IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 3, 2020, accepted June 18, 2020, date of publication July 1, 2020, date of current version July 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006154

An Integrated Approach for Ovarian Cancer Classification With the Application of Stochastic Optimization

SUNIL KUMAR PRABHAKAR[1], (Member, IEEE), AND SEONG-WHAN LEE[2], (Fellow, IEEE)
[1]Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea
[2]Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea

Corresponding author: Seong-Whan Lee (sw.lee@korea.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (Department of Artificial Intelligence, Korea University) under Grant 2019-0-00079.

ABSTRACT Ovarian Cancer is a type of cancer that begins in ovaries posing a serious threat to women. As a result, it leads to abnormal cells which has the ability to spread to other regions of the body. A highly useful diagnostic and prognostic data for ovarian cancer research is provided by the microarray data. Typically, genes with tens of thousands of dimension are present in the microarray data of ovarian cancer. There is a systematic methodology required to analyze this data and so it is important to select the most important genes or features for the entire data to avoid the computational complexity. In this work, an integrated approach to feature selection is done by two consecutive steps. Initially, the features are selected by the standard gene selection techniques such as Correlation Coefficient, T-Statistics and Kruskal-Wallis test. The selected genes or features will be further optimized by four suitable stochastic optimization algorithms chosen here such as Central Force Optimization (CFO), Lightning Attachment Procedure Optimization (LAPO), Genetic Bee Colony Optimization (GBCO) and Artificial Algae Optimization (AAO). Finally, it is classified with five different classifiers to analyze the ovarian cancer classification and the best results are projected when Kruskal Wallis test with GBCO is conducted and classified with Support Vector Machine – Radial Basis Function (SVM-RBF) Kernel technique giving a high classification accuracy of 99.48%. Similar results are also obtained when Correlation Coefficient test with AAO is conducted and classified with Logistic Regression giving a high classification accuracy of 99.48%.

INDEX TERMS Ovarian cancer, feature selection, optimization, classification.

I. INTRODUCTION

To treat ovarian cancer, surgery and chemotherapy are the general possible solutions used [1]. Usually the early stage ovarian cancer does not cause many symptoms [2]. Very few and nonspecific symptoms are caused by the advanced stage ovarian cancer that is mostly mistaken for common benign conditions [3]. The major symptoms of ovarian cancer include abdominal bloating/swelling, weight loss, discomfort in the pelvis area, changes in the bowel habits, quickly felling full after eating and an urgency to urinate [4]. Some of the types of ovarian cancer includes epithelial tumors, stromal tumors and germ kill tumors [5]. The factors that tend to increase the ovarian cancer risk is older age, family history of ovarian cancer, estrogen hormone replacement therapy, inherited gene mutations and consideration of the age when the menstruation begins and ends.

On a genomic scale, to globally analyze tens of thousands of genes, the most successful technique of the gene chip emerged and that is called microarray technology [6]. On a planar substrate, the layout of microarray data is a simple ordered array of microscopic elements. The relevant or irrelevant genes with respect to the cancer development is contained in the genotype. Thus, a microarray data can be considered as a device utilized to compare the expression level of the genes and genotype of the patients [7]. The expression level of genes under different conditions can be compared by this technology. As the microarray data contains a huge number of genes, tackling the genes is very difficult due to the curse of dimensionality problem [8]. Therefore, reducing the

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

size of it is highly essential. For the selection of informative genes from microarray data, a lot of approaches have been proposed in literature [9]. Many machine learning techniques with suitable genomic expressions have been successfully implemented in the past for the ovarian cancer classification from microarray data. As the genomic expression is with high dimension, there is a high time complexity involved in it also. Therefore, to get a better understanding into the global gene expression analysis and to get a higher classification accuracy, a systemic and integrated approach has been given in this paper.

The most important literature in the ovarian cancer classification is discussed as follows. For the human ovarian cancer, the microarray analysis of differentially expressed genes is given by Lee *et al.* [10]. The microarray-based gene expression studies in ovarian cancer was given by Chon and Lancaster [11]. The overview of biomarkers for the ovarian cancer diagnosis was done by Zhang *et al.* [12]. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer was done by Lee [13]. Intelligent systems were utilized to analyze the microarray data for classification of ovarian cancer by Jeng *et al.* [14]. The complementary learning fuzzy neural networks for ovarian cancer diagnosis was done by Tan *et al.* [15]. For the ovarian cancer microarray data, the dimension reduction was carried out by Chuang *et al.* [16]. The microarray analysis of ovarian cancer with machine learning was done by Huang *et al.* [17]. The gene expression patterns in the histopathological classification of epithelial ovarian cancer was done by Zhu and Yu [18]. A Bayesian neural network method was approached to ovarian cancer identification from high resolution mass spectrometry data by Yu and Chen [19]. The diagnosis of ovarian cancer utilizing decision tree classification of mass spectral data was done by Vlahou *et al.* [20]. The automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks was done by Wu *et al.* [21]. The machine learning techniques with Fourier transform was evaluated for the classification of ovarian tumors by Mas *et al.* [22]. A comparative study on various classification techniques for detection of ovarian cancer was done by Nuhic *et al.* [23]. An application of Artificial Neural Network (ANN) in the early detection of ovarian cancer was done by Zhang *et al.* [24]. The potentials and limitations of utilizing Bayesian networks for ovarian cancer diagnosis was done by Antal *et al.* [25]. Based on the historical data of ovary cancer patients, the detection of ovary cancer was done using Decision Trees (DT) classifiers by Osmanovic *et al.* [26]. With the help of feed forward ANN, the early detection of ovary cancer was done by Thakur *et al.* [27]. The recent progress in the diagnosis and treatment of ovarian cancer was done by Jelovac and Armstrong [28]. The application of SVM to ovarian cancer classification was done by Kusy [29]. The epithelial ovarian cancer stage subtype classification using gene expression approach was done by Nabawy *et al.* [30]. Using multicategory machine learning, the intraoperative diagnosis support tool for ovarian tumors based on Microarray data was done by Park *et al.* [31]. The ovarian cancer classification using bagging and random forest was done by Arfiani and Rustam [32]. A novel online paradigm for ovarian tumor characterization and classification using ultrasound was done by Acharya *et al.* [33]. To improve the diagnostic accuracy for prediction of ovarian cancer, a three-dimensional power Doppler ultrasound was used by Cohen *et al.* [34]. The ovarian cancer classification with missing data was done by Renz *et al.* [35]. For the early diagnosis of ovarian cancer, the efficient fuzzy if then rules from mass spectra of blood samples is extracted by Assareh and Moradi [36]. The feature extraction and analysis of ovarian cancer proteomic mass spectra was done by Meng *et al.* [37]. The multiple biomarker combination by Logistic Regression was explored for early screening of ovarian cancer by Kim *et al.* [38]. The organization of the work is as follow. The details of the dataset along with the first level feature selection techniques are explained in section 2. The stochastic optimization techniques used in this work for second level feature selection is employed in section 3. The classification details are given in section 4 followed by results and discussion in section 5 and concluded in section 6.

II. MATERIALS AND METHODS

For the Ovarian Cancer classification, a dataset was used which is publicly available online [39]. There are about 15154 genes here. There are 253 samples totally where Class 1 represents the normal class with 91 samples and Class 2 represents the cancer class with 162 samples. The details of the dataset are tabulated in Table 1. The illustration of the work is found in Fig. 1.

TABLE 1. Dataset details.

Dataset	Number of genes	Class 1 (Normal)	Class 2 (Cancer)	Total samples
Ovarian Cancer	15154	91	162	253

The pictorial representation of the work is depicted in Figure 1.

A. GENE SELECTION TECHNIQUES

A few top ranked features are selected here. The main intention of this work lies in extracting the best 5000 Genes from 15154 Genes. The selected 5000 genes will certainly undergo a second level optimization with the help of stochastic optimization techniques.

1) CORRELATION COEFFICIENT

The correlation of genes [40] is represented with various samples and is computed as follows:

$$C_c(g) = \frac{n_r \Sigma_{g*} IG - \Sigma G Ig}{\sqrt{n_r \Sigma G^2 - (\Sigma G)^2 \left(n_r \Sigma IG^2 - (\Sigma IG)^2\right)}} \quad (1)$$

where the number of records are represented as n_r, G indicates the particular gene value. The most useful and

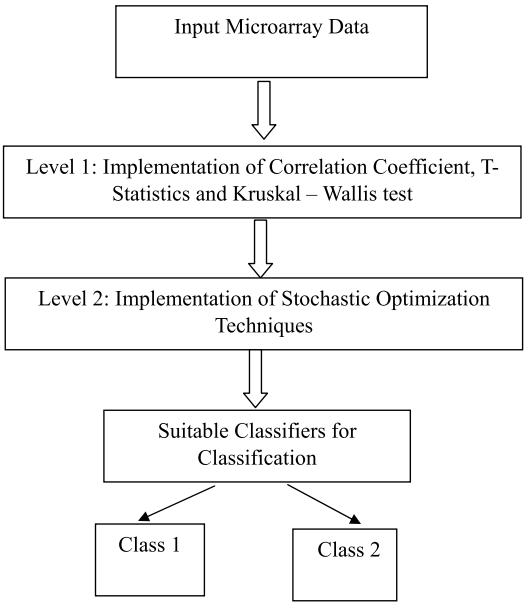

FIGURE 1. Illustration of the work.

informative gene in the dataset is represented as *Ig* and *IG* represents the most significant gene for the class level prediction. For the class partitioning problem, the most informative genes are the one having a higher correlation coefficient value.

