Genome feature optimization and coronary artery disease prediction using cuckoo search

E. Neelima[1], M. S. Prasad Babu[2]
[1]Department of Computer Science and Engineering, GITAM University, Visakhapatnam, Andhra Pradesh, India
[2]Department of Computer Science and System Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India

Article Info

Article history:

Received Jan 23, 2020
Revised Jun 11, 2020
Accepted Jun 28, 2020

Keywords:

Classification
Computer intelligence
Coronary artery disease
Cuckoo search
Gene expression coronary artery disease genes
Predictive analysis

ABSTRACT

Cardiovascular diseases are among the major health ailment issue leading to millions of deaths every year. In recent past, analyzing gene expression data, particularly using machine learning strategies to predict and classify the given unlabeled gene expression record is a generous research issue. Concerning this, a substantial requirement is feature optimization, which is since the overall genes observed in human body are closely 25000 and among them 636 are cardiovascular related genes. Hence, it complexes the process of training the machine learning models using these entire cardiovascular gene features. This manuscript uses bidirectional pooled variance strategy of ANOVA standard to select optimal features. Along the side to surpass the constraint observed in traditional classifiers, which is unstable accuracy at k-fold cross validation, this manuscript proposed a classification strategy that build upon the swarm intelligence technique called cuckoo search. The experimental study indicating that the number of optimal features those selected by proposed model is substantially low that compared to the other contemporary model that selects features using forward feature selection and classifies using support vector machine classifier (FFS&SVM). The experimental study evinced that the proposed model, which selects feature by bidirectional pooled variance estimation and classifies using proposed classification strategy that build on cuckoo search (BPVE&CS) outperformed the selected contemporary model (FFS&SVM).

Corresponding Author:

E. Neelima,
Department of Computer Science Engineering,
GITAM University, Visakhapatnam, Andhra Pradesh, India.
Email: eadha.neelima@gmail.com

1. INTRODUCTION

Among the various health aspects that lead to deaths, cardiovascular diseases (CVD) are one of the major factors that lead to millions of deaths globally every year [1]. Acute myocardial infarction (MI) is resultant element of the myocardial tissue formation because of reduced blood supply to the heart and it causes results in millions of deaths [1]. Many scientific studies have focused on solutions in terms of diagnosis, prevention, and cure for MI, but still the optimal success not accomplished in terms of mitigating the mortality ratio led resulting due to MI issues. In the present scenario, predominantly clinical symptoms are used for diagnosis of MI. Certain symptoms like complexities in breathing, inconvenience or uneasiness faced by the patients like chest pain, reports of abnormal electrocardiogram (ECG) results, abnormal fall in the circulation levels of cTns (cardiac troponins) [2]. Though there are many developments that has taken place in the domain, still there are certain limitations and constraints faced in attaining accurate analysis using the current diagnostic systems. For instance, the contemporary methods and solutions that were proposed in hs-c Tn assays has resulted in improved scope of detecting the lower circulating Tn concentrations (with increased sensitivity towards analysis). However, one of the key constraints from the process is the rise of false alarm rates as a greater number of non-diseased people are also shown as prone to conditions, because of change resulting in cTns due to the other complication (this reflects reduced sensitivity) [3]. The other diagnostic method used for detection is the cardiac mi RNAs that considered as sensitive biomarkers [4], but few limitations like the low abundance, tissue specific expression issues and the small size has impaired the reliability over the model. The role like the biomarkers has become more significant because of invention of fast, improved, and automated detection systems [5]. In some of the other studies that carried out in the lines of defining biomarkers for diagnosis, C-reactive protein (CRP), brain natriuretic peptide (BNP), and other such kind of inflammatory markers too considered, however only marginal improvements in the accuracy levels were attained as the outcome [6-8].

