NSGAIII (BCNSG3MI and BCNSG3E) based algorithms.

### A. RESULTS OF THE SINGLE-OBJECTIVE FILTER-BASED APPROACH BCOA-MI AND BCOA-E

The experimental outcomes are made known in Tables 2 and 3. From the tables, "Ave Size" speak for the mean of nominated features by all the algorithms in the 40 separate runs. Also, "Ave-Acc" along with "Best-Acc" serves as the mean accuracy and best accuracy respectively. "Std Dev" is the standard deviation for the 40 error rates tests. The outcome of the Wilcoxon Rank Test is denoted as "Sig Test" whereby a "+" or "−" symbolise that the classification performance of BCOA-MI or BCOA-E is good or poor than the full-length features, while "=" serve as the same classification performance.

Generally, it can be seen clearly from the outcomes that BCOA-MI achieved considerable well on the average size of selected features in the whole datasets, whereby nearly 75% of the whole feature size is minimised. Unlike BCOA-E, which done well on accuracy. It disclosed that both BCOA-MI and BCOA-E possibly would meaningfully minimise the feature size and accomplish the same or improve classification performance compared to full-length features.

Looking at Tables 2 and 3, it can be detected that the higher the values of $\beta_1$ and $\beta_2$ the better the accuracy for each datasets. If the values of $\beta$ are bigger, then the relevance is greater than the redundancy that leads to high accuracy. Nevertheless, sometimes if the difference negligible or similar. For example, looking at Lymphography dataset in Table 2, when$\beta_1 = 0.9$ and $\beta_1 = 0.8$ the best classification error rate are 14.00% and 16.00% respectively. Unlike in Dermatology dataset in Table 3, where $\beta_2 = 0.9$ and $\beta_2 = 0.8$ and the best classification error rate remains as 2.20% respectively. In either case, the number of features is minimised to the lowest level, almost around 60-70%.

On the other hand, the results in Tables 2 and 3 also shows that the higher the values of $\beta_1$ and $\beta_2$ the higher the size of chosen features. The decrease in the feature size is around 40% of the whole feature size. Moreover, the accuracy improves if both the values of $\beta_1$ and $\beta_2$ increase in all the datasets. The outcomes specified that several weight values might automatically inspire the goodness of the classifier specifically, those with smaller subsets of features compared to the full-length features.

Relating the performance of Tables 2 and 3, it is clear that $\beta_2$ performed well in terms of the error rate on each dataset as to $\beta_1$ and poor on chosen features and possibly the longest time in computation. Though, $\beta_1$ did well on chosen features along with the longer computational period owing to the only duo of features its works with which makes it faster in terms of computations.

In whichever way, it can be seen that engaging both $\beta_1$ and $\beta_2$ with suitable standards as fitness functions can derive a fewer number of features with improved classification performance compared to the full-length features. Therefore, BCOA-MI and BCOA-E with $\beta_1$ and $\beta_2$ values of 0.5 and 0.9 were employed for contrasting and evaluation in the

**TABLE 2.** Results of BCOA-MI with $\beta_1$.

| Datasets | B1 | Ave Size | Av Acc | Best Acc | Std Dev | Sig. Test |
|---|---|---|---|---|---|---|
| | | BCOA-MI | BCOA-MI | BCOA-MI | BCOA-MI | BCOA-MI |
| Lymphography | All | 18 | | 12.50 | | |
| | 0.9 | 7.8 | 14.00 | 11.20 | 0.013 | + |
| | 0.8 | 5.2 | 16.00 | 15.00 | 0.013 | - |
| | 0.75 | 4.9 | 16.60 | 16.60 | 0.000 | - |
| | 0.6 | 4.1 | 20.00 | 20.00 | 0.000 | - |
| | 0.5 | 3 | 22.00 | 20.10 | 0.001 | - |
| Spect | All | 22 | | 14.90 | | |
| | 0.9 | 9.2 | 11.20 | 10.60 | 0.012 | + |
| | 0.8 | 7.8 | 12.90 | 11.50 | 0.012 | + |
| | 0.75 | 5.6 | 15.60 | 14.50 | 0.001 | + |
| | 0.6 | 4.2 | 16.70 | 16.00 | 0.001 | - |
| | 0.5 | 4 | 17.00 | 17.00 | 0.000 | - |
| Leddisplay | All | 24 | | 0.00 | | |
| | 0.9 | 19.2 | 0.00 | 0.00 | 0.000 | = |
| | 0.8 | 18.6 | 0.00 | 0.00 | 0.000 | = |
| | 0.75 | 16.4 | 0.00 | 0.00 | 0.000 | = |
| | 0.6 | 13.9 | 0.00 | 0.00 | 0.000 | = |
| | 0.5 | 11.4 | 0.00 | 0.00 | 0.000 | = |
| Dermatology | All | 33 | | 5.25 | | |
| | 0.9 | 26.4 | 5.45 | 3.45 | 0.004 | + |
| | 0.8 | 16.2 | 3.81 | 3.12 | 0.004 | + |
| | 0.75 | 10.4 | 3.99 | 3.99 | 0.078 | + |
| | 0.6 | 8 | 4.12 | 4.12 | 0.089 | + |
| | 0.5 | 6 | 5.15 | 4.86 | 0.101 | + |
| Soybean Large | All | 35 | | 8.67 | | |
| | 0.9 | 21 | 7.78 | 4.23 | 0.000 | + |
| | 0.8 | 12 | 8.35 | 4.23 | 0.000 | + |
| | 0.75 | 8.8 | 8.35 | 6.35 | 0.079 | + |
| | 0.6 | 6.9 | 9.89 | 6.85 | 0.015 | + |
| | 0.5 | 5.1 | 9.89 | 7.00 | 0.015 | + |
| Chess (KrvskpEW) | All | 36 | | 10.80 | | |
| | 0.9 | 17.2 | 5.80 | 5.40 | 0.001 | + |
| | 0.8 | 16.7 | 6.50 | 6.00 | 0.002 | + |
| | 0.75 | 15.2 | 7.00 | 6.30 | 0.002 | + |
| | 0.6 | 14.2 | 7.60 | 7.50 | 0.001 | + |
| | 0.5 | 12.2 | 8.00 | 7.70 | 0.001 | + |
| Connect4 | All | 42 | | 25.00 | | |
| | 0.9 | 9 | 26.90 | 25.00 | 0.105 | = |
| | 0.8 | 8 | 26.90 | 25.00 | 0.205 | = |
| | 0.75 | 6.9 | 28.60 | 25.20 | 0.195 | - |
| | 0.6 | 5 | 29.00 | 25.90 | 0.185 | - |
| | 0.5 | 5 | 30.00 | 26.00 | 0.199 | - |
| Promoter | All | 57 | | 10.00 | | |
| | 0.9 | 42.8 | 7.55 | 7.55 | 0.015 | + |
| | 0.8 | 37.5 | 8.80 | 8.00 | 0.015 | + |
| | 0.75 | 34.7 | 8.80 | 8.00 | 0.015 | + |
| | 0.6 | 25.5 | 8.80 | 8.80 | 0.120 | + |
| | 0.5 | 19.5 | 11.02 | 10.50 | 0.125 | - |
| Splice | All | 60 | | 30.12 | | |
| | 0.9 | 9.46 | 29.75 | 22.45 | 0.478 | + |
| | 0.8 | 12.5 | 30.12 | 27.50 | 0.429 | + |
| | 0.75 | 10.5 | 30.15 | 28.20 | 0.347 | + |
| | 0.6 | 11.2 | 30.12 | 29.50 | 0.289 | + |
| | 0.5 | 11 | 28.75 | 23.02 | 0.395 | + |
| Optic | All | 64 | | 1.12 | | |
| | 0.9 | 13.3 | 1.12 | 1.12 | 0.362 | = |
| | 0.8 | 15.2 | 3.00 | 3.00 | 0.289 | - |
| | 0.75 | 16 | 6.45 | 4.45 | 0.245 | - |
| | 0.6 | 17.9 | 2.25 | 2.25 | 0.456 | - |
| | 0.5 | 18.1 | 2.11 | 1.33 | 0.262 | - |
| Audiology | All | 68 | | 34.44 | | |
| | 0.9 | 20.2 | 29.50 | 27.50 | 0.270 | + |
| | 0.8 | 21.4 | 30.00 | 28.00 | 0.289 | + |
| | 0.75 | 22.2 | 29.80 | 28.00 | 0.316 | + |
| | 0.6 | 21.8 | 30.45 | 28.45 | 0.471 | + |
| | 0.5 | 18.54 | 30.02 | 28.34 | 0.418 | + |
| Coil2000 | All | 85 | | 5.78 | | |
| | 0.9 | 30.1 | 7.44 | 5.78 | 0.145 | = |
| | 0.8 | 30.1 | 7.44 | 7.44 | 0.242 | - |
| | 0.75 | 32.2 | 7.80 | 6.40 | 0.178 | - |
| | 0.6 | 25.6 | 9.00 | 9.00 | 0.129 | - |
| | 0.5 | 19.09 | 8.02 | 6.01 | 0.167 | - |
| DNA | All | 180 | | 17.22 | | |
| | 0.9 | 56.5 | 18.00 | 16.80 | 0.250 | + |
| | 0.8 | 57.2 | 17.50 | 16.80 | 0.011 | + |
| | 0.75 | 60.1 | 17.50 | 17.02 | 0.016 | + |
| | 0.6 | 55.6 | 12.50 | 15.35 | 0.000 | + |
| | 0.5 | 52.55 | 11.45 | 11.45 | 0.125 | + |
| Madelon | All | 500 | | 20.00 | | |
| | 0.9 | 202.5 | 19.12 | 19.12 | 0.000 | + |
| | 0.8 | 189.4 | 19.56 | 19.56 | 0.000 | + |
| | 0.75 | 175.6 | 20.00 | 20.00 | 0.000 | = |
| | 0.6 | 201.5 | 20.50 | 20.50 | 0.001 | - |
| | 0.5 | 185.85 | 21.02 | 20.02 | 0.002 | - |