2) T-STATISTICS

The gene expression dataset with the help of t-statistics approach [41], and hence the ranking of the differentially expressed genes are done. The calculation of the t-statistic for a particular gene is expressed as follows:

$$t_g = \frac{mean_1 - mean_2}{\sqrt{\frac{\text{var}_1}{n_1} + \frac{\text{var}_2}{n_2}}} \tag{2}$$

With respect to the sample type, the relative difference of gene value is represented by t_g value. Thus, with the help of top ranked t value, the extraction of some differentially expressed genes is done. The mean and variance of one category of patient samples of size n_1 is represented by $mean_1$ and var_1 respectively. The mean and variance of the second/other category of patient samples of size n_2 is represented by $mean_2$ and var_2 respectively.

3) KRUSKAL–WALLIS TEST

For testing and analyzing about the samples spreading from the same partition, a famous non parametric ranking based technique is developed and known as Kruskal-Wallis test [42]. A one-way Analysis of Variance (ANOVA) is equivalent to Kruskal – Wallis. The class membership is usually ignored, and the ranking of the data is done from 1 to N. The ranking of all the data from all the classes together is done here. The score for the Kruskal-Wallis test K_w is computed as

$$K_w = \frac{12}{N(N+1)} \sum_{j=1}^{g} n_j \left(\bar{r}_j - \frac{N+1}{2} \right)^2 \tag{3}$$

where the number of examination in class j is represented as n_j. The total number of experiments in all classes is represented as N.

$$\bar{r}_j = \frac{\sum_{k=1}^{n_j} r_{jk}}{n_j} \tag{4}$$

Among all the observations, r_{jk} is the rank of experiment k from class j.

III. OPTIMIZATION TECHNIQUES

The 5000 features selected through the standard genes selection techniques are further optimized with a second level feature selection by means of utilizing stochastic optimization techniques to select the top 50–150 genes. Stochastic optimization techniques are the optimization methods where the random variables are generated and used. In the formulation of the optimization problem itself, the random variables appear for the stochastic problems. Some stochastic optimization techniques have methodologies involved with random iteration too. The generalization of deterministic method for deterministic problems is done by Stochastic optimization methods.

A. CENTRAL FORCE OPTIMIZATION

Here in the entire population, every individual is termed as a probe. Based on the defined masses, the probes are attracted by gravitation. The objects are considered as probes and fitness function is utilized to assess its performance [43]. A solution is represented by each mass and based on newton's universal law of gravitation; the position is adjusted properly by means of navigation. The algorithm mainly consists of 3 steps such as (a) initialization (b) calculation of the acceleration of the probe (c) motion factor.

Initially, in the search space, the creation of a population of probes is done. The start position and acceleration vector is assigned to zero. Secondly, based on the Newton universal law, the calculation of the compound acceleration vector of one probe from appropriate components in every direction is done. The user-defined function is Mass and it is obtained from the objective function which has to be minimized. In a ND-dimensional search space with NQ probes $y^q = \left(y_1^q, y_2^q, \ldots, y_{ND}^q\right), (q = 1, 2, \ldots, NQ)$, the operation of the q^{th} probe is computed based on the following formula as

$$Z_t^q = V \sum_{\substack{n=1 \\ n \neq q}}^{NQ} P\left(F_{t-1}^n - F_{t-1}^q\right)\left(F_{t-1}^n - F_{t-1}^q\right)^{\alpha} \times \left(y_{t-1}^n - y_{t-1}^q\right) \left\| y_{t-1}^n - y_{t-1}^q \right\|^{-\beta} \tag{5}$$

where y_t^q, Z_t^q represent the position and acceleration vectors for the q^{th} probe at the t^{th} generation. $F_{t-1}^q = f\left(y_1^{q,t-1}, y_2^{q,t-1}, \ldots, x_{ND}^{q,t-1}\right)$ indicates the fitness of the q^{th} probe which gives the objective value and $P(w)$ indicates a piece wise function and the number of iterations

TABLE 2. Performance analysis of classifiers in terms of classification accuracies with CFO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	95.69897	97.3975	97.395
	100	97.395	84.21453	78.125
	150	91.1475	91.67	94.795
KNN	50	93.23	96.355	91.66875
	100	86.78594	94.2725	84.961
	150	92.71	81.77	85.67875
LR	50	83.594	95.06961	86.48672
	100	81.25	98.4375	91.96164
	150	90.1125	85.42	91.6675
SVM (with RBF Kernel)	50	96.7125	93.29656	92.71
	100	91.4075	97.36406	97.3975
	150	92.19	85.42	93.7525
MLP	50	89.6	86.71875	95.315
	100	87.5	93.75	96.87625
	150	85.48219	98.96	80.21
Average		**90.32107**	**92.00773**	**90.60004**

as represented as t.

$$P(w) = \begin{cases} 1 & w \geq 0 \\ 0 & otherwise \end{cases} \tag{6}$$

In (5), V, α and β do not include the fundamentals of concrete gravitational elements. In order to avoid the probe from flying far away from the search space, detection of the probe positions is absolutely necessary. The probe is pulled back to its original search space if it is out of range.

Thirdly, based on the previous calculation of acceleration, depending on the Newtonian formula, the updation of the position vectors of the probes are done. Once the exertion of the acceleration Z_t^q is done, then the movement of the probe q from y_t^q to y_{t+1}^q is done in accordance to the motion equation as

$$y_{t+1}^q = y_t^q + 0.5 \times Z_t^q \Delta t^2 \tag{7}$$

where the step is represented as Δt

Depending on the last ''mass'' information, the updation of probe position is done as a specific deterministic gradient algorithm. The revelation of the convergence conditions of CFO that it would converge to the optimal point seeked so far is done. The performance of it is pretty good then the predefined one present in the initial distributions is considered.

The algorithm is shown in the following procedure:

Step 1: Initialization of parameters:

The dimension boundaries ND of the objective function is set, (i.e.), $y^{\min}$, $y^{\max}$, number of probes NQ, acceleration parameters α, β, gravitational constant V.

Step 2: Initialization of population:

The initial probe distribution Y, acceleration vector Z, fitness matrix F and position vectors S is set.

Step 3: (Loop on time step) The following are considered.

3.1 The probe position vector S is computed

3.2 The retrieval of the probe which flies out of the boundary can be done

3.3 For the current probes, the fitness matrices are updated

3.4 For the next time step, the acceleration vectors Z is computed.

Step 4: The time step is increased and unless the stopping criterion has been met, the step 3 is repeated.

B. LIGHTNING ATTACHMENT PROCEDURE OPTIMIZATION

A new and famous nature inspired global optimization algorithm is LAPO where the lightning attachment process is mimicked which includes both the upward leader propagation and downward leader movement [44]. Between electrically charged regions of a cloud, the occurrence of a sudden electrostatic discharge happens, and it is called lightning.

TABLE 3. Performance analysis of classifiers in terms of classification accuracies with LAPO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	81.77	96.355	97.915
	100	97.16906	98.69875	97.655
	150	94.7925	97.9175	85.80813
KNN	50	92.97375	98.4375	87.5
	100	91.93	75.5	91.1475
	150	85.9175	95.055	93.23
LR	50	84.8975	90.625	94.53375
	100	76.54	94.53375	81.77
	150	91.1475	95.055	93.49
SVM (with RBF Kernel)	50	98.66594	98.17625	94.01125
	100	97.1375	95.835	85.67875
	150	93.7525	82.616	83.78925
MLP	50	94.2725	80.21	93.23
	100	84.375	87.10938	95.315
	150	89.0625	92.19	94.53375
Average		**90.29358**	**91.88761**	**91.30716**

In a step wise movement, the lightning always progresses towards or away from the ground. The downward leader stops after every step and then to a randomly selected potential point it moves. The potential point which is randomly selected always has a very high value of electrical field. From the sharp points, the upward leader is started and then progresses towards the downward leader. The effect of branch fading lightning feature takes place when the branch charge is lower than a specific value. A final strike occurs when the two leaders joins together, and the neutralization of the cloud changes takes place.