Domain knowledge of the pathological and physiological aspects the key aspects relied upon for developing many of the earlier cardiac biomarkers. Whereas, the microarray platforms consider the expression of large number of genes in simultaneous manner, that focuses on enabling gene expression profiling across varied pathways in simultaneous. The aforesaid method has the capability to indicate broad range of pathophysiological processes of CVD in more economic and efficient manner [9]. Gene expression profiling extends deeper than the biomarkers to identify more potential biomarkers that earlier reported to be associated with CVD. Gene expressions usually enable us to identify and discover insightful and more sensitive biomarkers that can reflect upon CVD. Majority of the studies that have focused on this section has provided significant results from the process. In [10], a study carried out for gene expression analysis to understand and discover contemporary and sensitive biomarkers of CVD identified 482 genes that are in association to composition of coronary atherosclerotic plaques and majority of them never tagged to the atherosclerosis [10]. In [11], wide scale gene expression profiling comprising 56 divergent genes for atherosclerotic and non-atherosclerotic human coronary arteries explored, of which 49 of them were associated with coronary artery disease (CAD) earlier [11]. In [12], the authors have focused on identifying a set of classifying genes based on demographics and it has strongly depicted the obstructive CAD in non-diabetic patients [12]. Divergent range of gene expressions identified that differentiated the ischemic and non-ischemic cardiomyopathy conditions of the patients confronting end-stages [13, 14]. In [15], the authors have worked on microarray analysis and gene expression profiling that are used for discovering genes related to heart failures based on expression profiles of patients with heart failure complications. In [16], the study has targeted on normal controls and MI patients have found that the genetic markets and the deregulated pathways that are associated with the disease recurrence in first time MI patients [16].

It is imperative that the efficacy with which the blood transcriptase denotes the changes of transcriptional elements in heart, improves the accuracy of diagnosis. In [17], the authors have reported that upon conducting a genome wide survey by using microarrays and the expressed sequence tags having the peripheral blood transcript me to the transcript me of nine other human tissues including the ones of heart, more than 80% of overlapping is estimated at tissue levels. 84% of overlapping with heart, indicating that study of peripheral blood transcriptase can be an economic and readily accessible tool for proxy gene expression in other tissues [17]. Though many studies have focused on the domain of differential expression in CVD outcomes, in [18] the authors have focused on using differential expression for classifying the patient record outcomes. Such an approach provides efficacy to improve the diagnosis to sub-classify patients. Also, the discriminatory features for differentiating over normal profiles and the patients with MI, CAD and the ones comprising unstable angina over gene expression in blood cells. Blood transcriptase that used with easily accessible tissue for the diagnostic purposes and majority of such contributions depict that the computational overhead resulting from dense number of gene features are adapted in the learning process. In this paper, a t-test dependent feature optimization model proposed which could effectively reduce the count of features used for analysis. The solution uses lesser number of features when compared to many of the earlier models. Despite of using limited set of features, the accuracy levels of diagnosis with reduced false alarm rates has been the outcome for the proposed solution.

2. RELATED WORKS

In [19], the study has detected varied issues of imbalances that might creep up in the usage of microarray, because of noisy, huge volume and irrelevant samples. Because of the afore-stated complexities, researchers focused to use swarm intelligence techniques for addressing the issue. The study used the technique of ant colony optimization (ACO) sampling, which developed based on ACO algorithm for eliminating the noisy and irrelevant features in the process of feature selection. Support vector machine (SVM) classifiers were adapted because of its prominence for high dimensional data classification even with small set of samples. The issues of unstable classification performance identified in cross validation process are a major factor. In [20], the authors have adapted a hybrid model for selecting optimal features by using artificial bee colony (ABC) and the classification carried out using the SVM classifiers. ABC used for clustering and selecting optimal features, which reduces the search space. Experimental studies depict that the unstable accuracy at the level of 10-fold classification. In [21], the model proposes the usage of ACO, and rough set theory (RST) in combination for achieving the optimized feature count. Accuracy of feature selection is inversely proportionate to the level of dimensionality in the feature set. In [22], it explores the usage of BAT algorithm for reducing the dimensionality of feature and selection of optimal features.

In [23], fuzzy based model depicting the rules depending on relationship among the features developed, using the combination of ACO and BAT technique. In addition, the rules can be in use for selecting optimal features in dynamic manner. Among the constraints that envisaged in the model, there is need for exposure to ensure selection of prior attributes that supports in selecting the dependent attributes, based on devised fuzzy rules. RST and BCO combined in [24] wherein, the emphasis is on clustering the features based on phenotype or the pattern that identifies the optimal features. It used the locality sensitive discriminant analysis (LSDA) for reducing dimensionality of feature sets, which further clusters, the outcome using fuzzy c-means (FCM) algorithm. FCM used in combination with ABC approach for feature similarity assessment whilst forming the clusters. The FCM incorporated with ABC approach that used for feature similarity assessment during cluster formation. Other contemporary models in the feature optimization are binary bat algorithm and ABC were used [25], and in [26] minimum redundancy and maximum relevance (M-RMR), and particle swam organization and decision tree in [27]. The M-RMR [26] is an effective method for reduction of noise and irrelevant features apart from reducing the dimensionality.