**TABLE 3.** Results of BCOA-E with $\beta_2$.

| Datasets | $\beta_1$ | Ave Size BCOA-E | Av Acc BCOA-E | Best Acc BCOA-E | Std Dev BCOA-E | Sig. Test BCOA-E |
|---|---|---|---|---|---|---|
| **Lymphography** | All | 18 | | 12.50 | | |
| | 0.9 | 12.6 | 11.00 | 11.00 | 0.000 | + |
| | 0.8 | 10.5 | 12.00 | 11.20 | 0.000 | + |
| | 0.75 | 8.9 | 12.60 | 12.10 | 0.001 | + |
| | 0.6 | 6.4 | 14.00 | 12.80 | 0.001 | - |
| | 0.5 | 5.1 | 14.50 | 14.50 | 0.001 | - |
| **Spect** | All | 22 | | 9.60 | | |
| | 0.9 | 10.2 | 10.10 | 8.60 | 0.004 | + |
| | 0.8 | 8.7 | 10.90 | 10.50 | 0.001 | - |
| | 0.75 | 6.8 | 11.60 | 11.10 | 0.001 | - |
| | 0.6 | 5.2 | 12.90 | 12.00 | 0.002 | - |
| | 0.5 | 5 | 13.80 | 13.80 | 0.001 | - |
| **Leddisplay** | All | 24 | | 0.00 | | |
| | 0.9 | 9 | 0.00 | 0.00 | 0.000 | = |
| | 0.8 | 9 | 0.00 | 0.00 | 0.000 | = |
| | 0.75 | 7 | 0.00 | 0.00 | 0.000 | = |
| | 0.6 | 7 | 0.00 | 0.00 | 0.000 | = |
| | 0.5 | 7 | 0.00 | 0.00 | 0.000 | = |
| **Dermatology** | All | 33 | | 2.10 | | |
| | 0.9 | 9.2 | 7.50 | 2.20 | 0.040 | + |
| | 0.8 | 8 | 8.00 | 2.20 | 0.050 | + |
| | 0.75 | 7.2 | 8.25 | 3.10 | 0.099 | - |
| | 0.6 | 6.2 | 8.85 | 3.60 | 0.105 | - |
| | 0.5 | 5.7 | 10.00 | 5.00 | 0.019 | - |
| **Soybean Large** | All | 35 | | 8.67 | | |
| | 0.9 | 19.4 | 8.50 | 5.00 | 0.057 | + |
| | 0.8 | 18.2 | 9.60 | 5.00 | 0.057 | + |
| | 0.75 | 16.8 | 10.20 | 5.00 | 0.057 | + |
| | 0.6 | 15.4 | 10.90 | 5.50 | 0.099 | + |
| | 0.5 | 13.2 | 11.50 | 5.90 | 0.107 | + |
| **Chess (KrvskpEW)** | All | 36 | | 10.80 | | |
| | 0.9 | 19.2 | 2.80 | 2.40 | 0.001 | + |
| | 0.8 | 18.4 | 3.50 | 2.00 | 0.005 | + |
| | 0.75 | 16.3 | 5.00 | 2.30 | 0.005 | + |
| | 0.6 | 15.4 | 5.60 | 5.50 | 0.001 | + |
| | 0.5 | 13.9 | 7.10 | 6.70 | 0.001 | + |
| **Connect4** | All | 42 | | 21.10 | | |
| | 0.9 | 27.6 | 17.44 | 11.50 | 0.004 | + |
| | 0.8 | 24.5 | 20.60 | 18.50 | 0.003 | + |
| | 0.75 | 21.2 | 21.40 | 20.50 | 0.003 | + |
| | 0.6 | 20.1 | 23.10 | 22.10 | 0.001 | - |
| | 0.5 | 19.2 | 24.00 | 23.80 | 0.000 | - |
| **Promoter** | All | 57 | | 90.00 | | |
| | 0.9 | 25 | 92.45 | 93.00 | 0.010 | + |
| | 0.8 | 21.5 | 92.20 | 92.20 | 0.001 | + |
| | 0.75 | 17.5 | 92.20 | 92.20 | 0.011 | + |
| | 0.6 | 15.5 | 93.00 | 93.00 | 0.001 | + |
| | 0.5 | 8.5 | 93.45 | 94.00 | 0.000 | + |
| **Splice** | All | 60 | | 10.00 | | |
| | 0.9 | 9.2 | 7.55 | 7.00 | 0.012 | + |
| | 0.8 | 9.1 | 7.80 | 7.80 | 0.210 | + |
| | 0.75 | 8.8 | 7.80 | 7.80 | 0.092 | + |
| | 0.6 | 8.2 | 7.00 | 7.00 | 0.089 | + |
| | 0.5 | 7.51 | 6.55 | 6.00 | 0.000 | + |
| **Optic** | All | 64 | | 1.12 | | |
| | 0.9 | 12.4 | 1.02 | 1.02 | 0.215 | + |
| | 0.8 | 14.6 | 2.45 | 2.00 | 0.350 | - |
| | 0.75 | 15.2 | 4.40 | 3.45 | 0.250 | - |
| | 0.6 | 11.2 | 1.80 | 1.25 | 0.650 | - |
| | 0.5 | 10.9 | 0.90 | 0.00 | 0.987 | + |
| **Audiology** | All | 68 | | 34.44 | | |
| | 0.9 | 19.2 | 28.80 | 26.80 | 0.210 | + |
| | 0.8 | 19 | 29.11 | 27.15 | 0.250 | + |
| | 0.75 | 19 | 29.02 | 28.00 | 0.123 | + |
| | 0.6 | 18.2 | 30.00 | 26.50 | 0.326 | + |
| | 0.5 | 17.2 | 25.60 | 24.60 | 0.289 | + |
| **Coil2000** | All | 85 | | 5.78 | | |
| | 0.9 | 18 | 6.80 | 5.78 | 0.122 | = |
| | 0.8 | 17.2 | 6.50 | 5.15 | 0.185 | + |
| | 0.75 | 18.5 | 6.00 | 6.15 | 0.174 | - |
| | 0.6 | 18.2 | 5.80 | 4.40 | 0.000 | + |
| | 0.5 | 17.5 | 4.50 | 3.90 | 0.011 | + |
| **DNA** | All | 180 | | 17.22 | | |
| | 0.9 | 50.2 | 17.50 | 15.65 | 0.150 | + |
| | 0.8 | 49.5 | 17.22 | 15.35 | 0.460 | + |
| | 0.75 | 48.6 | 16.80 | 14.80 | 0.325 | + |
| | 0.6 | 47.2 | 14.35 | 13.90 | 0.095 | + |
| | 0.5 | 45.2 | 12.80 | 11.50 | 0.158 | + |
| **Madelon** | All | 500 | | 20.00 | | |
| | 0.9 | 185.5 | 19.00 | 19.00 | 0.000 | + |
| | 0.8 | 150.6 | 18.44 | 17.44 | 0.099 | + |
| | 0.75 | 120.4 | 18.00 | 18.00 | 0.001 | + |
| | 0.6 | 118.6 | 17.89 | 17.00 | 0.018 | + |
| | 0.5 | 104.65 | 17.45 | 15.50 | 0.055 | + |