1) TEST POINTS PARAMETER INITIALIZATION

The vital parameters here are the maximum number of iterations $Iter_{\max}$, the number of decision variables n, the number of test points N_{pop}, the upper bounds $Z_{\max}$ and lower bounds $Z_{\min}$ for decision variables. At the start of the algorithm, these parameters are assigned. An initial population is mandatorily required as it is similar to other Nature Inspired optimization algorithms. In the feasible search space, every population is considered as a test point which serves as an emitting point of either a downward or upward leader. The random initialization of the test points are as follows:

$$Z_{j,k} = Z_{\min} + rand\,() * (Z_{\max} - Z_{\min}),\quad j = 1, 2, .., N_{pop}, k = 1, 2, \ldots, n \tag{8}$$

$rand()$ indicates a uniformly distributed random number in the range of [0,1]. Based on the objective function the fitness value $f = f_1, f_2, \ldots, f_{Npop}$ of every test point is computed as

$$f_j = obj\left(Z_{j,1}, Z_{j,2}, \ldots., Z_{j,n}\right)\quad j = 1, 2, \ldots, N_{pop} \tag{9}$$

2) MOVEMENT OF DOWNWARD LEADER TOWARDS THE GROUND

The consideration of the test points as the downward leader in this phase is done and it moves down towards the group. For all the test points and its respective fitness value, the average values is computed as follows

$$Z_{ave} = mean\left(Z_{j,k}\right)$$
$$f_{ave} = obj(Z_{ave}) \tag{10}$$

A random behavior is present in the lightning process, for test point j, the selection of a random point s among the population ($j \neq s$) is done. Based on the following rules, the updation of the new test points is done.

(i) If the electric field of point s is greater than the average electric field, then

$$Z_{j,k}^{new} = Z_{j,k} + rand\,()^* \left(Z_{ave} - rand\,()^* Z_{s,k}\right) \tag{11}$$

(ii) If the electric field of point s is less than the average electric field, then

$$Z_{j,k}^{new} = Z_{j,k} - rand\,()^* \left(Z_{ave} - rand\,()^* Z_{s,k}\right) \tag{12}$$

TABLE 4. Performance analysis of classifiers in terms of classification accuracies with GBCO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	84.375	95.835	76.27
	100	94.2725	90.1125	85.80813
	150	92.19	94.795	92.71
KNN	50	87.10938	93.23	75.75
	100	98.4375	94.2725	82.942
	150	95.315	85.54938	77.72875
LR	50	82.03	76.54	90.36875
	100	81.38	77.6025	77.34125
	150	79.43	94.53375	92.19
SVM (with RBF Kernel)	50	98.17625	92.19	99.48
	100	88.55	86.13281	81.77
	150	76.54	76.54	91.67
MLP	50	93.88063	81.315	81.25
	100	87.5	89.075	96.875
	150	85.42	90.625	75.75
Average		**88.30708**	**87.8899**	**85.19359**

(iii) The branch sustains if the electric field of the new test points is better than the old one, otherwise it gradually fades. Mathematically, this feature is formulated as

$$Z_{j,k} = \begin{cases} Z_{j,k}^{new}, & if \ f\left(Z_{j,k}^{new}\right) < f\left(Z_{j,k}\right) \\ Z_{j,k}, & otherwise \end{cases} \tag{13}$$

3) MOVEMENT OF UPWARD LEADER

The consideration of all the test points in the upward movement phase as the upward leader towards the cloud is done. The generation of the new test points are done as follows:

$$Z_{j,k}^{new} = Z_{j,k} + rand\ ()^{*}\ E^{*}\ (Z_{best} - Z_{worst}) \tag{14}$$

where Z_{best} and Z_{worst} represent the between and within solutions of the population, E represents the exponent factor that is a function of both the number of iterations $Iter$ and the maximum number of iterations $Iter_{\max}$ as:

$$E = 1 - \left(\frac{Iter}{Iter_{\max}}\right)^{*} \exp\left(\frac{Iter}{Iter_{\max}}\right) \tag{15}$$

For balancing both the exploration and exploitation capabilities of the algorithms, the iteration dependent exponent factors are significant from a computational point of view. The branch fading feature also happens in this phase similar to the downward movement.

4) PERFORMANCE ENHANCEMENT

For the performance enhancement of LAPO, in every iteration, the average test point replaces the worst test point if the fitness factor is worse.

$$Z_{worst} = Z_{ave} \quad if \ \ f_{ave} < f\ (Z_{worst}) \tag{16}$$

5) STOP CRITERION

If the maximum number of iterations is satisfied, then the termination of algorithm is done. Otherwise the procedure of downward and upward leader movements is repeated to enhance the performance.

6) THE PROCEDURE OF LAPO

The Procedure of LAPO is given as follows.

(1) Set $Iter_{\max}, n, N_{pop}, Z_{\max}, Z_{\min}$

(2) Test Points Random Initialization

(3) $Z_{j,k} = Z_{\min} + rand()^{*}\ (Z_{\max} - Z_{\min})\ , j = 1, 2, \ldots, N_{pop}$, $k = 1, 2, \ldots, n$

(4) Fitness value computation

(5) $f_j = obj(Z_{j,1}, Z_{j,2}, \ldots, Z_{j,n}), j = 1, 2, \ldots, N_{pop}$

(6) While $Iter < Iter_{\max}$

(7) Compute the average values for all the test point and fitness value

(8) $Z_{ave} = mean\left(Z_{j,k}\right)$

(9) $f_{ave} = obj\left(Z_{ave}\right)$

TABLE 5. Performance analysis of classifiers in terms of classification accuracies with AAO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	93.36	86.52344	98.96
	100	85.22406	89.6	87.10938
	150	76.3375	93.75	76.97875
KNN	50	78.51688	91.47406	76.3375
	100	95.055	95.575	96.875
	150	91.67	91.47406	92.97
LR	50	99.48	92.71	94.66438
	100	76	93.36	81.9
	150	77.08	85.84047	90.625
SVM (with RBF Kernel)	50	81.9	88.025	90.49688
	100	87.5	91.93	93.23
	150	98.96	85.74344	98.4375
MLP	50	94.53375	91.93	76.135
	100	80.73	89.20625	94.66438
	150	76.3375	78.45156	85.9375
Average		**86.17898**	**89.70622**	**89.02142**

(10) if $f_{ave} < f\ (Z_{worst})$
(11) $Z_{worst} = Z_{ave}$
(12) end
(13) Movement of downward leader towards the group
(14) for $j = 1 : N_{pop}$
(15) Random selection $Z_{s,k}\ (Z_{s,k} \neq Z_{j,k})$
(16) if $f_{ave} < f\ (Z_{s,k})$
(17) $Z_{j,k}^{new} = Z_{j,k} + rand\ ()^* \left(Z_{ave} - rand\ ()^* Z_{s,k}\right)$
(18) end
(19) $Z_{j,k}^{new} = Z_{j,k} - rand\ ()^* \left(Z_{ave} - rand\ ()^* Z_{s,k}\right)$
(20) end
(21) Compute fitness value of new test points
(22) if $f\left(Z_{j,k}^{new}\right) < f\left(Z_{j,k}\right)$
(23) $Z_{j,k} = Z_{j,k}^{new}$
(24) end
(25) end
(26) Movement of upward leader
(27) for $j = 1 : N_{pop}$
(28) $E = 1 - \left(Iter/Iter_{\max}\right)^* \exp\left(Iter/Iter_{\max}\right)$
(29) $Z_{j,k}^{new} = Z_{j,k} + rand\ ()^* E^* \left(Z_{best} - Z_{worst}\right)$
(30) if $f\left(Z_{j,k}^{new}\right) < f\left(Z_{j,k}\right)$
(31) $Z_{j,k} = Z_{j,k}^{new}$
(32) end
(33) end
(34) $Iter = Iter + 1$
(35) end

C. GENETIC BEE COLONY OPTIMIZATION

By incorporating the advantages of Genetic Algorithm and Artificial Bee Colony (ABC) algorithm, a new optimization algorithm called GBC was developed for the sake of optimizing numerical problems [45]. The colony of the artificial bees in the ABC algorithm is classified into three different kinds such as employed artificial bees, onlookers' bees, and scouts' artificial bees. The following steps are done in ABC as follows:

1) ABC PARAMETER SETTINGS

Initialization of the main parameters of the algorithm should be done. The population size or solution, the limit parameter (L) and the number of bees that are considered to be double the size of the population size are the parameters considered.

2) INITIALIZATION OF THE ENTIRE POPULATION OF SOLUTION

By means of random generation, the solution with equal size to population size is expressed as

$$w = w_{j,k}^{\min} + rand[0, 1]\left(w_{j,k}^{\max} - w_{j,k}^{\min}\right) \quad (17)$$

TABLE 6. Performance analysis of classifiers in terms of performance index with CFO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	89.75052	94.725	94.49
	100	94.49	48.63844	22.22
	150	78.465	80.01	88.38
KNN	50	84.315	92.455	79.9675
	100	64.44	87.04	56.86075
	150	82.93	42.58	59.85
LR	50	51.163	88.56547	62.70266
	100	40	96.76	78.77609
	150	77.925	58.83	79.925
SVM (with RBF Kernel)	50	93.14563	84.40375	82.93
	100	79.195	94.30813	94.4575
	150	81.47	58.83	85.655
MLP	50	78.93	63.765	88.94
	100	66.66	85.7	92.8025
	150	58.60375	97.87	34.295
Average		**74.76553**	**78.29872**	**73.48347**

where the solution index is represented as j, the decision variable is defined as k, the generation of a random variable between 0 and $w_{j,k}^{\min} - w_{j,k}^{\max}$ is done as $rand[0, 1]$. The lower and upper limits of the k^{th} decision variable is represented as $w_{j,k}^{\min}$ and $w_{j,k}^{\max}$.