In order to surpass the constraints observed in existing meta-heuristic swarm intelligence-based feature selection models, a couple of feature selection techniques called forward feature selection, forward feature inclusion, and backward feature elimination discussed in [28]. The experimental study indicating that, among these three strategies forward feature selection is optimal. However, the performance observed in 10-fold classification done by SVM, maximum classification accuracy limited to 89% and not consistent between divergent folds. The classifiers depicted above have varied levels of performance efficacy that influenced by pre-processing stages for datasets. Pre-processing stages depicted in feature selection process could lead to better performances for classifiers. Features reduction in the datasets is one of the critical aspects facing the classifier. Many of the earlier techniques of feature selection or reduction has depicted that it could be a resourceful solution for classification purposes. In addition, the accuracy and performance of classification might depend on the quality of feature selection techniques adapted.

3. CORONARY ARTERY DISEASE PREDICTION FROM GENOME FEATURES USING CUCKOO SEARCH

In this section of study, the process of feature optimization for genome features and in terms of predicting the CAD heuristic scale-based defining based on Cuckoo search is proposed. The further sections, firstly the methods and materials used in the devised model discussed. Further, the method of feature optimization based on ANOVA standard termed as bidirectional pooled variance estimation discussed. In furtherance, the search process and label prediction based on cuckoo search discussed.

3.1. Methods and materials

3.1.1. The feature set

The 636 genomes among the total 25000 genomes are related to the CVD [29], which is usually depicted as CAD genes. In terms of evaluating the correlation among the 636 genome features, high levels of process complexity are imperative, and it causes significant range of false alarming over the prediction models. Hence, in order to ensure liner and lower levels of complexity, ensuring truncation of false alarm rates to minimal levels is very essential. Every record of the dataset adapted for training and testing phases comprise the single nucleotide polymorphism (SNP) of every gene, denoting genetic variation of various genes. In addition, the initial length of every record is 636 values depicting the SNPs of all the 636 genomes that listed in CAD genes. Initial dataset comprises the set of records that either labelled as prone to CVD or the ones that are salubrious with no trace of any CVD implications. In addition, the dimensionality of genes count has to reduce from the current number of 636 to considerably lesser values. ANOVA standard termed as bidirectional pooled variance estimation is adapted for the process of reducing the dimensionality to optimize the gene count and building the proposed scale. In addition, the details of bidirectional pooled variance estimation that is adapted for feature optimization process explored in the following section.

3.1.2. Bidirectional pooled variance estimation

Attributes of every record in the chosen dataset denotes each gene of CAD genes for the count of 636. Hence, every record comprises 636 SNPs as values pertaining to all the genomes. To defuse the number of genes that considered for optimal features, the covariance amidst values denoting every gene in the record labelled either as prone or salubrious for all the features. Genes are optimal features comprising effective covariance amidst values pertaining to prone or the salubrious records chosen. For estimating variance of SNPs, comprising values of a gene related to prone or salubrious records of the chosen training set, the method adapts ANOVA standard bidirectional pooled variance estimation. Based on results envisaged in [30, 31], the method is chosen for analysis. The bidirectional pooled variance estimation is adapted for selecting optimal features pertaining to every record (both prone and salubrious) for a training set chosen. Differential values amidst two distinct vectors depicted by the usage of bidirectional pooled variance estimation as follows:

$$bpve = \frac{(\langle v_1 \rangle - \langle v_2 \rangle)}{\sqrt{stdv(v_1) + stdv(v_2)}}$$

In the equation above

- $\langle v_1 \rangle, \langle v_2 \rangle$ indicates the mean values identified for relevant vectors $v1, v2$ and these vectors indicate the SNPs constituted as values to a gene pertaining to records labelled as prone and salubrious respectively in given training set.
- The representation s$stdv(v_1)$, $stdv(v_2)$ signify the mean square distance of the vectors $v1, v2$ respectively.