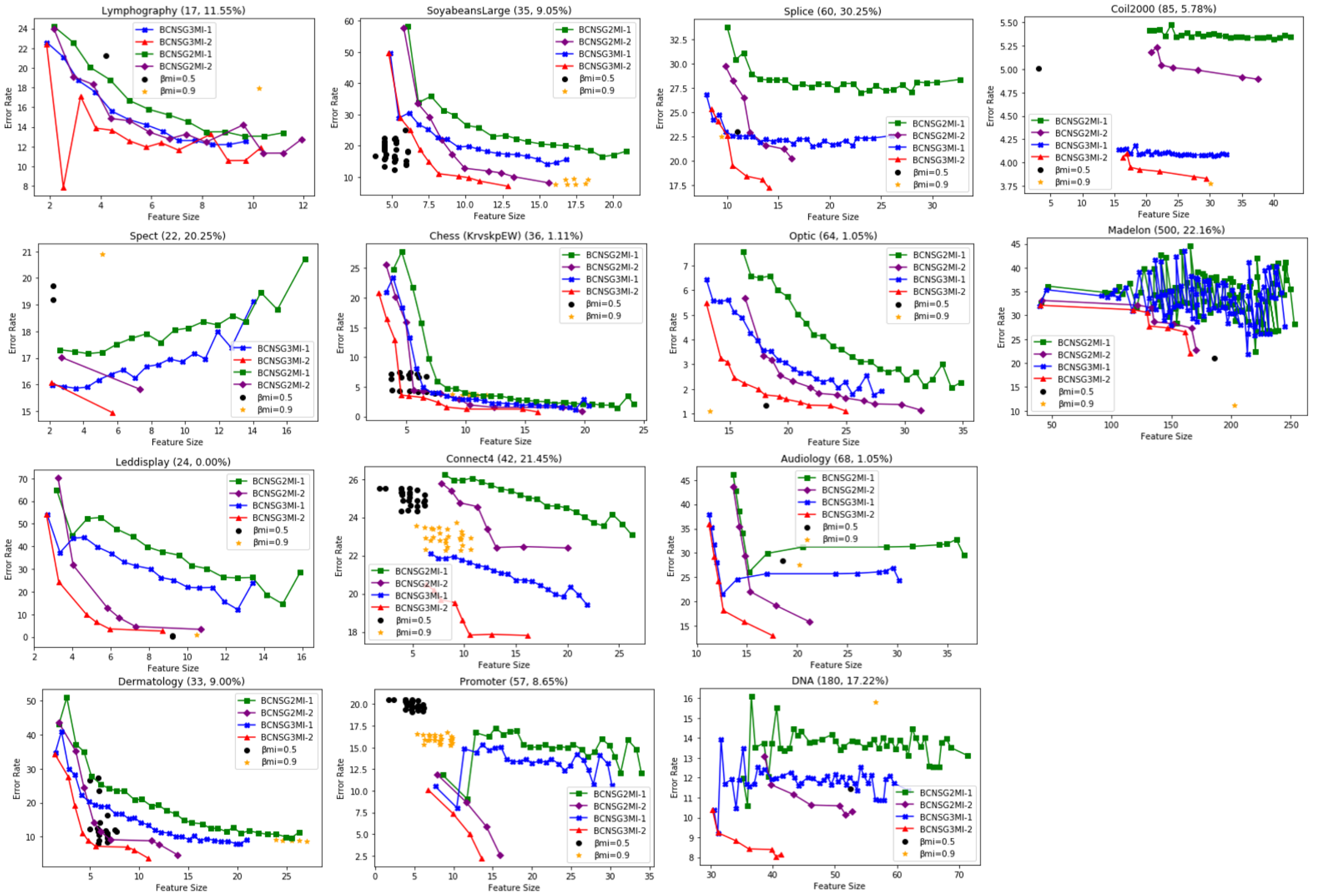


**FIGURE 3.** Experimental Results of BCOA-MI, BCNSG2MI and BCNSG3MI.

following segment to investigate the goodness of the proposed filter-based multi-objective FS algorithms.

## *B. RESULTS OF THE MULTI-OBJECTIVE FILTER-BASED APPROACH*

The experimental results of the filter-based BCOA-MI and BCOA-E with different weights show that its useful criteria for the filter-based FS. Nonetheless, the values of weights assigned in the fitness functions of the BCOA-MI and BCOA-E needs to be defined. In this segment, a filter-based multi-objective FS using the same concepts of MI and entropy is proposed. The main objectives are to minimise feature size and consequently improve the relevance amongst the features and their target class label. By so doing, it is expected to discover the Pareto front during the FS processes.

The results obtained from the experiments of BCNSG2, BCNSG3 and BCOA are depicted in Fig 3 and 4. At the top middle of each of the graph is the title of the dataset and inside the bracket is the entire feature size followed by the error rate of the SVM classifier used on all the features. Like every other graph, the x-axis displays the feature size, whereas the y-axis displays the error rate of the SVM classifier. The legend in each of the charts contains three elements whereas the first two that end with "$-1$" and "$-2$" represents average non-dominated solutions and Pareto front for the 40 independent runs respectively. The last element in the legend is BCOA-MI with either $\beta mi = 0.5$ or $\beta mi = 0.9$ and BCOA-E with either $\beta E = 0.5$ or $\beta E = 0.9$ which represents the 40 solutions achieved by the single objective filter-based feature selection algorithms with both MI and entropy.

The results of BCOA-MI and BCOA-E shows that some of the datasets evolve the same feature size in different runs as depicted in the graph. Despite the 40 independent runs applied, there are less than 40 distinct points shown in each of the charts. Similarly, the set of nondominated solutions ("$-2$") may have the same subsets of features that are revealed at the matching point in the graph.

### 1) RESULTS OF BCNSG3MI AND BCNSG2MI

The results obtained by the Pareto front solutions of BCNSG2MI and BCNSG3MI in the filter-based FS real objective space is shown clearly in Fig 3, where MI is employing as the evaluation measures. It's well known that in any multi-objective filter-based FS methods, the goodness of the Pareto front features is assessed by its error rate on the hidden test data. The same is applied to these experiments. Thus,

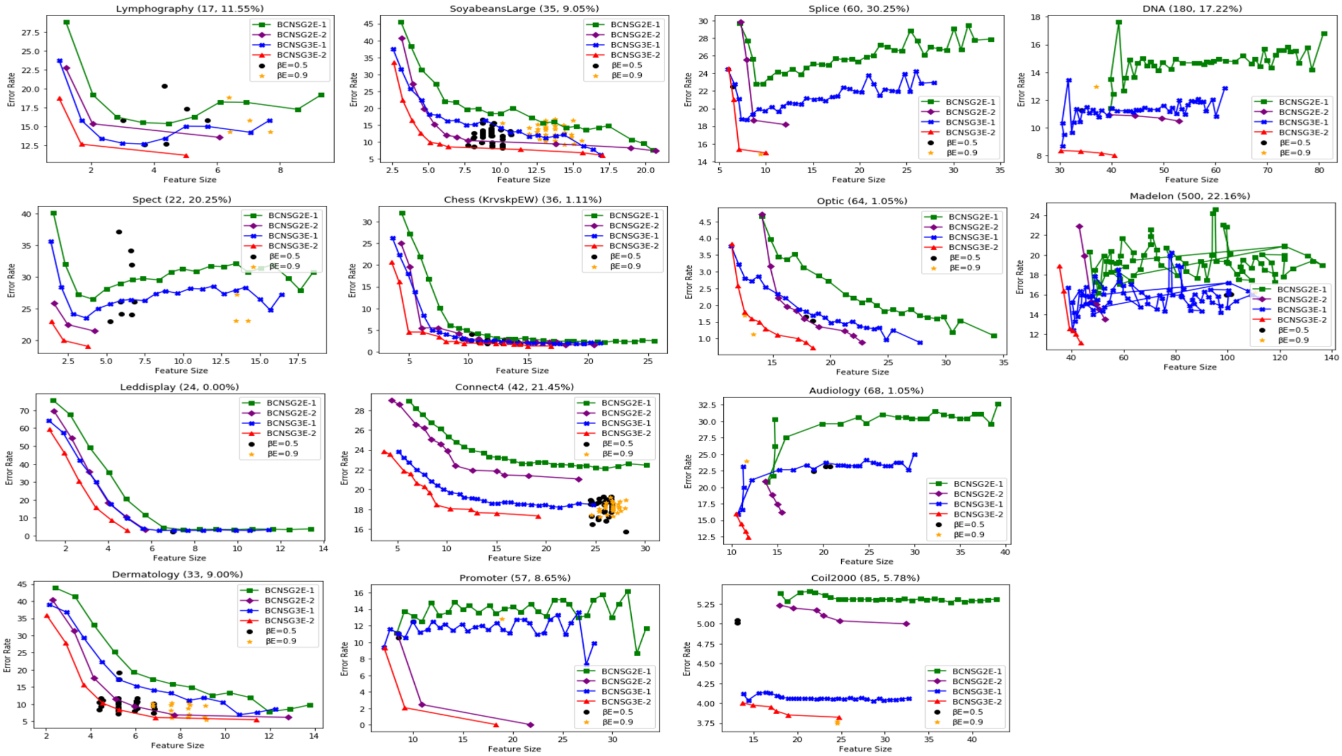


**FIGURE 4.** Experimental Results of BCOA-E, BCNSG2E and BCNSG3E.

the solution used in Fig. 3 is the Pareto solutions obtained in the MI space. Nevertheless, the error rate display in the charts was evaluated using SVM on the test data.