3) POPULATION SOLUTION EVALUATION

To assess the obtained generated solutions, the objective functions are utilized.

4) THE EMPLOYEE BEE

In this phase, a new source of food is being discovered by every employed bee in the surrounding area of its location. Then the movement of the employed bees into its candidate neighbour solutions is done, so that the food source is there to every employed bees in the surrounding environment. The evaluation of the nectar amount in the detected food sources is done. If the nectar amount of the detected source of food is greater than the nectar amount of the present resources of food, then the memorization of the detected food source is done immediately. By the modification of the j^{th} solution, a neighbour solution 'n' can be obtained as expressed in the equation as follows:

$$n_{j,k} = w_{j,k} + \theta_{j,k}\left(w_{j,k} - w_{s,k}\right) \tag{18}$$

where s is a solution which is selected from population size randomly and θ is also randomly selected in the range of $[-1,1]$.

5) ONLOOKER BEE

To detect the new food source in the neighbourhood area, onlooker bees are used. The information that has been obtained from the previous phase from the employed bees is made effective use of here in the exploitation process. By means of exploration of their neighbourhood using equation (18), the current solutions are tried to be improved by both the onlooker bees and employee bees. The onlooker bees can select the solutions by exploiting the fitness values according to the following equation as

$$q_j = \frac{fit_j}{\sum_{k=1}^{PS} fit_k} \tag{19}$$

6) SCOUT BEE

When the detection of food source is done, the employee bee becomes a scout bee so that a new source of food is found in the solution space. To indicate the number of trials, a parameter named limit is used to control the number of scout bees. When the source of food cannot be improvised or developed,

TABLE 7. Performance analysis of classifiers in terms of performance index with LAPO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	42.58	92.015	95.65
	100	94.14719	97.2675	95.07
	150	88.13	95.79	60.36
KNN	50	84.7525	96.76	66.66
	100	80.74	3.92	78.465
	150	49.635	89.18	84.315
LR	50	52.755	76.92	87.71
	100	11.525	87.71	42.58
	150	77.52	89.18	85.0075
SVM (with RBF Kernel)	50	97.18344	96.205	86.37
	100	93.925	91.58	59.85
	150	85.655	46.66075	52.00725
MLP	50	87.04	34.295	84.315
	100	54.54	65.2125	89.98
	150	70.12	81.47	87.71
Average		**71.34988**	**76.27772**	**77.06998**

then a random determination of a new source of food needs to be done. Thus, in the search space, exploitation and exploration processes should be carried out together.

7) INVOLVEMENT OF GENETIC OPERATORS

By utilizing some genetic operations such as cross over and swap, a new binary version of ABC algorithm is proposed as GBC. To the equations of (17) and (18) in ABC algorithm, the modifications are made, and the generations of the initial solutions is done by equation (20) instead of (17).

$$W_j : j = 1, \ldots, SN \quad w_{jk} = \begin{cases} 0, & if \;\; V(0,1) \leq 0.5 \\ 1, & if \;\; V(0,1) > 0.5 \end{cases} \tag{20}$$

where $V(0, 1)$ is a generated uniformly value.

Within the following four steps, the integration of ABC with GA for search mechanism is utilized.

i) Two sources of food from the population is selected randomly in the neighbourhood of current food source, so that a proposed solution can be found out
ii) Between the current two neighborhoods along with the best and zero food sources the two-point cross over operators are applied in order to generate the sources of children food
iii) To the sources of children food, the second operator called swap operator is applied so that the grandchildren sources of food is found out.
iV) Among the children and grandchildren food sources, the selection of best sources of food as a neighbourhood source of food is done. Thus, in a binary optimization problem, the performance of this ABC is improved with the inclusion of GA.

D. ARTIFICIAL ALGAE OPTIMIZATION

Mimicking the living styles and behaviours of microalgae, AAA was developed [46]. Microalgae lifestyles such as algal tendency, adaption of the surrounding quality, reproduction etc are considered as the major simulation factors by this algorithm. Three vital process called evolutionary process, helical movement and adaption phase are present in this algorithm. Algal colonies are comprising in this population of this algorithm. When enough light is received by the algal cells in algal colonies, then it grows into a bigger size. When insufficient light condition occurs, then there may not be sufficient algal colony growth. In the helical movement, only towards the best algal colony, the movement of every algal colony would be present. To explain the main process of AAA, assume $x_j = (x_{j1}, x_{j2}, \ldots, x_{1m})$, where $j = 1, 2, \ldots, n$

TABLE 8. Performance analysis of classifiers in terms of performance index with GBCO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	54.54	91.58	9.6075
	100	87.04	77.925	60.36
	150	81.47	88.38	82.93
KNN	50	65.2125	84.315	5.805
	100	96.76	87.04	48.1615
	150	89.98	59.34	19.54625
LR	50	43.87	11.525	77.4225
	100	40.645	18.79	17.075
	150	30.01625	87.71	81.47
SVM (with RBF Kernel)	50	96.205	81.47	98.935
	100	72.795	61.59375	42.58
	150	11.525	11.525	80.01
MLP	50	86.035	40.3225	40
	100	66.66	75.8625	93.33
	150	58.83	76.92	5.805
Average		**65.43892**	**63.61992**	**50.86918**

and the solution in search space is expressed by x_j. The following matrix is utilized to represent the algae population as follows:

$$Population = \begin{bmatrix} x_{11} & x_{12} & \ldots. & x_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots. & x_{nm} \end{bmatrix} \tag{21}$$

Assuming that the algal colony size of j^{th} colony is C_j, where $j = 1, 2, \ldots, n$, and the $f(x_j)$ is represented as the objective function, then C_j is updated as follows using the mathematical equations as

$$C_j = size\left(x_j\right) \tag{22}$$

$$\mu_j = \frac{C_j + 4f\left(x_j\right)}{C_j + 2f\left(x_j\right)} \tag{23}$$

$$C_j^{t+1} = \mu_j C_j^t, \quad j = 1, 2, \ldots, n \tag{24}$$

where the update coefficient of C_j represents μ_j and the current generation is represented as t.

1) HELICAL MOVEMENT PHASE

The movement of the algal colony is usually in 3D. Using the following equations, the movement of algal colony in 3D is expressed as follows

$$X_{jg}^{t+1} = X_{jg}^t + \left(X_{kg}^t - X_{jg}^t\right)\left(cf - \sigma_j\right)p \tag{25}$$

$$X_{jl}^{t+1} = X_{jl}^t + \left(X_{kl}^t - X_{jl}^t\right)\left(cf - \sigma_j\right)\cos\alpha \tag{26}$$

$$X_{jh}^{t+1} = X_{jh}^t + \left(X_{kh}^t - X_{jh}^t\right)\left(cf - \sigma_j\right)\sin\beta \tag{27}$$

where the movement in 1D is given by (25), say x, (26) and (27) indicates the movement in other dimensions y and z, say l, g and h indicate the random integers generated uniformly between 1 and d, X_{jg}, X_{jl}, X_{jh} simulates the three coordinates of the j^{th} algal colony, k represents the index of a neighbor algal colony, p represents the independent random number in the range of $(-1,1)$, α, β represent the random degrees between 0 and 2π, shear force is represented as sf, σ_j indicates the friction surface area of the j^{th} algal colony.

2) EVOLUTIONARY PROCESS PHASE

To get a most feasible solution, the algal colony X_j becomes larger as it progresses towards a feasible solution. This simulation process is expressed in the following equations as

$$Biggest = \arg\max\left\{size\left(X_j\right)\right\}, \quad j = 1, 2, \ldots., n \tag{28}$$

$$Smallest = \arg\min\left\{size\left(X_j\right)\right\}, \quad j = 1, 2, \ldots., n \tag{29}$$

$$Smallest_k = Biggest_k. \quad k = 1, 2, \ldots, d \tag{30}$$

TABLE 9. Performance analysis of classifiers in terms of performance index with AAO for different gene selection techniques using 50-100-150 selected genes.

Classifiers	Number of Genes selected	Gene Selection Techniques		
		Correlation Coefficient	T-Statistic	Kruskal-Wallis Test
LDA	50	84.66125	63.04125	97.87
	100	58.02563	78.93	65.2125
	150	4.626875	85.7	14.64094
KNN	50	24.60875	79.43063	20.93375
	100	89.18	90.78	93.33
	150	80.01	79.43063	83.6225
LR	50	98.935	82.93	88.045
	100	4.39	84.66125	43.225
	150	15.36	60.4875	76.92
SVM (with RBF Kernel)	50	43.225	69.7275	77.17125
	100	66.66	80.72125	84.315
	150	97.87	60.105	96.76
MLP	50	87.71	80.72125	8.64875
	100	37.1475	76.62938	88.045
	150	18.14688	24.21063	60.87
Average		**54.03713**	**73.16708**	**66.64065**

where the biggest algal colony is expressed as *Biggest* and the smallest algal colony is expressed as *Smallest*. The algal cell which is randomly selected is indicated by a random value k.