The bidirectional pooled variance estimation is the ratio amidst the mean variation of relative vectors and the square root of sum of mean square distances of the relative vectors. In furtherance, the p-value (degree of probability) [32] is attained based on t-table [33]. P-value is much lesser than the probability threshold, which reflects that the vectors vary. Hence, the feature denoting respective vectors are of optimal feature.

3.1.3. Cuckoo search

The natural elements based meta-heuristics models developed are among the best set of algorithms to address the issues of optimization. The proposed work evaluates the fitness for a given gene vector for CAD prone set and the salubrious sets based on contemporary meta-heuristic model of cuckoo search (CS) [34]. CS developed based on obligate brood parasitism of the cuckoo species. Its main characteristic is to let the eggs in the nests of other bird species that are relatively matching. Three key fundamentals based on such nesting process followed by Cuckoo are: Cuckoo egg denotes a solution to the issue and it drops randomly in a chosen nest. However, only one egg left at every instance. The nests that comprise higher quality of eggs have to pass to the future generation Nest owner shall identify a cuckoo egg based on probability ∈ [0, 1]. In the instance of such occurrence, the nest owner leaves the nest and develops other nest in a varied location. The cumulative number of nests is the fixed value. Not all the previously mentioned rules are essential, as the cuckoo search used in the proposed model, only for identifying the fitness of features for a chosen input gene record. Hence, the proposal is to develop nests in a traditional manner and the search performed using random approach. Traditional search drops only one egg in the chosen nest, but in the proposed solution, it clones the egg to varied number of compatible nests and places one egg in every compatible nest. It also estimates fitness of every egg for entire nest hierarchy.

3.1.4. The dataset

Data set generated based on records denoting coronary artery susceptibility mode (NCBI GEO Dataset ID: GDS4527) and atherosclerotic CAD prone (NCBI GEO Dataset ID: GDS3690) are gathered from NCBI gene expression omnibus (NCBI GEO) [35], authenticated as gene expression dataset repository. The dataset GDS4527 comprise gene expressions of 20 subjects. Among them 10 records are categorized as salubrious and rest of the records are categorized as prone to coronary artery disease. The other dataset GDS 3690 comprises 153 records of which 66 records categorized as salubrious and rests of them as prone to coronary artery disease. Based on the records of two datasets representing 173 subjects, values observed for CAD genes, which are collected as record for every subject. Statistics of final datasets that generated from the process depicted in the following Table 1.

Table 1. Classification of datasets

Data	Subjects
Cumulative records	173
Length of each record	636
Records with disease prone labeling	97
Records of salubrious labeling	76

3.2. Optimizing genome features

As a part of portioning process of labelled records in the dataset$G$, which classified to two sets $PandS$ indicting CAD prone and salubrious records respectively. The sets$PandS$are in the form of matrix size of records counting as row count and CAD genes counted as column count, which are fixed to 636 [29]. Every row of the matrix shall be a vector denoting SNPs attained for all the CAD genes pertaining to individual case and every column in the vector indicates SNPs gathered from specific gene in the chosen cases. Context of optimal feature selection is about a gene comprising a varied vector of SNPs pertaining to prone and salubrious record sets$PandS$. In addition, it applies bidirectional pooled variance estimation test over the attained value for a gene pertaining to both labelling sets using the following process.

```
step 1: $\overset{i=1}{\overset{|C|}{\forall}} \{pc_i \exists pc_i \in P \wedge sc_i \exists sc_i \in S\}$ Begin
step 2: $\langle pc_i \rangle = \frac{\sum_{k=1}^{|pc_i|}\{pc_i(k)\}}{|pc_i|}$// observing the mean$\langle pc_i \rangle$ of all values comprised in column vector $pc_i$ of the set
    $P$ denoting SNPs found in records of $P$ for gene $pc_i$
step 3: $\langle sc_i \rangle = \frac{\sum_{k=1}^{|sc_i|}\{sc_i(k)\}}{|sc_i|}$// observing the mean$\langle sc_i \rangle$of all values comprised in column vector $sc_i$ of the set $S$
    that denotes SNPs observed in all records of the set $S$ for gene $sc_i$
step 4: $rmsd_{pc_i \rightleftharpoons sc_i} = \sqrt{\frac{\sum_{k=1}^{|pc_i|} pc_i(k) - \langle pc_i \rangle}{|pc_i|-1} + \frac{\sum_{k=1}^{|sc_i|} sc_i(k) - \langle sc_i \rangle}{|sc_i|-1}}$ // observing the root mean square
    distance$rmsd_{pc_i \rightleftharpoons sc_i}$of the vectors $pc_i$ and $sc_i$
step 5: $t_{pc_i \rightleftharpoons sc_i} = \frac{(\langle pc_i \rangle - \langle pc_i \rangle)}{rmsd_{pc_i \rightleftharpoons sc_i}}$// Estimating the bidirectional pool variance score of the vector $pc_i$and vector $sc_i$
    comparison
step 6: $if\left(p\left(t_{pc_i \rightleftharpoons sc_i}\right) < pt\right)$ // Upon instance of degree of probability $p\left(t_{pc_i \rightleftharpoons sc_i}\right)$ identified for $t_{pc_i \rightleftharpoons sc_j}$is
    lesser than the probability threshold (usually 0.01, 0.05 or 0.1) given
step 7: $oGene \leftarrow C\{i\}$ // then the $i^{th}$ gene of the CAD genes set $C$is deliberated as optimal and moved to the
    optimal gene set $oGene$
step 8: End
```