Besides, Fig 3 compared the results obtained by the BCNSG2MI, BCNSG3MI and BCOA-MI with $\beta mi = 0.5$ and $\beta mi = 0.9$, that use MI to assess the relevancy as well as the redundancy amongst a couple of features. Both BCOA-MI and BCOA-E may evolve similar subsets of features in different runs on some of the datasets, besides they are revealed at a similar point in the graph. Even though 40 results have been offered, most likely there will be not up to 40 separate points display in the graph. For example, both the BCNSG2MI and BCNSG3MI nondominated solutions possibly will have an identical subset of features and are displayed in the same point in the chart. It is like the charts in Fig 4.

### a: RESULTS OF BCNSG3MI

Results displayed in Fig. 3 indicated that BCOA-MI was able to reduce about 70% of the total feature size in almost all the datasets. Similarly, the classification error is mostly moderate and low on the Splice, Leddisplay, Chess (KrvskpEW), Optic, Audiology, Dermatology and Madelon datasets while is quite high on Connect4, Promoter and Spect datasets. The termed BCNSG3MI-2 means the average Pareto front whereas as BCNSG3MI-1 represents the nondominated features served from the 40 separate runs mentioned earlier.

In BCNSG3MI-1 the graphs showed that the nondominated features contain greater than or equal to one subset of features that choose almost halved of the total features and yet accomplish a minimum error rate in comparison to the full-length features. A typical example can be seen in DNA dataset, where a single nondominated solution carefully chosen 58 features out of the 180 full features. Besides, the error rate was diminished drastically from 17.22% to 10.75%. This can be seen on the graphs of the other datasets as well.

The graph in BCNSG3MI-2, shows that there are two or more solutions that chose fewer features and yet attained a minimum error rate as to the full-length features. In most cases, for equal feature size, there exist a various combination of features with different error rate. As such, the subset of features obtained in different runs may have different error rate for the feature size. Thus, some of the solutions in the average Pareto front will likely dominate others, even though the solutions obtained in all run are nondominated.

The results indicate that BCNSG3MI is a multi-objective algorithm that would spontaneously derive a subset of features that can decrease the feature size and consequently enhance the goodness of the classifier.

BCNSG3MI performed better than BCOA-MI in the majority of the datasets on the classification error rate. However, despite the fewer features size recorded by BCOA-MI with $\beta mi = 0.5$ and $\beta mi = 0.9$ than BCNSG3MI-1 on some few datasets, the majority of the solutions in BCNSG3MI-2 choose the fewer number of features and yet achieved an improved performance. Therefore, comparisons proved that using MI as the evaluation measures, the planned filter-based multi-objective FS (BCNSGMI) outperformed the filter-based single objective feature selection BCOA-MI with both $\beta mi = 0.5$ and $\beta mi = 0.9$.

#### *b: RESULTS OF BCNSG2MI*

Observing at the results in Fig 3, the average Pareto fronts of BCNSG2MI especially BCNSG2MI-1 comprise more than or equal to two solutions which choose the least features and consequently attained the comparable or improved performance compared to the full-length features in all the datasets. For example, similar performance was recorded in BCNSG2MI-1 and BCNSG2MI-2 on some few data points on the Chess dataset. In the majority of the datasets, BCNSG2MI-2 select the minimum number of feature subsets containing almost half of the total feature size then obtained boosted error rate compared to the full-length features. For instance, in Splice dataset BCNSG2MI-2 selects 17 features out of 60 and the error rate reduced from 30.25% to 20.00%. Almost similar results are achieved on the other datasets.

This is a testimony that BCNSG2MI as a filter-based multi-objective optimisation algorithm can automatically discover the Pareto front of an FS problem and minimise the error rate as well as the feature size required for the classification.

#### *c: COMPARISONS AMONG BCNSG3MI, BCNSG2MI AND BCOA-MI*

Relating the results obtained by BCNSG2MI with BCOA-MI, it can be noticed that in most cases, BCNSG2MI (BCNSG2MI -2) obtained an improved classification performance than BCOA-MI, For example, the charts in Fig. 3 shows that BCNSG2MI-2 outperformed BCOA-MI with $\beta mi = 0.5$ and $\beta mi = 0.9$ on all the datasets except on led-display, Madelon and Optic datasets. Moreover, BCOA-MI with $\beta mi = 0.5$ performed better than the BCNSG2MI-2 on the Soyabeans large dataset. In most cases, BCNSG2MI-2 outperformed BCOA-MI with $\beta mi = 0.5$ and $\beta mi = 0.9$ on the number of chosen features except in Connect4 and Promoter datasets where BCOA-MI recorded few numbers of selected features but with non-promising error rate.

The contrasting suggest that with MI in the fitness function, obtaining the best classification performance usually requires more features. However, occasionally there are some subsets of features that have fewer number features and yet attained better classification performance. Also, both BCNSG2MI and BCNSG3MI might acquire a set of nondominated solutions that used fewer feature size and achieved the best results. Thus, BCNSG2MI and BCNSG3MI as a filter-based multi-objective optimisation algorithm could better search for the solution region compared with the single-objective algorithm, BCOA-MI.

### 2) RESULTS OF BCNSG3E AND BCNSG2E

Fig. 4 displays the Pareto front solutions achieved by the BCNSG2E together with BCNSG3E in the entropy zone. Nevertheless, their error rate shown in the charts were assessed by SVM on the test data. The outcomes in Fig. 4 compares the results obtained by BCNSG2E, BCNSG3E and BCOA-E with $\beta E = 0.5$ and $\beta E = 0.9$, that used entropy to estimate the relevancy as well as redundancy of a set of features in contrast to the MI, that evaluates for a pair of features.

#### *a: RESULTS OF BCNSG3E*

From Fig. 4 BCOA-E with $\beta E = 0.5$ and $\beta E = 0.9$ decreased almost 70% of the total features on most of the datasets and yet attained an equivalent or higher classification performance than using the full-length features.

From the results, one can observe that BCNSG3E-2 perform well on all the dataset. It includes greater than or equal to a single solution which chose fewer features and yet attained a higher level of performance compared to the complete features. On the other hand, BCNSG3E-1 attained a better classification error rate in the majority of the datasets. Correspondingly, the feature size reduces drastically to almost 50% on all the datasets. For instance, in Soyabeans Large dataset, the feature size reduced from 35 to 17.5 (exactly 50% feature size reduction) besides the error rate from 9.05% to 5.00%. Likewise, the other datasets in Fig. 4 confirmed the assertions.

This result advocates that the advanced BCNSG3E algorithm could derive a set of feature subsets that might enhance the goodness of the classifier simultaneously and yet decrease the feature size.

Comparing BCNSG3E with BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.9$ one can observe that BCNSG3E performed better than BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.9$ in terms classification error rate in all the datasets except Connect4 where BCOA-E with $\beta E = 0.5$ performed better than BCNSG3MI. Moreover, a similar result was perceived on Dermatology, Leddisplay and Chess datasets. Similarly, BCNSG3E selected fewer features than BCOA-E in all the dataset.

Therefore, a comparison using gain ratio based-entropy as per the evaluation measure, the planned filter-based multi-objective FS (BCNSG3E) can attain better solutions and do well than the filter-based single-objective (BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.9$.

#### *b: RESULTS OF BCNSG2E*

According to the results in Fig. 4, all together, the average Pareto fronts of BCNSG2E (BCNSG2E -1) have many solutions that nominated smaller features and realised the best error rate compared to the entire full-length features. In most of the datasets, BCNSG2E-2 was able to minimise the error rate by choosing nearly half of the total features. Looking at the Promoter dataset, BCNSG2E decreased the error rate as of 8.65% to 0.00% by picking just 22 features out of the whole 57 features.

Moreover, the results indicated that the planned BCNSG2E together with entropy as the assessment condition could successfully choose a subset of features that can concurrently decrease the feature size and enhance the classification performance than the full-length features.