3) ADAPTATION PHASE

When the growth of the algal colony is not sufficient, it can adapt itself to the surrounding environment. After the adaptation movement, the objective function value is considered as inferior or superior. The highest starvation value is obtained after the algal colony movement is obtained after the algal colony movement is completed as shown in (31). With an adaptation probability A_p, the adaptation to the biggest algal colony is represented as:

$$X_c = \arg\max \left\{ starvation\left(X_j\right) \right\}, \quad j = 1, 2, \ldots, n \tag{31}$$

For the algal colony phase, the adaptation phase of AAA is expressed as follows:

$$X_{ck}^{t+1} = \begin{cases} X_{ck}^{t} + \left(Biggest_k - X_{ck}^{t}\right).Rand1, \\ \quad if \;\; Rand2 < M_p.k = 1, 2, \ldots, d \\ X_{ck}^{t} \\ \quad otherwise \end{cases} \tag{32}$$

where the algal colony index which has the highest starvation value is expressed by c. To assess the starvation level of algal colony X_j, starvation X_j is utilized, where the algal cell index is represented as k. A_p represents the adaptation probability and it considers value between 0.3 and 0.7. The $Rand1$ and $Rand2$ generates random values between 0 and 1.

IV. CLASSIFICATION PROCEDURES

The following classification models are utilized here. All the models used here belongs to well established group of Machine Learning Algorithms.

A. LINEAR DISCRIMINANT ANALYSIS (LDA)

A very simple with great utility is possessed by LDA [47]. When the two classes $c = \{0, 1\}$ is assumed to have a Gaussian distribution with a specific mean μ_c, and the similar covariance matrix Σ is shared by them, then the Linear discriminant function $\delta_c(y)$, $c = \{0, 1\}$ is expressed by

$$\delta_c(y) = y^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log\left(\pi_c\right) \tag{33}$$

where the frequency of occurrence of class labels is expressed by π_c. The class labels which are predicted is expressed as

$$f(y) = \arg\max_{c=(0,1)} \left\{ \delta_c\left(y\right) \right\} \tag{34}$$

LDA is thus conceptually very robust, simple, and fast. It is very popular in high dimensional problems too.

B. NEAREST NEIGHBOUR CLASSIFIER

A weighted average over the labels y_i is considered for these observations y_i in the training set that are near to the query point q. This is expressed as

$$f(q) = \frac{1}{\sum d_i} \sum_{y_i \in N_k(q)} d_i y_i \tag{35}$$

where the k-element neighborhood of q is denoted by $N_k(q)$, d_i denotes the related distance in the given metric. The metric choice and the number of neighbours are the parameters of the model. As this approach is based on concept of similarity, a very intuitive approach is provided to classification problems by K-Nearest Neighbor (KNN) [48].

C. LOGISTIC REGRESSION MODEL

For binomial distributed dependent variables, Logistic Regression model is extensively used for medical applications [49]. It is almost similar to LDA except for the way of estimation of the linear coefficients. The probability of the dichotomic variable v can be computed by the binary Log. Reg. model from the n independent variables z:

$$v = \frac{1}{1 + \exp(y)} \tag{36}$$

where

$$y = a_0 + \sum_{i=1}^{n} a_i z_i \tag{37}$$

With the help of a second order gradient descent the estimation of model coefficients is done. As these calculations are highly memory and time consuming, it could be a hassle to apply it in high dimensional problems.

D. SVM

In machine learning, SVMs are one of the powerful tools utilized for many applications [50]. A hyperplane is created by the SVM in a feature space so that the data is separated into 2 classes with the maximum margin. Using a positive semi definite function, the mapping of a feature space of the original features (y, y') into a high dimensional space is done as

$$(y, y') \mapsto k(y, y') \tag{38}$$

The kernel function is represented by $K(\bullet, \bullet)$ function and Mercer's condition is used by Kernel trick, which explains that the representation of a dot product in a high dimensional space is mentioned by any positive semi-definite kernel $k(y, y')$. The standard kernels utilized generally are as follows:

$$k(y, y') = (y \bullet y') \text{ Linear} \tag{39}$$

$$k(y, y') = (y \bullet y' + 1)^d \text{ Poly} \tag{40}$$

$$k(y, y') = \exp\left(-\frac{\|y - y'\|}{\sigma^2}\right) \text{RBF} \tag{41}$$

The model parameters are with respect to the Kernel type, the polynomial degree d and the width of the RBF σ^2. In this work only SVM – RBF kernel is used.

E. MULTI-LAYER PERCEPTRON (MLP)

With a sigmoid activation function, the training of a multilayer feed forward Neural Network is done [51]. With Gaussian distributed random numbers, the initialization of weights is done which has scaled variance and a zero mean. Gradient descent with Back propagation is used to train the weights. Common weight decay is found in MLP which has a penalty term represented as

$$P(\vec{v}) = \lambda \sum_{i=1}^{N} -\frac{v_i^2}{1 + v_i^2} \tag{42}$$

where the N-dimensional weight vector of the MLP is expressed as $\vec{v}$, λ represents a small regularization parameter. During the cross-validation training, the number of neurons, number of regularization parameters and the number of hidden layers is adjusted to get a minimum error loss.

V. RESULTS AND DISCUSSION

It is classified with a 10-fold cross validation method and the performance of it is shown in tables below. The mathematical formulae for computing the Performance Index (PI), Sensitivity, Specificity and Accuracy is mentioned in literature and using the same, the values are computed and exhibited [52]. PC is Perfect Classification; MC is Missed Classification and FA is False Alarm in the expressions below.

The sensitivity is computed as

$$Sensitivity = \frac{PC}{PC + FA} \times 100 \tag{43}$$

Specificity is computed as

$$Specificity = \frac{PC}{PC + MC} \times 100 \tag{44}$$

Accuracy is expressed as

$$Accuracy = \frac{Sensitivity + Specificity}{2} \tag{45}$$

Performance Index (PI) is expressed as

$$PI = \left(\frac{PC - MC - FA}{PC}\right) \times 100 \tag{46}$$

Table 2 depicts the Performance Analysis of Classifiers in terms of Classification Accuracies with CFO for different gene selection techniques using 50-100-150 selected genes. As reported in the Table 2 that MLP with 150 genes selected for T-Statistic gene selection Method achieved high accuracy of 98.96%. Low accuracy value of 78.125% is attained at Kruskal-Wallis Test gene selection in LDA Classifier 100 selected genes. T-Statistic gene selection technique attained at high average accuracy of 92.007% across the classifiers.

TABLE 10. Consolidated average performance analysis of classifiers in terms of classification accuracy and performance index with four optimization techniques for Average of gene selection techniques using 50-100-150 selected genes for Ovarian Cancer.

Classifiers	Optimization Method	CFO		LAPO		GBCO		AAO	
	Parameters / Number of Genes selected	Accuracy (%)	Performance Index (%)	Accuracy (%)	Performance Index (%)	Accuracy (%)	Performance Index (%)	Accuracy (%)	Performance Index (%)
LDA	50	96.83049	92.98851	92.01333	76.74833	85.49333	51.90917	92.94781	81.8575
	100	86.57818	55.11615	97.84094	95.4949	90.06438	75.10833	87.31115	67.38938
	150	92.5375	82.285	92.83938	81.42667	93.23167	84.26	82.35542	34.98927
KNN	50	93.75125	85.57917	92.97042	82.72417	85.36313	51.7775	82.10948	41.65771
	100	88.67315	69.44692	86.1925	54.375	91.884	77.3205	95.835	91.09667
	150	86.71958	61.78667	91.40083	74.37667	86.19771	56.28875	92.03802	81.02104
LR	50	88.38344	67.47704	90.01875	72.46167	82.97958	44.2725	95.61813	89.97
	100	90.54971	71.84536	84.28125	47.27167	78.77458	25.50333	83.75333	44.09208
	150	89.06667	72.22667	93.23083	83.9025	88.71792	66.39875	84.51516	50.9225
SVM (with RBF Kernel)	50	94.23969	86.82646	96.95115	93.25281	96.61542	92.20333	86.80729	63.37458
	100	95.38969	89.32021	92.88375	81.785	85.48427	58.98958	90.88667	77.23208
	150	90.45417	75.31833	86.71925	61.441	81.58333	34.35333	94.38031	84.91167
MLP	50	90.54458	77.21167	89.2375	68.55	85.48188	55.4525	87.53292	59.02667
	100	92.70875	81.72083	88.93313	69.91083	91.15	78.6175	88.20021	67.27396
	150	88.2174	63.58958	91.92875	79.76667	83.93167	47.185	80.24219	34.40917
Average		**90.97628**	**75.5159**	**91.16278**	**74.89919**	**87.13019**	**59.97601**	**88.3022**	**64.61495**

TABLE 11. Consolidated average performance analysis of classifiers in terms of classification accuracy and performance index with three gene selection techniques for average of four optimization techniques across the classifiers for Ovarian Cancer.