3.3. Cuckoo search for fitness assessment

This section explores the process of fitness assessment through cuckoo search. The overall process includes nest formation, hierarchical search to notify the fitness of the optimal features of the given record towards prone to CAD and salubrious state. Nest formation, search and label prediction process explored in following sections

3.4. Nest formation

In order to perform the cuckoo search, the hierarchy of the nests should generate for corresponding disease prone and salubrious sets P, S. The optimal gene features represent the nests in a hierarchy such that each set of optimal gene features represents a unique nest in hierarchy that referred further as nest representative setn_i. The optimal gene feature sets explored such that each set contains more than one gene feature that are highly correlate in regard to the their respective SNPs as values found in records of the corresponding sets P, S.Further, these nest representative sets referred as NRS and let the hierarchies PH, SH formed respective to disease prone and salubrious setsP, S using these nest representative sets NRS Further, the unique value sets $\{e_1, e_2, .., e_n\}$as eggs, such that each egg represents the values of a gene features in nest representative set $\{n_i \exists n_i \in NRS\}$ and exists in at least one record of the respective records-set, should place in to the nest represented by $\{n_i \exists n_i \in NRS\}$.

3.5. Assessing fitness by nest search

The fitness of the given record estimates based on the number of compatible nests noticed in respective hierarchies PH, SH. Concerning this, for each nest, any egg of the respective nest is identical to the values observed in given record for the gene features in corresponding nest representative set then the fitness of the given record in related to corresponding hierarchy will increment by 1. This practice delivers the fitness related to disease prone and salubrious state for given record. Further the fitness ratio of the given record about to both hierarchies will measure, which is the average of the fitness related to number of nests in corresponding hierarchies. Then the root mean square distance of the fitness values corresponding to both hierarchies should be measure. Then these fitness ratios and root mean square distances corresponding to both hierarchies will use to confirm the state of the given record is prone to coronary vascular disease or not that explored in following section. The mathematical model to assess the fitness follows:

step 1: Let NRS be the nest representative sets (see sec 3.C) of disease prone and salubrious hierarchies PH, SH respectively, such that each nest representative set contains a set of highly correlated features obtained from optimal gene features discovered (see sec 3.B)

step 2: Let Rbe the record contains SNPs respective to all optimal feature genes selected (see sec 3.B)

step 3: Let eR be the set representing the sets of values as eggs to place in nests, such that each egg contains the values observed in R for the genes of respective nest representative set $\{n_i \exists n_i \in NRS\}$.

step 4: $pf = \sum_{i=1}^{|eR|}\{1 \exists e_i \in n_i \exists n_i \in NRS \wedge n_i \in PH\}$ //add 1 to disease prone fitness$pf$of the given record R related to prone hierarchy PH if egg $e_i$is compatible to place in nest n_i in prone hierarchyPH.