#### *c: COMPARISONS AMONG BCNSG3E, BCNSG2E AND BCOA-E*

Comparing the results of BCNSG2E with BCOA-E, in most cases, BCNSG2E (BCNSG2E-2) attained the best results compared to BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.9$. Despite, the feature size is a bit bigger in some few cases. Still, BCNSG2E outpaced BCOA-E since improving the error rate is considered more important than reducing feature size.

Furthermore, relating the results of BCNSG3E with BCOA-E, it can be observed that, in the majority of the datasets, BCNSG3E select the fewer features and obtained an improved result than BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.9$. A near similar result was achieved on Chess datasets, where BCNSG3E accomplished the same results to BCOA-E with both $\beta E = 0.5$ and $\beta E = 0.59$.

The comparisons of the methods show that using entropy as the assessment measure, the planned filter-based multi-objective FS algorithms (BCNSG2E and BCNSG3E) could well discover the exploration space and accomplish good solutions compared to single-objective FS algorithm, (BCOA-E).

### 3) COMPARISONS BETWEEN PROPOSED MULTI-OBJECTIVE APPROACHES

To be fair in the comparison between the proposed methods. This study adopts the pattern of the existing works in [10], [11]. In their papers, a comparison is first made with the single-objective then between the proposed methods and lastly with the state-of-the-art approaches (if exist). Based on that, this study also compared the proposed multi-objective filter-based approaches BCNSG2 (BCNSG2MI and BCNSG2E) and BCNSG3 (BCNSG3MI and BCNSG3E) with a single objective (BCOA-MI and BCOA-E). Then a comparison between the proposed multi-objective approaches is made based on the evaluation measures—for example, BCNSG2MI Vs BCNSG3MI since they all used MI as the filter evaluation measure. Also, BCNSG2E Vs BCNSG3E because they all used gain-ratio based entropy as the filter evaluation measures. However, it will not be fair to compare MI-based with entropy-based approaches

Relating between the MI as well as entropy-based algorithms in Figs 3 and 4 respectively. It shows that BCOA-E along with BCNSG2E and BCNSG3E, mainly attained an excellent classification performance with minimum error rate compared to the BCOA-MI, BCNSG2MI and BCNSG3MI.

On the other hand, BCOA-MI chose the least number of features than BCOA-E. Simply because MI deals with two pairs of features in contrast to the gain ratio based-entropy that deals with a group of features in finding both relevance and redundancy. Hence the reason why the number of features in BCOA-E are many compared with the BCOA-MI. Alternatively, the features selected by the planned multi-objective optimisation algorithms is quite lesser compared to the single-objective algorithms. Thus, BCNSG2E and BCNSG3E with entropy as the evaluation criterion can attain an excellent result because it can use multiple ways relevancy and redundancy to improve both the classification performance and some selected features compared to BCNSG2MI and BCNSG3MI with MI as the evaluation condition.

Comparing among the algorithms in Figs 3 and 4, one can observe that BCNSG3MI and BCNSG3E based on NSGAIII framework outperformed the BCNSG2MI and BCNSG2E based on NSGAII framework on all the datasets both on error rate and the selected features. The results of both BCNSG3MI and BCNSG3E is 10-20% better than the BCNSG2MI and BCNSG2E in the majority of the datasets.

The results are not surprising because NSGAII is reported to lacks some reference point; instead, it used the crowding distance and mutation operators for its computation. Also, a full crowded comparison can restrict the convergence of NSGAII [50]. Therefore, the maintenance of diversity among population members in NSGAIII is supported by supplying and adaptively updating several well-spread reference points. Hence, the reason why BCNSG3MI and BCNSG3E outperformed both BCNSG2MI and BCNSG2E and can search for the better zone of the solutions and attained best classification performance using fewer features than all the other methods.

### 4) COMPARISON AMONG PROPOSED APPROACHES BASED ON TIME

The results in Table 4 analyses the average time spent in seconds by all the proposed algorithms. The four filter-based multi-objective algorithms are compared with the two filter-based methods BCOA-MI (with $\beta_{MI} = 0.5$ and $\beta_{MI} = 0.9$) along with BCOA-E (with $\beta_E = 0.5$ and $\beta_E = 0.9$).

The table (Table 4) displays that usually, majority of the pair-wise multi-objective algorithms, BCNSG2MI and BCNSG3MI complete their metamorphic training process in less than four seconds except on the Connect4, DNA and Madelon datasets. The Madelon dataset generally recorded much lengthier time compared to other datasets since it has the highest number of features than the remaining datasets. Likewise, in Connect4 dataset because of its large number of instances.

On the other hand, while applying the gain-ratio based entropy (group-based measures), all the multi-objective algorithms, BCNSG2E and BCNSG3E, completed the metamorphic training procedure around one minute in all the datasets excluding the Madelon dataset. Thus, there is some little variation on the time spent by the multi-objective algorithms. So, the BCNSG3E outperformed all others. The single objective algorithm BCOA-E with $\beta_E = 0.5$ and $\beta_E = 0.9$ spent lengthier time compared to the multi-objective algorithms, that is almost ten times lengthier on all the dataset. The wisdom behind it is that the feature size in the multi-objective algorithms is calculated as a single objective, which requires minor time compared to the redundancy measure $RedB_E$ in

**TABLE 4.** Computational time for the proposed methods.

| Datasets | $\beta_{MI}$ = 0.5 | $\beta_E$ = 0.5 | $\beta_{MI}$ = 0.9 | $\beta_E$ = 0.9 | BCNSG2MI | BCNSG2E | BCNSG3MI | BCNSG3E |
|---|---|---|---|---|---|---|---|---|
| Lymph | 0.120 | 3.16 | 0.130 | 3.19 | 0.120 | 0.1 | 0.110 | 0.11 |
| Spect | 0.140 | 7.5 | 0.140 | 10.6 | 0.150 | 0.11 | 0.130 | 0.1 |
| Leddisplay | 0.210 | 31.25 | 0.210 | 33.95 | 0.200 | 4.25 | 0.200 | 4.15 |
| Dermatology | 0.210 | 14.25 | 0.210 | 15.65 | 0.190 | 2.68 | 0.180 | 1.71 |
| Soybeanlarge | 0.340 | 36.52 | 0.360 | 40.25 | 0.200 | 1.15 | 0.220 | 1.13 |
| Chess | 0.560 | 212.11 | 0.570 | 248.52 | 17.050 | 60 | 17.050 | 55.25 |
| Connect4 | 102.600 | 996.87 | 102.570 | 1002.43 | 102.640 | 704.52 | 102.250 | 650.05 |
| Promoter | 0.440 | 42.25 | 0.480 | 44.65 | 0.320 | 3.49 | 0.310 | 2.63 |
| Splice | 0.550 | 26.86 | 0.510 | 28.96 | 0.490 | 2.28 | 0.480 | 2.25 |
| Optic | 0.730 | 121.65 | 0.640 | 136.46 | 0.620 | 7.55 | 0.610 | 7.45 |
| Audiology | 0.470 | 97.56 | 0.490 | 99.25 | 0.430 | 6.52 | 0.180 | 4.98 |
| Coil2000 | 1.550 | 106.11 | 1.510 | 97.65 | 1.490 | 6.85 | 1.470 | 5.25 |
| DNA | 10.560 | 212.9 | 12.120 | 294.09 | 9.210 | 6.51 | 8.050 | 6.42 |
| Madelon | 210.560 | 2905.56 | 299.540 | 2999.54 | 235.250 | 2196.52 | 221.550 | 2026.78 |
| **Average Time** | **23.503** | **343.896** | **29.963** | **361.085** | **26.311** | **214.466** | **25.199** | **197.733** |

the fitness function of the BCOA. Generally, the combined entropy-based algorithms spent lengthier time compared to its MI-based counterpart.

### 5) DISCUSSIONS

In Figs. 3 and 4, the solutions employed in the graph are the Pareto front solutions gotten through the filter-based evaluation measure. Nevertheless, the classification performances display in the graphs were assessed using SVM on the test sets. While Fig. 3 displays the Pareto fronts attained by the BCNSG2MI and BCNSG3MI via MI as the assessment condition. Fig. 4 displays the Pareto fronts attained by BCNSG2E and BCNSG3E via entropy as the assessment measure.

It can be observed from Figs. 3 and 4, that some of the solutions in the average Pareto front (denoted by '$-1$') influence others though they are nondominated solutions in the filter-based assessment condition. Hence, this confirms that the Pareto front in the filter-based assessment condition zone on the training set have not included similar subsets as per the Pareto front in the SVM-based assessment on the test set. Just because the superiority of a feature subset assessed by MI or entropy on the training set does not automatically display its meticulous goodness on the test set.