Parameters	Gene Selection Techniques					
	Correlation Coefficient		T-Statistic		Kruskal-Wallis test	
Optimization Methods	Accuracy(%)	Performance Index (%)	Accuracy (%)	Performance Index (%)	Accuracy (%)	Performance Index (%)
CFO	90.32107	74.76553	92.00773	78.29872	90.60004	73.48347
LAPO	90.29358	71.34988	91.88761	76.27772	91.30716	77.06998
GBCO	88.30708	65.43892	87.8899	63.61992	85.19359	50.86918
AAO	86.17898	54.03713	89.70622	73.16708	89.02142	66.64065
Average	**88.77518**	**66.39786**	**90.37286**	**72.84086**	**89.03055**	**67.01582**

Table 3 displays the Performance Analysis of Classifiers in terms of Classification Accuracies with LAPO for different gene selection techniques using 50-100-150 selected genes. As shown in the Table 3 that LDA Classifier with 100 genes selected for T-Statistic gene selection Method achieved high accuracy of 98.698%. Low accuracy value of 75.5% is arrived at same T-Statistic gene selection in KNN Classifier 100 selected genes. T-Statistic gene selection technique maintained at high average accuracy of 91.887% across the classifiers.

Table 4 demonstrates the Performance Analysis of Classifiers in terms of Classification Accuracies with GBCO for different gene selection techniques using 50-100-150 selected genes. As mentioned in the Table 4 that SVM (with RBF Kernel) Classifier with 50 genes selected for Kruskal-Wallis Test gene selection Method achieved high accuracy of 99.48%. Low accuracy of 75.75% is ebbed at same Kruskal-Wallis Test gene selection in KNN Classifier 50 selected genes. Correlation Coefficient gene selection technique maintained at high average accuracy of 88.307% across the classifiers.

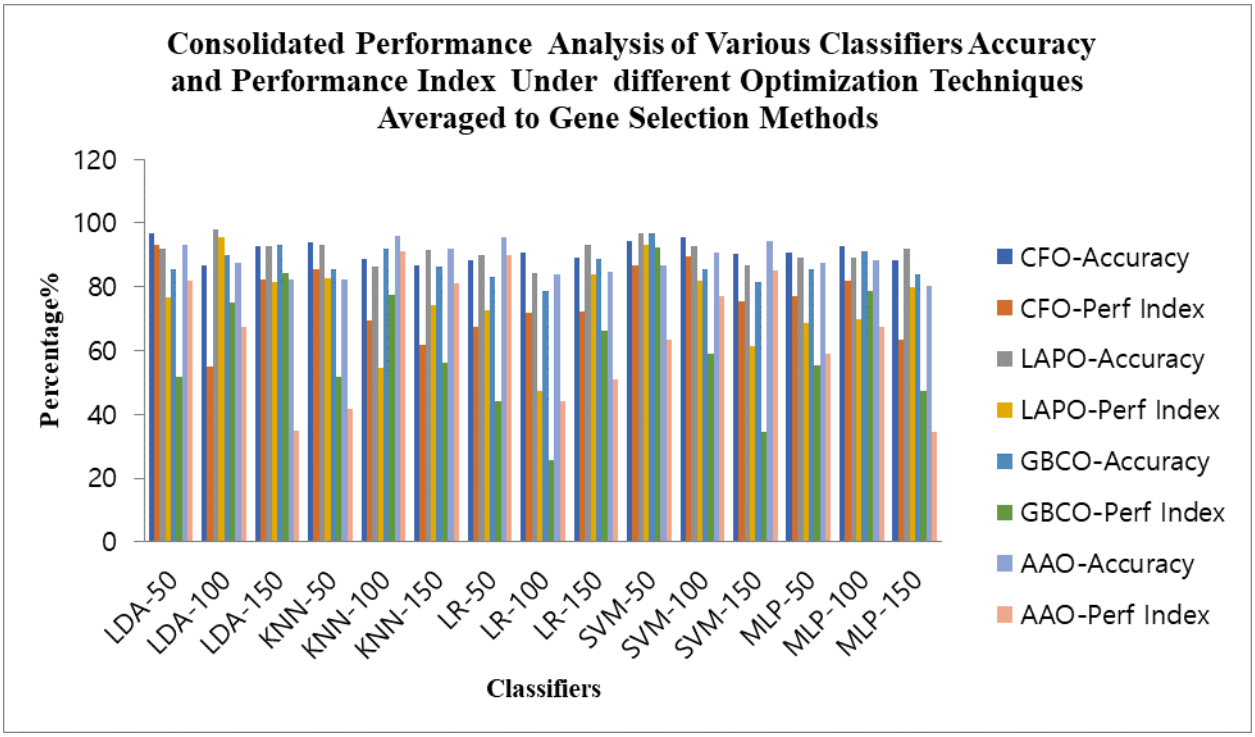

FIGURE 2. Consolidated performance analyses of various classifiers accuracy and performance index under different optimization techniques averaged to gene selection methods for ovarian cancer.

Table 5 indicates the Performance Analysis of Classifiers in terms of Classification Accuracies with AAO for different gene selection techniques using 50-100-150 selected genes. As reported in the Table 5 that LR Classifier with 50 genes selected for Correlation Coefficient gene selection Method achieved high accuracy of 99.48%. Low accuracy of 76% is arrived at same correlation coefficient gene selection in LR Classifier 100 selected genes. T-Statistic gene selection technique maintained at high average accuracy of 89.706% across the classifiers.

Table 6 reports the Performance Analysis of Classifiers in terms of Performance Index with CFO for different gene selection techniques using 50-100-150 selected genes. As reported in the Table 6 that MLP with 150 genes selected for T-Statistic gene selection Method achieved high Performance Index of 97.87%. Low Performance Index of 22.22% is attained at Kruskal-Wallis Test gene selection in LDA Classifier 100 selected genes. T-Statistic gene selection technique attained at high average Performance Index of 78.29% across the classifiers.

Table 7 demonstrates the Performance Analysis of Classifiers in terms of Performance Index with LAPO for different gene selection techniques using 50-100-150 selected genes. As indicated in the Table 7 that LDA Classifier with 100 genes selected for T-Statistic gene selection Method achieved high Performance Index of 97.267%. Low Performance Index of 3.92% is retained at same T-Statistic gene selection in KNN Classifier 100 selected genes. T-Statistic gene selection technique maintained at high average Performance Index of 76.277% across the classifiers.

Table 8 reveals the Performance Analysis of Classifiers in terms of Performance Index with GBCO for different gene selection techniques using 50-100-150 selected genes. As mentioned in the Table 8 that SVM (with RBF Kernel) Classifier with 50 genes selected for Kruskal-Wallis Test gene selection Method achieved high Performance Index of 98.935%. Low Performance Index of 5.805% is also ebbed at same Kruskal-Wallis Test gene selection in KNN Classifier 50 selected genes. Correlation Coefficient gene selection technique maintained at high average Performance Index of 65.4389% across the classifier.

Table 9 indicates the Performance Analysis of Classifiers in terms of Performance Index with AAO for different gene selection techniques using 50-100-150 selected genes. As demonstrated in the Table 9 that LR Classifier with 50 genes selected for Correlation Coefficient gene selection Method achieved high Performance Index of 98.935%. Low Performance Index of 4.39% is arrived at same correlation coefficient gene selection in LR Classifier 100 selected genes. T-Statistic gene selection technique maintained at high average Performance Index of 73.167% across the classifiers.

Table 10 shows the Consolidated Average Performance Analysis of Classifiers in terms of Classification Accuracy and Performance Index with four optimization techniques for Average of gene selection techniques using 50-100-150 selected genes for Ovarian Cancer. As indicated in the Table 10 that LAPO method with LDA classifier of 100 genes selected achieved high accuracy of 97.84% and Performance Index of 95.49%. The LR classifier with 100 genes selected for GBCO optimization reached low value