step 5: $\langle pf \rangle = \frac{pf}{|NRS|}$ //Finding the fitness ratio $\langle pf \rangle$ related to disease prone hierarchy

step 6: $rmsd_{pf} = \frac{\sum_{i=1}^{|pf|}\left\{\sqrt{(1-\langle pf \rangle)^2}\right\}+\langle pf \rangle * (|NRS|-pf)}{|NRS|}$ // The fitness ratio discards from 1 for number of nests compatible to the eggs exist in eR and the fitness ratio multiplies by the number of incompatible nests, which is the difference between total number of nests and number of compatible nests that denoted as $|NRS| - pf$

step 7: $sf = \sum_{i=1}^{|eR|}\{1 \exists e_i \in n_i \exists n_i \in NRS \wedge n_i \in SH\}$ //add 1 to salubrious fitness$sf$related to salubrious hierarchy SH if egg $e_i$is compatible to place in nest n_i in salubrious hierarchySH.

step 8: $\langle sf \rangle = \frac{sf}{|NRS|}$ //Finding the fitness ratio $\langle sf \rangle$ related to salubrious hierarchy

step 9: $rmsd_{sf} = \frac{\sum_{i=1}^{sf}\left\{\sqrt{(1-\langle sf \rangle)^2}\right\}+\langle sf \rangle * (|NRS|-sf)}{|NRS|}$// finding the root mean square distance of the salubrious

step 10: fitness using the similar process defined for prone fitness rmsd calculation in step 6

3.6. Discovering the record state

The fitness ratios $\langle pf \rangle, \langle sf \rangle$ and root mean square distances $rmsd_{pf}, rmsd_{sf}$ obtained in respective to disease prone and salubrious hierarchies$PH, SH$for given input record Rshould use to label the record $R$is prone to disease or salubrious. The label should define using the conditional flow that follows:

```
step 1: $if(\langle pf \rangle \cong \langle sf \rangle)$ Begin
step 2: $if(rmsd_{pf} < rmsd_{sf})$ Begin
step 3: Label the record as disease prone
step 4: End //of step 2
step 5: Else $if(rmsd_{pf} > rmsd_{sf})$Begin
step 6: Label the record as salubrious
step 7: End // of step 5
step 8: Else //of condition in step 5
step 9: Record state is ambiguous// since the fitness ratios and root mean square distance obtained for both
        hierarchies is same
step 10: End //of step 1
step 11: Else Begin // of condition in step 1
step 12: $if(\langle pf \rangle > \langle sf \rangle)$Begin
step 13: Label the record as disease prone
step 14: End //of step 11
step 15: Else $if(\langle pf \rangle < \langle sf \rangle)$Begin
step 16: Label the record as salubrious
step 17: End //of step 14
step 18: Else Begin//of condition in step 15
step 19: Record state is ambiguous// since the fitness ratios and root mean square distance obtained for both
         hierarchies are not meeting the prescribed conditions
step 20: End // of step 18
step 21: End //step 11
```

3.7. Empirical analysis of the proposed model

The experimental study conducted on dataset explored in section 3.4). In order to explore the performance significance of the proposed model that incorporated the feature optimization by bidirectional pooled variance and cuckoo search model classifier (BPVE&CS), the experimental results obtained and compared to the other contemporary model [28] that selects optimal features using forward selection technique and classifies using SVM classifier (FFS&SVM).

The statistics of the dataset used can depict in table 1 that explored in 3.4). The classification process on said dataset using both models done in 4 folds. In addition, the performance assessment of the proposed model and contemporary model depicted using classification assessment metrics [36] such as precision, sensitivity, specificity, and accuracy. The results obtained for both the proposed and contemporary model depicted in Table 2. The notation used as row and column headers are:

– PPV: Positive predictive value (or) precision
– TPR: True positive rate (or) sensitivity
– TNR: True negative rate (or) specificity
– FNR: False negative rate (or) missing rate
– FPR: False positive rate (or) fallout
– ACC: Accuracy

Table 2. The metrics and the values obtained from 4-fold classification using proposed and contemporary model