Furthermore, the right Pareto front accomplished by the comprehensive exploration in the twofold filter-based assessment measures, the objective space cannot be the right Pareto front of the SVM-based assessment on the test set. The subsets of features that have similar filter-based results cannot essentially accomplish similar (good or poor) error rate on the hidden test set assessed by SVM. Let takes dual subsets of the feature as an example, both may have equal feature size, but diverse mixtures of the features. Those two subsets of feature possibly will have similar goodness assessed by the filter-based assessment condition on the training set. Therefore, they are nondominated with all others. Though, if SVM is applied or any available classifier to assess their error rate on the hidden test set, their error rate possibly will be somewhat dissimilar. The subsets of features that have the best classification performance will influence others. Also, like other filter-based conditions and other classifiers. As such, the Pareto front in the filter assessment condition region is mostly not similar to the Pareto front in the SVM-based assessment.

In an ideal world, the algorithms would recognise the right Pareto front in all the filter-based assessment condition zone. Since it is not possible to carry out a complete search for the datasets with huge feature size to detect the right Pareto fronts. The proposed multi-objective algorithms BCNSG3MI and BCNSG3E will recognise the right Pareto fronts gotten by the complete search; nonetheless, BCNSG2MI and BCNSG2E to some extends cannot. The main reason is that the BCNSG2MI and BCNSG2E lack some reference point; instead, it used the crowding distance and mutation operators for its computation. Moreover, a full crowded comparison restricts its convergence due to the used of NSGAII. BCNSG3MI and BCNSG3E attained the right Pareto front for the datasets with huge feature size.

### 6) COMPARISONS WITH OTHER EXISTING APPROACH

To be fair in the comparison, only filter-based multi-objective FS approaches that use the concepts of nondominated sorting and information theory and yet have similar datasets. For example, in the work of [10] we have eight related datasets, they used the concepts of nondominated sorting and crowding distance as well as MI and entropy all embedded in PSO. Similarly, the work of [11] has eleven datasets in common with this study. In addition to that, MI and relief f are used as filter evaluation measures in multi-objective DE.

The detailed comparison with the existing works is shown in the subsequent sections.

#### *a: COMPARISONS WITH BPSO*

To further investigate the performances of the proposed BCNSG3MI, BCNSG2MI, BCNSG3E and BCNSG2E algorithms we, first of all, compared them with four

**TABLE 5. Comparison between proposed multi-objective filter-based with existing works.**

| Datasets | BCNSG2MI | BCNSG3MI | MODE_MI | BCNSG2E | BCNSG3E | MODE_MIRF |
|---|---|---|---|---|---|---|
| Lymph | 0.12 | 0.11 | 0.12 | 0.1 | 0.11 | 0.1 |
| Spect | 0.15 | 0.13 | 0.14 | 0.11 | 0.1 | 0.11 |
| Leddisplay | 0.2 | 0.2 | 0.39 | 4.25 | 4.15 | 0.27 |
| Soybeanlarge | 0.2 | 0.22 | 0.22 | 1.15 | 1.13 | 0.13 |
| Connect4 | 102.64 | 102.25 | 754.09 | 704.52 | 650.05 | 705.03 |
| Promoter | 0.32 | 0.31 | 0.31 | 3.49 | 2.63 | 0.1 |
| Splice | 0.49 | 0.48 | 2.42 | 2.28 | 2.25 | 2.29 |
| Optic | 0.62 | 0.61 | 8.43 | 7.55 | 7.45 | 7.56 |
| Audiology | 0.43 | 0.18 | 0.43 | 6.52 | 4.98 | 0.18 |
| Coil2000 | 1.49 | 1.47 | 34.77 | 6.85 | 5.25 | 29.92 |
| DNA | 9.21 | 8.05 | 8.16 | 6.51 | 6.42 | 6.91 |
| Average | 10.53364 | **10.36455** | 73.58909 | 67.57545 | **62.22909** | 68.41818 |

multi-objective PSO filter-based feature selections NSfsMi and NSfsE in [10] based on nondominated sorting based multi-objective PSO in [51]. Moreover, the results are compared with CMDfsMI and CMDfsE based on multi-objective PSO in [10]. All the eight datasets used in [10] except Mushroom dataset are compared with the results obtained in this study. The proposed approaches performed better than both CMDfsMI and CMDfsE as well as the NSfsMI and NSfsE on all the datasets with around 5-15% and 15-20% better in terms of both selected features and classification performance for BCNSG2MI and BCN-SG2E respectively. Moreover, the proposed BCNSG3MI and BCNSG3E performed even better with almost 20-35% reduction on both error rate as well as selected features.

The comparisons above clearly indicate that multi-objective BCOA with both NSGAII and NSGAIII has more advanced search mechanisms and have the potential of achieving even better performance.

### b: COMPARISONS WITH DE

Besides, the results obtained are also compared with $MODE_{mi}$ as well as $MODE_{mirf}$ in [11] on the eleven datasets that are common to this study. Although, $MODE_{mirf}$ performed much better than $MODE_{mi}$ on all the datasets. The proposed BCNSG2MI and BCNSG3E outpaced both $MODE_{mi}$ and $MODE_{mirf}$ on most of the datasets except on Leddisplay datasets that they attained the same performance both on the selected features and error rate. Conversely, the proposed BCNSG3MI and BCNSG3E outpaced both $MODE_{mi}$ and $MODE_{mirf}$ on all the datasets. Therefore, the proposed multi-objective approaches have the potential to evolve the Pareto front features subsets automatically. Also, simultaneously, select the minimum and most relevant features and consequently attain the best results than the existing methods.

### c: COMPARISON WITH EXISTING METHODS BASED ON TIME

The existing methods are also compared based on the CPU execution time (in seconds) as shown in Table 5. To be fair in the comparison $MODE_{MIRF}$ is compared with BCNSG2E and BCNSG3E while $MODE_{MI}$ is compared with BCNSG2MI and BCNSG3MI. The Average in the last row of the table represents the average number of time spent on all the datasets for each method. The bold signifies the best methods which are BCNSG3MI and BCNSG3E. Comparing amongst BCNSG2MI, BCNSG3MI and $MODE_{MI}$ it can be seen that the proposed methods performed faster than $MODE_{MI}$ on all the datasets except on SoyabeanLarge and Promoter datasets where similar execution time was recorded on BCNSG3MI with the $MODE_{MI}$.

On the other hand, the proposed BCNSG2E and BCNSG3E are faster than $MODE_{MIRF}$ on five out of the eleven datasets. Similar execution time is realised on Lymph and Spect datasets. Alternatively, $MODE_{MIRF}$ is better than the proposed methods on Audiology, Promoter, Soyabean-Large and Leddisplay datasets. Meanwhile, the proposed methods are faster on the majority of the datasets. The average time of both BCNSG2E and BCNSG3E is 67.58 and 62.23 compared to 68.42 recorded by the $MODE_{MIRF}$.

### 7) LIMITATION OF THE PROPOSED METHODS

This paper presents the first study of filter-based multi-objective FS using the concepts of NSGAII, NSGAIII with BCOA along with MI and gain ratio based entropy as the filter-based evaluation measures. Even though the results obtained are competitive to other existing works, however, the proposed methods have some limitations as follows:

1) The crowding-distance strategy of the BCNSG2MI and BCNSG2E restricted in the same front, can't exhibit the real superiority in the same front. Hence, there is a need to improve the dummy fitness strategy while considering the crowding within a different front.
2) Although the standard NSGA-II algorithm uses the crowding distance-based method for maintaining solutions diversity, the limitation of the crowding distance-based approach according to [63] is that it, selects two nearer solutions from the Pareto front for the mating. As such, the proposed methods BCNSG2MI and BCNSG2E sometimes may not preserve extreme solutions in Pareto front.

3) Using reference points in NSGA-III has difficulty in maintaining the diversity of the solutions in the discrete multi-objective optimisation problems [64]. Therefore, the proposed BCNSG3MI and BCNSG3E were likely unable to link all reference points, especially the best reference points in each objective.
4) The use of gain-ratio based entropy as the filter-based evaluation measure provides the best solutions compared to its MI counterpart. However, the gain ratio based entropy is computationally expensive. Hence the use of other faster filter-based approaches that can handle a group of features at a time as an evaluation measure may likely solve the problem.
5) Rajabioun in [16] stated that "it should be noted that the higher performance of COA in reaching better results for these five benchmark functions and areal case study does not necessarily mean that COA is the ever best evolutionary method developed. It just can be considered as a successful mimicking of nature; suitable for some sort of optimisation problems." As such other EAs may be fine-tuned and use for filter-based multi-objective FS.

### C. RESULTS ANALYSIS

The results show that information theory concept can be successfully used as a filter-based evaluation measure with BCOA to select fewer number of features and better classification performance. A relevance was employed to measure the classification performance of the selected features to the class labels. On the other hand, the number of the selected features is measured by the redundancy amongst features chosen. Based on that, two different relevance and redundancy measures are established, which are a pair-wise based on MI and a group-based using the concept of gain ratio based entropy.