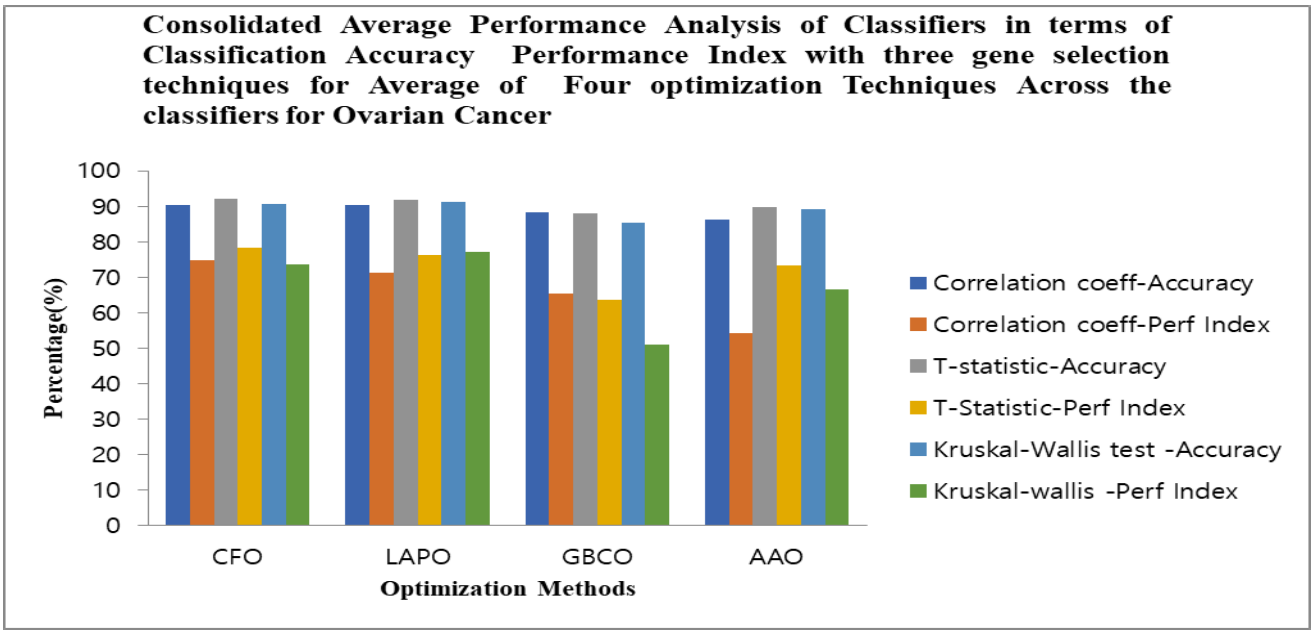

FIGURE 3. Consolidated average performance analysis of classifiers in terms of classification accuracy performance index with three gene selection techniques for average of four optimization techniques across the classifiers for ovarian cancer.

of accuracy 78.77% and Performance Index of 25.503%. Across the classifiers the LAPO optimization method reaches high average accuracy of 91.162% and along with Performance Index of 74.89%. The LAPO method outperforms other three optimization methods in terms of accuracy and Performance Index.

Table 11 displays the Consolidated Average Performance Analysis of Classifiers in terms of Classification Accuracy and Performance Index with three gene selection techniques for Average of four optimization techniques across the classifiers for Ovarian Cancer. As indicated in the Table 11 that T-Statistic gene selection technique in CFO optimization method retained at high accuracy of 92.0077% along with Performance Index of 78.29%. Once again T-Statistic gene selection methods scores high average accuracy of 90.37% and Performance Index of 72.84% across the four optimization methods.

Fig. 2. displays the Consolidated Performance Analyses of Various Classifiers Accuracy and Performance Index under different Optimization Techniques Averaged to Gene Selection Methods for Ovarian Cancer. As demonstrated in the Figure 2 that LAPO method with LDA classifier of 100 genes selected achieved high accuracy of 97.84% and Performance Index of 95.49%. The LR classifier with 100 genes selected for GBCO optimization reached low value of accuracy 78.77% and Performance Index of 25.503%. Across the classifiers the LAPO optimization method reaches high average accuracy of 91.162% and along with Performance Index of 74.89%.

Fig. 3. depicts the Consolidated Average Performance Analysis of Classifiers in terms of Classification Accuracy Performance Index with three gene selection techniques for Average of Four optimization techniques Across the classifiers for Ovarian Cancer. As shown in the Fig. 3. that T-Statistic gene selection technique in CFO optimization method retained at high accuracy of 92.077% along with Performance Index of 78.29%. Once again T-Statistic gene selection methods scores high average accuracy of 90.37% and Performance Index of 72.84% across the four optimization methods.

VI. CONCLUSION AND FUTURE WORK

The most common gynecological malignancy is ovarian cancer. To determine the diagnosis correctly, Computer Aided Diagnosis is absolutely necessary. Monitoring the expression levels of thousands of genes in a simultaneous manner under specific conditions is enables by Micro array technology. Microarray technology makes it possible for the analysis of gene expressions and tremendous amount of data is generated. As a result, due to the curse of dimensionality problem along with a small sample space, processing it further is very difficult. Therefore, in this paper, a two-level feature selection process is proposed, first with the standard gene selection techniques and then the with the implementation of optimization techniques before proceeding to classification. The second-best results are produced when T-static test results are further optimized with both CFO and LAPO and classified with MLP and LDA giving a classification accuracy of 98.96% and 98.69% respectively. the best results are projected when Kruskal Wallis test with GBCO is conducted and classified with SVM -RBF Kernel technique giving a high classification accuracy of 99.48%. Similar results are also obtained when Correlation Coefficient test with AAO is conducted and classified with Logistic Regression giving

a high classification accuracy of 99.48%. Future works is to utilize a variety of other stochastic optimization techniques for the analysis of ovarian cancer classification.

REFERENCES

[1] A. K. Folkins, E. A. Jarboe, A. Saleemuddin, Y. Lee, M. J. Callahan, R. Drapkin, and C. P. Crum, "A candidate precursor to pelvic serous cancer (p53 signature) and its prevalence in ovaries and fallopian tubes from women with BRCA mutations," *Gynecologic Oncol.*, vol. 109, no. 2, pp. 168–173, 2008.

[2] K. Levanon, C. Crum, and R. Drapkin, "New insights into the pathogenesis of serous ovarian cancer and its clinical impact," *J. Clin. Oncol.*, vol. 26, no. 32, pp. 5284–5293, Nov. 2008.

[3] F. Medeiros, M. G. Muto, Y. Lee, J. A. Elvin, M. J. Callahan, C. Feltmate, J. E. Garber, D. W. Cramer, and C. P. Crum, "The tubal fimbria is a preferred site for early adenocarcinoma in women with familial ovarian cancer syndrome," *Amer. J. Surgical Pathol.*, vol. 30, no. 2, pp. 230–236, Feb. 2006.

[4] R. Wooster and B. L. Weber, "Breast and ovarian cancer," *New England J. Med.*, vol. 348, no. 23, pp. 2339–2347, 2003.

[5] W.-H. Wen, A. Reles, I. B. Runnebaum, J. Sullivan-Halley, L. Bernstein, L. A. Jones, J. C. Felix, R. Kreienberg, A. El-Naggar, and M. F. Press, "p53 mutations and expression in ovarian cancers: Correlation with overall survival," *Int. J. Gynecol. Pathol.*, vol. 18, no. 1, pp. 29–41, Jan. 1999.

[6] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, Mar. 2005.

[7] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," *Comput. Statist. Data Anal.*, vol. 48, no. 4, pp. 869–885, Apr. 2005.

[8] J. J. Liu, W. S. Cai, and X. G. Shao, "Cancer classification based on microarray gene expression data using a principal component accumulation method," *Sci. China Chem.*, vol. 54, no. 5, pp. 802–811, 2011.

[9] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[10] B.-C. Lee, K. Cha, S. Avraham, and H. Avraham, "Microarray analysis of differentially expressed genes associated with human ovarian cancer," *Int. J. Oncol.*, vol. 4, pp. 847–851, Apr. 2004.

[11] H. S. Chon and J. M. Lancaster, "Microarray-based gene expression studies in ovarian cancer," *Cancer Control*, vol. 18, no. 1, pp. 8–15, Jan. 2011.

[12] B. Zhang, F. F. Cai, and X. Y. Zhong, "An overview of biomarkers for the ovarian cancer diagnosis," *Eur. J. Obstetrics Gynecol. Reproductive Biol.*, vol. 158, pp. 119–123, Oct. 2011, doi: 10.1016/j.ejogrb.2011.04.023.

[13] Z.-J. Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," *Artif. Intell. Med.*, vol. 42, no. 1, pp. 81–93, Jan. 2008.

[14] J.-T. Jeng, T.-T. Lee, and Y.-C. Lee, "Classification of ovarian cancer based on intelligent systems with microarray data," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Waikoloa, HI, USA, Oct. 2005, pp. 1053–1058.

[15] T. Z. Tan, C. Quek, and G. S. Ng, "Ovarian cancer diagnosis using complementary learning fuzzy neural network," in *Proc. IEEE Int. Joint Conf. Neural Netw.* Piscataway, NJ, USA: IEEE Service Center, Jul. 2005, pp. 3034–3039.

[16] C.-C. Chuang, S.-F. Su, and J.-T. Jeng, "Dimension reduction with support vector regression for ovarian cancer microarray data," in *Proc. IEEE Int. Conf. Syst., Man Cybern.* New York, NY, USA: IEEE Systems, Man, and Cybernetics Society, Oct. 2005, pp. 1048–1052.

[17] G. S. Huang, A. Chen, Y.-C. Hung, and M.-Y. Hong, "Microarray analysis of ovarian cancer," in *Proc. IEEE Int. Conf. Syst., Man Cybern.* New York, NY, USA: IEEE Systems, Man, and Cybernetics Society, Oct. 2005, pp. 1036–1042.

[18] H. Zhu and J. J. Yu, "Gene expression patterns in the histopathological classification of epithelial ovarian cancer," *Exp. Therapeutic Med.*, vol. 1, no. 1, pp. 187–192, 2010, doi: 10.3892/etm_00000030.