		PPV	TPR	TNR	FNR	FPR	ACC
BPVE&CS	Fold#1	0.958	0.958	0.947	0.042	0.053	0.953
	Fold#2	0.923	1	0.895	0	0.105	0.953
	Fold#3	1	0.917	1	0.083	0	0.953
	Fold#4	0.957	0.917	0.947	0.083	0.053	0.93
FFS&SVM	Fold#1	0.88	0.917	0.842	0.083	0.158	0.884
	Fold#2	0.875	0.875	0.842	0.125	0.158	0.86
	Fold#3	0.917	0.917	0.895	0.083	0.105	0.907
	Fold#4	0.84	0.875	0.789	0.125	0.211	0.837

The prediction accuracy of the both the models observed from the experiments depicted in Figure 1. The results depicted in Figure 1 evincing that the classification accuracy observed for BPVE&CS is stable and substantially high with greater than 93% that compared to FFS&SVM, which observed as inconsistent and less than 90%. Figures 2 and 3 depicts the performance advantage of the BPVE&CS over FFS&SVM towards sensitivity and specificity those refers the significance of disease scope prediction and significance of salubrious state prediction respectively. The proposed model clearly outperformed the FFS&SVM in this regard. The Figures 4 and 5 evincing the false negative rate or missing rate, false positive rate or fall-out those indicates prediction failure rate of disease scope and salubrious state respectively observed for BPVE&CS and FFS&SVM. From the depicted results of prediction failure rate for disease scope and salubrious state is much low for proposed model that compared to FFS&SVM.

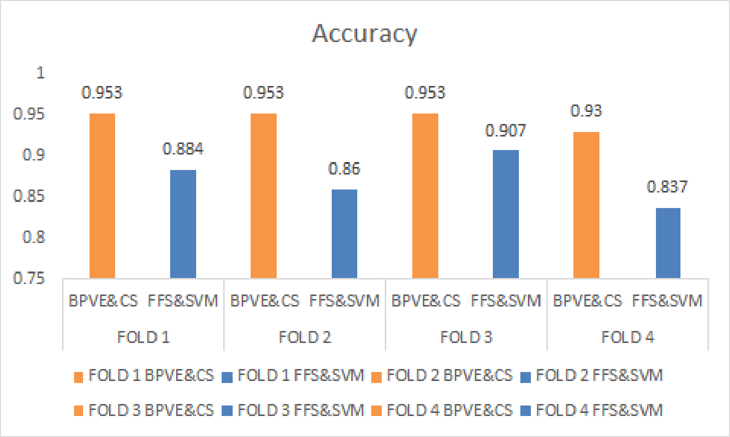

Figure 1. The classification accuracy ratios of BPVE&CS and FFS&SVM observed from 4-fold classification

The process complexity observed in both training and testing phases depicted in Figures 6 and 7 respectively. From the depicted figures, it is obvious to conclude that process completion time of BPVE&CS in training and testing phases is significant, since they are much lesser than the process completion time observed for FFS&SVM in respective training and testing phases.

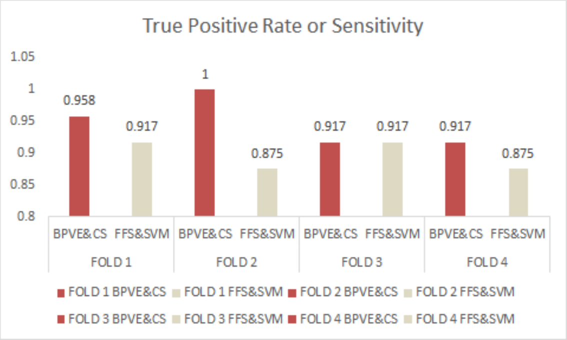

Figure 2. The sensitivity (disease prediction rate) of BPVE&CS and FFS&SVM observed from 4-fold classification

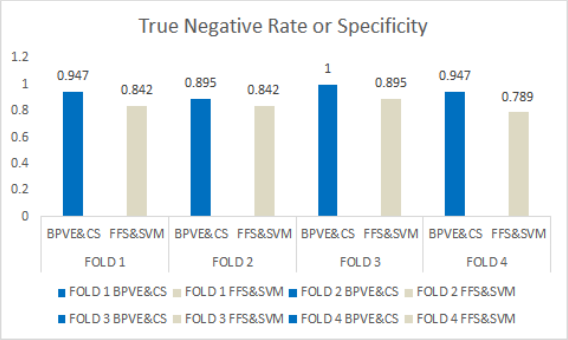

Figure 3. The specificity (salubrious state prediction rate) of BPVE&CS and FFS&SVM observed from 4-fold classification