In the pairwise based measure, it shows that BCOA-MI is faster than its BCOA-E counterpart and the optimal fitness values comprise of a few numbers of features, whereas the classification performance is in favour of the BCOA-E. The reason behind this is that BCOA-MI used pairwise evaluation to measure the relationship between two features, which does not involve complex computation of relevance and redundancy. Thus, no complex interactions amongst a group of features, which is considered a challenge in FS problems.

Alternatively, BCOA-E using the group-based measure is slower but yet recorded better classification performance and choose more features than the BCOA-MI. The reason is that it deals with subsets or group features while computing the relevance and redundancy. Also, it considers the selected features as a whole which leads to better feature interaction that consequently leads to improve classification performance.

A weight $\beta$ values were employed to balance between the relevance (accuracy) and redundancy (selected features) in the fitness function for both BCOA-MI and BCOA-E, respectively. It is challenging to choose pre-determine best value of the $\beta$. The reason is that a considerable weight value on the redundancy in BCOA-MI or BCOA-E may reduce the feature size and affect the classification performance or vice versa. Similarly, larger weight value on the relevance may improve the classification performance and consequently affect the number of selected features and vice versa. To avoid this problem, both the relevance and redundancy are treated as two separate objectives in a multi-objective FS. It is hypothesised that it will solve the task better and obtain a set of nondominated solutions instead of a single solution, where the gathered Pareto front can assist users in choosing their preferred solutions to meet their requirements.

Based on that the concepts of NSGAII and NSGAIII are embedded in BCOA-MI and BCOA-E respectively to form BCNSG2 (BCNSG2MI and BCNSG2E) and BCNSG3 (BCNSG3MI and BCNSG3E). Both BCNSG3MI and BCNSG3E achieved the best performance than BCNSG2MI and BCNSG2E with regards to both selected features and the classification error rate on most of the datasets. It is because FS tasks are complicated problems with various local optimal. And BCNSG3MI, along with BCNSG3E, uses multiple mechanisms for maintenance of variety among population members and is supported by providing and adaptively updating several well-spread reference points. Precisely, it picks and screens out jam-packed leaders and applies various mutation operators to preserve the variety of the crowd to evade stagnancy in local optimal.

Also, both BCNSG2MI and BCNSG2E are not as good as BCNSG3MI and BCNSG3E regarding stagnancy avoidance in a local optimum. They handle various stages of Pareto fronts to keep the previously found nondominated solutions. Hence, the entire nondominated solutions are saved in the habitat from one iteration to another. The nondominated solutions would be replicated, and the habitat may miss variety faster, that might cause the problem of early convergence. Hence the reason why both BCNSG3MI and BCNSG3E are faster than their BCNSG2MI and BCNSG2E counterparts.

## VI. CONCLUSION

This study aimed was to examine the use of FS specifically, filter-based utilising BCOA and information theory concepts for both single and multi-objective FS. The aims have been accomplished by developing two new filter-based single objective FS using MI and entropy, which are BCOA-MI along with BCOA-E. The BCOA-MI apply MI for all the couples of features to assesses the relevance as well as redundancy of the chosen couple of features. Whereas BCOA-E applies entropy to all the set of features to assess the relevance and redundancies of the chosen feature subsets. Besides, diverse weights values are assigned to evaluate the relevance and redundancy.

The outcome of the filter-based single objective disclosed that using a suitable value for the weight the BCOA-MI and BCOA-E could decrease the feature size and subsequently attain or accomplish comparable classification performance. BCOA-MI selected the smaller subsets of features while BCOA-E gets the best classification performance. However, neither BCOA-MI nor BCOA-E balance between the error rate as well as features size. As a result, a multi-objective filter-based BCOA is also proposed to find the set of non-dominated solutions.

The aim of developing a filter-based multi-objective FS has also been achieved, in which the novel idea of NSGAIII, as well as NSGAII, are employed to hunt for fewer features with best error rate. Four filter-based multi-objective FS BCNSG2MI, BCNSG2E, BCNSG2MI and BCNSG3E were developed and evaluated also based on MI and entropy. The algorithms are first compared with BCOA-MI and BCOA-E, on fourteen benchmark datasets of varying degree of complexities. The multi-objective algorithms outperformed the single-objective algorithms in most of the datasets and can easily evolve a set of nondominated solutions with fewer feature size and improve performance

In addition to that, the presented multi-objective algorithms are also related to filter-based multi-objective BPSO (NSfsMI and NSfsE) with MI and entropy as evaluation criteria. Also, with filter-based multi-objective DE approach ($MODE_{mi}$ and $MODE_{mirf}$). Whereby, the proposed multi-objective approach outperformed all the existing approaches and can easily evolve the Pareto subset of features with the least feature size and yet attained an improve classification performance.

Even though the proposed multi-objective approach would derive the best subsets of features, it is not clear whether the Pareto front, together with the set of nondominated solutions, can be improved or otherwise. Thus, in the future, filter-based multi-objective will address such problems and compared with other popular evolutionary algorithms for better solutions. Wrappers have better classification performance than the filter-based, but most of them are single objective that works by combining the aims of the FS into one single fitness function. Thus, future work on balancing those conflicting aims using the wrapper-based along with the novel concept of NSGAIII is not fully studied.

Moreover, recently filter-wrapper approaches are combined to benefits from the advantages of both. For example, filters are faster and scalable to large datasets but lack good classification performance. Whereas wrappers got good classification performance but not fast. Although, there are filter-wrapper approaches proposed in the literature to augment the problems of each approach and consequently benefits from their advantage. However, the work on multi-objective filter-wrapper FS is still an open issue, since the problem of each approach still exists in the single-objective.

## REFERENCES

[1] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, p. 94, 2016.

[2] E. Hancer, "Differential evolution for feature selection: A fuzzy wrapper–filter approach," *Soft Comput.*, vol. 23, no. 13, pp. 5233–5248, Jul. 2019.

[3] S. J. Russell and P. Russell, *Artificial Intelligence: A Modern Approach*. London, U.K.: Pearson Education Limited, 2016.

[4] A. M. Usman, U. K. Yusof, and S. Naim, "Cuckoo inspired algorithms for feature selection in heart disease prediction," *Int. J. Adv. Intell. Informat.*, vol. 4, no. 2, p. 95, 2018.

[5] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[6] M. Tavana, S. Shahdi-Pashaki, E. Teymourian, F. J. Santos-Arteaga, and M. Komaki, "A discrete cuckoo optimization algorithm for consolidation in cloud computing," *Comput. Ind. Eng.*, vol. 115, pp. 495–511, Jan. 2018.

[7] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6345–6398, 2017.

[8] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.

[9] L. Cervante, B. Xue, M. Zhang, and L. Shang, "Binary particle swarm optimisation for feature selection: A filter based approach," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2012, pp. 1–8.

[10] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "A multi-objective particle swarm optimisation for filter-based feature selection in classification problems," *Connection Sci.*, vol. 24, nos. 2–3, pp. 91–116, Sep. 2012.

[11] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowl.-Based Syst.*, vol. 140, pp. 103–119, Jan. 2018.

[12] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013.

[13] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[14] X. Yuan, H. Tian, Y. Yuan, Y. Huang, and R. M. Ikram, "An extended NSGA-III for solution multi-objective hydro-thermal-wind scheduling considering wind power cost," *Energy Convers. Manage.*, vol. 96, pp. 568–578, May 2015.

[15] H. Jain and K. Deb, "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach,—Part II: Handling constraints and extending to an adaptive approach," *IEEE Trans. Evol. Comput.*, vol. 18, no. 4, pp. 602–622, Aug. 2014.

[16] R. Rajabioun, "Cuckoo optimization algorithm," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 5508–5518, Dec. 2011.

[17] K. Kira and L. A. Rendell, *A Practical Approach to Feature Selection*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.

[18] C. Cardie, "Using decision trees to improve case-based learning," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 25–32.

[19] H. Almuallim and T. G. Dietterich, "Learning Boolean concepts in the presence of many irrelevant features," *Artif. Intell.*, vol. 69, nos. 1–2, pp. 279–305, Sep. 1994.

[20] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[21] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[22] C. Bishop and C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

[23] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 2000.

[24] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.

[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[26] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2006, pp. 68–82.

[27] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, vol. 2. Springer, 2009.

[28] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognit.*, vol. 42, no. 7, pp. 1330–1339, Jul. 2009.

[29] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[30] G. Brown, "A new perspective for information theoretic feature selection," in *Proc. Artif. Intell. Statist.*, 2009, pp. 49–56.