[19] J. Yu and X.-W. Chen, "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data," *Bioinformatics*, vol. 21, no. 1, pp. i487–i494, Jun. 2005.

[20] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, "Diagnosis of ovarian cancer using decision tree classification of mass spectral data," *J. Biomed. Biotechnol.*, vol. 2003, no. 5, pp. 308–314, 2003.

[21] M. Wu, C. Yan, H. Liu, and Q. Liu, "Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks," *Bioscience Rep.*, vol. 38, no. 3, Jun. 2018, Art. no. BSR20180289, doi: 10.1042/BSR20180289.

[22] J. Martínez-Más, A. Bueno-Crespo, S. Khazendar, M. Remezal-Solano, J.-P. Martínez-Cendán, S. Jassim, H. Du, H. Al Assam, T. Bourne, and D. Timmerman, "Evaluation of machine learning methods with Fourier transform features for classifying ovarian tumors based on ultrasound images," *PLoS ONE*, vol. 14, no. 7, Jul. 2019, Art. no. e0219388, doi: 10.1371/journal.pone.0219388.

[23] J. Nuhic, L. Sphahic, S. Cordic, and J. Kevric, "Comparative study on different classification techniques for ovarian cancer datasets," in *Proc. Int. Conf. Med. Biol. Eng. (CMBEBIH)*, 2019, pp. 511–518.

[24] Z. Zhang, H. Zhang, and R. C. Bast, "An application of artificial neural networks in ovarian cancer early detection," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN) Neural Comput., New Challenges Perspect. New Millennium*, vol. 4, Jul. 2000, pp. 107–112.

[25] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote, "Bayesian networks in ovarian cancer diagnosis: Potentials and limitations," in *Proc. 13th IEEE Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2000, pp. 103–108.

[26] A. Osmanovic, L. Abdel-Ilah, A. Hodžic, J. Kevric, and A. Fojnica, "Ovary cancer detection using decision tree classifiers based on historical data of ovary cancer patients," in *Proc. CMBEBIH*, 2017, pp. 503–510.

[27] A. Thakur, V. Mishra, and S. K. Jain, "Feed forward artificial neural network: Tool for early detection of ovarian cancer," *Scientia Pharmaceutica*, vol. 79, no. 3, pp. 493–505, 2011.

[28] D. Jelovac and D. K. Armstrong, "Recent progress in the diagnosis and treatment of ovarian cancer," *CA, A Cancer J. Clinicians*, vol. 61, no. 3, pp. 183–203, May 2011.

[29] M. Kusy, "Application of SVM to ovarian cancer classification problem," in *Proc. Int. Conf. Artif. Intell. Soft Comput. (ICAISC)*, 2004, pp. 1020–1025.

[30] A. El-Nabawy, N. El-Bendary, and N. A. Belal, "Epithelial ovarian cancer stage subtype classification using clinical and gene expression integrative approach," *Procedia Comput. Sci.*, vol. 131, pp. 23–30, Jan. 2018.

[31] J. S. Park, S. B. Choi, H. J. Kim, N. H. Cho, S. W. Kim, Y. T. Kim, E. J. Nam, J. W. Chung, and D. W. Kim, "Intraoperative diagnosis support tool for serous ovarian tumors based on microarray data using multicategory machine learning," *Int. J. Gynecol. Cancer*, vol. 26, no. 1, pp. 104–113, Jan. 2016, doi: 10.1097/IGC.0000000000000566.

[32] A. Arfiani and Z. Rustam, "Ovarian cancer data classification using bagging and random forest," in *Proc. 4th Int. Symp. Current Prog. Math. Sci. (ISCPMS)*, vol. 2168, 2019, Art. no. 020046, doi: 10.1063/1.5132473.

[33] U. R. Acharya, S. V. Sree, L. Saba, F. Molinari, S. Guerriero, and J. S. Suri, "Ovarian tumor characterization and classification using ultrasound—A new online paradigm," *J. Digit. Imag.*, vol. 26, no. 3, pp. 544–553, Jun. 2013, doi: 10.1007/s10278-012-9553-8.

[34] L. S. Cohen, P. F. Escobar, C. Scharm, B. Glimco, and D. A. Fishman, "Three-dimensional power Doppler ultrasound improves the diagnostic accuracy for ovarian cancer prediction," *Gynecol. Oncol.*, vol. 82, no. 1, pp. 40–48, Jul. 2001.

[35] C. Renz, J. C. Rajapakse, K. Razvi, and S. K. C. Liang, "Ovarian cancer classification with missing data," in *Proc. 9th Int. Conf. Neural Inf. Process. (ICONIP)*, vol. 2, 2002, pp. 809–813.

[36] A. Assareh and M. H. Moradi, "Extracting efficient fuzzy if-then rules from mass spectra of blood samples to early diagnosis of ovarian cancer," in *Proc. IEEE Symp. Comput. Intell. Bioinf. Comput. Biol.*, Apr. 2007, pp. 502–506.

[37] H. Meng, W. Hong, J. Song, and L. Wang, "Feature extraction and analysis of ovarian cancer proteomic mass spectra," in *Proc. 2nd Int. Conf. Bioinf. Biomed. Eng.*, May 2008, pp. 668–671.

[38] Y. S. Kim, M. K. Jang, C. Y. Park, H. J. Song, and J. D. Kim, "Exploring multiple biomarker combination by logistic regression for early screening of ovarian cancer," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 2, pp. 67–75, 2013.

[39] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002.

[40] A. Roy, "Estimating correlation coefficient between two variables with repeated observations using mixed effects model," *Biometrical J.*, vol. 48, no. 2, pp. 286–301, Apr. 2006.

[41] S. Zhang and J. Cao, "A close examination of double filtering with fold change and t test in microarray analysis," *BMC Bioinf.*, vol. 10, no. 1, p. 402, Dec. 2009.
[42] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU Feature Selection Repository*, pp. 1–28, 2010. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.642.5862
[43] D. Ding, D. Qi, X. Luo, J. Chen, X. Wang, and P. Du, "Convergence analysis and performance of an extended central force optimization algorithm," *Appl. Math. Comput.*, vol. 219, no. 4, pp. 2246–2259, Nov. 2012.
[44] A. F. Nematollahi, A. Rahiminejad, and B. Vahidi, "A novel physical based meta-heuristic optimization method known as lightning attachment procedure optimization," *Appl. Soft Comput.*, vol. 59, pp. 596–621, Oct. 2017.
[45] R. Kumar and A. Rajasekhar, "Speed control of PMSM by hybrid genetic artificial bee colony algorithm," in *Proc. Int. Conf. Commun. Control Comput. Technol.*, Oct. 2010, pp. 241–246.
[46] M. A. Tawhid and V. Savsani, "A novel multi-objective optimization algorithm based on artificial algae for multi-objective engineering design problems," *Int. J. Speech Technol.*, vol. 48, no. 10, pp. 3762–3781, Oct. 2018.
[47] M.-H. Lee, S. Fazli, and S.-W. Lee, "Optimal channel selection based on statistical analysis in high dimensional NIRS data," in *Proc. Int. Winter Workshop Brain-Comput. Interface (BCI)*, Jeongseon-gun, South Korea, Feb. 2013, pp. 95–97.
[48] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0172578.
[49] S. K. Prabhakar, H. Rajaguru, and S.-W. Lee, "A comprehensive analysis of alcoholic EEG signals with detrend fluctuation analysis and post classifiers," in *Proc. 7th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Seoul, South Korea, Feb. 2019, pp. 18–20.
[50] H. Byun and S.-W. Lee, "Applications of support vector machines for pattern recognition: A survey," in *Pattern Recognition With Support Vector Machines* (Lecture Notes in Computer Science), vol. 2388. Berlin, Germany: Springer-Verlag, 2002, pp. 213–236.
[51] M.-S. Lee, Y.-M. Yang, and S.-W. Lee, "Automatic video parsing using shot boundary detection and camera operation analysis," *Pattern Recognit.*, vol. 34, no. 3, pp. 711–719, Mar. 2001.
[52] S. K. Prabhakar, H. Rajaguru, and S.-W. Lee, "A framework for schizophrenia EEG signal classification with nature inspired optimization algorithms," *IEEE Access*, vol. 8, pp. 39875–39897, 2020.

SUNIL KUMAR PRABHAKAR (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Anna University, Chennai, India, in 2012, 2014, and 2017, respectively. He is currently a Postdoctoral Research Fellow with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea. His research interests include signal processing, pattern recognition, and machine learning.

SEONG-WHAN LEE (Fellow, IEEE) received the B.S. degree in computer science and statistics from Seoul National University, Seoul, in 1984, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, Seoul, South Korea, in 1986 and 1989, respectively. He is currently the Hyundai-Kia Motor Chair Professor and the Head of the Department of Artificial Intelligence, Korea University. His research interests include pattern recognition, artificial intelligence, and brain engineering. He is a Fellow of the IAPR and the Korea Academy of Science and Technology.