[31] R. Liu, N. Yang, X. Ding, and L. Ma, "An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure," in *Proc. 3rd Int. Symp. Intell. Inf. Technol. Appl.*, Nov. 2009, pp. 65–68.

[32] L. Zhu, L. Miao, and D. Zhang, "Iterative Laplacian score for feature selection," in *Proc. Chin. Conf. Pattern Recognit.* Springer, 2012, pp. 80–87.

[33] S. Foithong, O. Pinngern, and B. Attachoo, "Feature subset selection wrapper based on mutual information and rough sets," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 574–584, Jan. 2012.

[34] K. Yu, W. Ding, and X. Wu, "LOFS: A library of online streaming feature selection," *Knowl.-Based Syst.*, vol. 113, pp. 1–3, Dec. 2016.

[35] J. Fulcher, *Computational Intelligence: An Introduction*. Springer, 2008, pp. 3–78.

[36] M. Moghadasian and S. P. Hosseini, "Binary cuckoo optimization algorithm for feature selection in high-dimensional datasets," in *Proc. Int. Conf. Innov. Eng. Technol. (ICIET)*, 2014, pp. 18–21.

[37] B. Xue, L. Cervante, L. Shang, and M. Zhang, "A particle swarm optimisation based multi-objective filter approach to feature selection for classification," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Springer, 2012, pp. 673–685.

[38] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "Multi-objective evolutionary algorithms for filter based feature selection in classification," *Int. J. Artif. Intell. Tools*, vol. 22, no. 4, Aug. 2013, Art. no. 1350024.

[39] Y. Zhang, D.-W. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 64–75, Jan. 2017.

[40] H. B. Nguyen, B. Xue, I. Liu, P. Andreae, and M. Zhang, "New mechanism for archive maintenance in PSO-based multi-objective feature selection," *Soft Comput.*, vol. 20, no. 10, pp. 3927–3946, Oct. 2016.

[41] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay, "Pareto front feature selection based on artificial bee colony optimization," *Inf. Sci.*, vol. 422, pp. 462–479, Jan. 2018.

[42] M. Gorjestani, E. Shadkam, M. Parvizi, and S. Aminzadegan, "A hybrid COA-DEA method for solving multi-objective problems," 2015, *arXiv:1509.00595*. [Online]. Available: http://arxiv.org/abs/1509.00595

[43] Z. Borhanifar and E. Shadkam, "The new hybrid COAW method for solving multi-objective problems," 2016, *arXiv:1611.00577*. [Online]. Available: http://arxiv.org/abs/1611.00577

[44] A. Konak, D. W. Coit, and A. E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," *Rel. Eng. Syst. Saf.*, vol. 91, no. 9, pp. 992–1007, Sep. 2006.

[45] S. Mahmoudi, R. Rajabioun, and S. Lotfi, "Binary cuckoo optimization algorithm," in *Proc. 1st Nat. Conf. New Approaches Comput. Eng. Inf. Retr. Young Researchers Elite Club Islamic Azad Univ., Roudsar-Amlash Branch*, 2013.

[46] A. Tsanas, M. A. Little, and P. E. McSharry, "A simple filter benchmark for feature selection," *J. Mach. Learn. Res.*, vol. 1, nos. 1–24, 2010.

[47] M. Amoozegar and B. Minaei-Bidgoli, "Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism," *Expert Syst. Appl.*, vol. 113, pp. 499–514, Dec. 2018.

[48] K. Deb, *Multi-Objective Optimization*. Springer, 2014, pp. 403–449.

[49] T. M. Hamdani, J.-M. Won, A. M. Alimi, and F. Karray, "Multi-objective feature selection with NSGA II," in *Proc. Int. Conf. Adapt. Natural Comput. Algorithms*. Springer, 2007, pp. 240–247.

[50] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using Reference-Point-Based nondominated sorting approach,—Part I: Solving problems with box constraints," *IEEE Trans. Evol. Comput.*, vol. 18, no. 4, pp. 577–601, Aug. 2014.

[51] X. Li, "A non-dominated sorting particle swarm optimizer for multiobjective optimization," in *Proc. Genetic Evol. Comput. Conf.* Springer, 2003, pp. 37–48.

[52] A. Frank and A. Asuncion, "Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california," *School Inf. Comput. Sci.*, vol. 213, no. 11, p. 2, 2010.

[53] R. Caruana and D. Freitag, "Greedy attribute selection," in *Machine Learning Proceedings*. Amsterdam, The Netherlands: Elsevier, 1994, pp. 18–36.

[54] Y. Zhang, S. Cheng, Y. Shi, D.-W. Gong, and X. Zhao, "Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm," *Expert Syst. Appl.*, vol. 137, pp. 46–58, Dec. 2019.

[55] X.-H. Wang, Y. Zhang, X.-Y. Sun, Y.-L. Wang, and C.-H. Du, "Multi-objective feature selection based on artificial bee colony: An acceleration approach with variable sample size," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 106041.

[56] Y. Zhang, D.-W. Gong, X.-Z. Gao, T. Tian, and X.-Y. Sun, "Binary differential evolution with self-learning for multi-objective feature selection," *Inf. Sci.*, vol. 507, pp. 67–85, Jan. 2020.

[57] Y. Zhang, H.-G. Li, Q. Wang, and C. Peng, "A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection," *Int. J. Speech Technol.*, vol. 49, no. 8, pp. 2889–2898, Aug. 2019.

[58] Y. Liu, D. Gong, J. Sun, and Y. Jin, "A many-objective evolutionary algorithm using a One-by-One selection strategy," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2689–2702, Sep. 2017.

[59] Y. Liu, D. Gong, X. Sun, and Y. Zhang, "Many-objective evolutionary optimization based on reference points," *Appl. Soft Comput.*, vol. 50, pp. 344–355, Jan. 2017.

[60] Y. Zhang, X.-F. Song, and D.-W. Gong, "A return-cost-based binary firefly algorithm for feature selection," *Inf. Sci.*, vols. 418–419, pp. 561–574, Dec. 2017.

[61] Z. Yong, G. Dun-wei, and Z. Wan-qiu, "Feature selection of unreliable data using an improved multi-objective PSO algorithm," *Neurocomputing*, vol. 171, pp. 1281–1290, Jan. 2016.

[62] Y. Zhang, Q. Wang, D.-W. Gong, and X.-F. Song, "Nonnegative Laplacian embedding guided subspace learning for unsupervised feature selection," *Pattern Recognit.*, vol. 93, pp. 337–352, Sep. 2019.

[63] V. L. Vachhani, V. K. Dabhi, and H. B. Prajapati, "Improving NSGA-II for solving multi objective function optimization problems," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2016, pp. 1–6.

[64] H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima, "Performance comparison of NSGA-II and NSGA-III on various many-objective test problems," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 3045–3052.

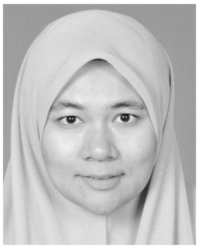
[65] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Statist. Data Anal.*, vol. 143, Mar. 2020, Art. no. 106839.

**ALI MUHAMMAD USMAN** (Member, IEEE) received the B.Tech. degree in computer science from Abubakar Tafawa Balewa University, in 2005, and the M.Sc. degree in computer science from Bauchi and Adamawa State University, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Sciences, Universiti Sains Malaysia.

He is also a Lecturer with the Department of Computer Science, Federal College of Education (Technical) Gombe, Nigeria. He has published several international and journal conferences. His research interests include data mining, computational intelligence, Web/software engineering, optimization, multiobjective optimization, and data security. He is a member of the Computational Intelligence Society of the IEEE.

**UMI KALSOM YUSOF** received the B.Sc. degree from Western Illinois, Macomb, IL, USA, the M.Sc. degree from Universiti Sains Malaysia (USM), Penang, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM) Skudai, Johor.

She is currently an Associate Professor and a Lecturer with the School of Computer Sciences, USM. Her research interests are related to data mining, Web engineering, computational intelligence, artificial intelligence, multiobjective optimization, evolutionary computing, computer security, and grid computing. She has published research articles at national and international journals, conference proceedings, as well as chapters of books.



**SYIBRAH NAIM** (Member, IEEE) received the B.Sc. degree in financial mathematics and the M.Sc. degree in applied mathematics from Universiti Malaysia Terengganu, in 2007 and 2010, respectively, and the Ph.D. degree in computer science from the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K., in 2014. She is currently a Lecturer with the Technology Department, Endicott College of International Studies (ECIS), Woosong University, South Korea. She has published research articles at national and international journals, conference proceedings, as well as chapters of books. Her research interests are related to optimization, computational intelligence, artificial intelligence, multiobjective optimization, evolutionary computation, soft computing fuzzy clustering, fuzzy logic, fuzzy set theory, and fuzzy theory.