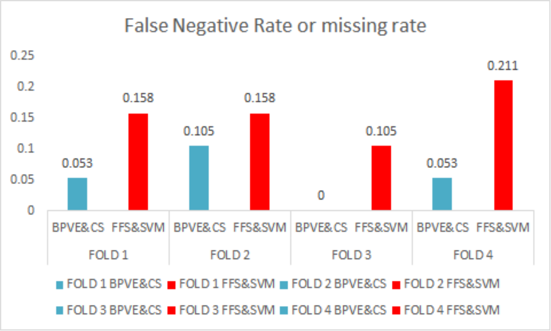

Figure 4. The disease prediction failure rate (false negative rate) of BPVE&CS and FFS&SVM observed from 4-fold classification

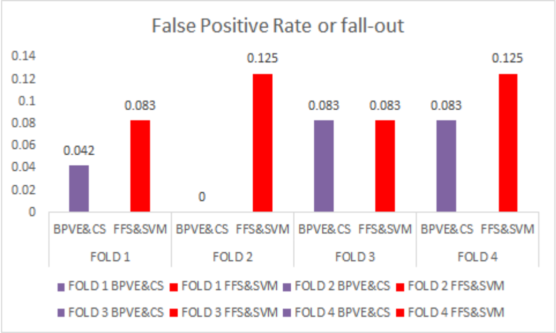

Figure 5. The salubrious state prediction failure rate (false positive rate) of BPVE&CS and FFS&SVM observed from 4-fold classification

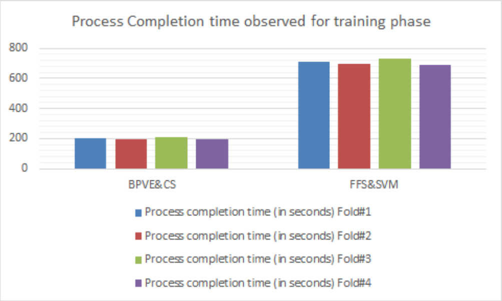

Figure 6. Process completion time of training phase observed for both BPVE&CS and FFS&SVM in 4-fold classification

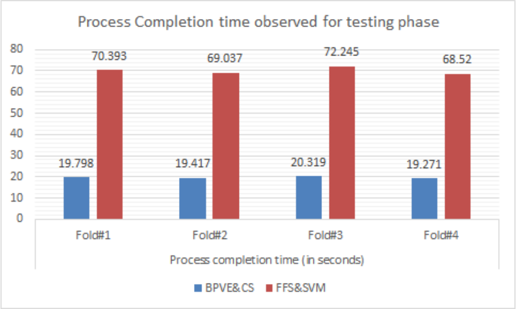

Figure 7. Process completion time of testing phase observed for both BPVE&CS and FFS&SVM in 4-fold classification

4. CONCLUSION

Gene expressions forms as the combination of many genes among the thousands of genes defined until now. Among these thousands of genes, 636 genes identified as cardiovascular related that are usually refers as CAD genes. Still this count of genes is high dimension to apply machine-learning methods to learn cardiovascular related information. Concerning this, reducing the dimensionality of the CAD genes is essential factor to improve the performance of the machine learning process that applied on these CAD genes set. This manuscript depicted a novel optimal feature selection technique that uses bidirectional pooled variance estimation (BPVE) for CAD prediction. Learnings from the contemporary literature stating that existing classifiers are unstable towards classification accuracy and inconstant to label the individual record, hence the label prediction for given record of the individual is highly false alarmed. Considering this, a novel classifier as prediction scale proposed here in this article. The depicted classifier built over the swarm intelligence technique called cuckoo search. The experimental study stating that the proposed model BPVE&CS is the best to reduce dimensionality of the CAD genes among the models found in recent literature. The experimental study compared the results obtained from proposed model with the results obtained from contemporary model that selects features through forward feature selection and classifies using SVM (FFS&SVM). The proposed model evinced 18 genes as optimal features, which is best count that compared to any of the contemporary model found in recent literature. The label prediction strategy through the proposed classifier that build over cuckoo search is consistent in classification accuracy, evinced less fallout and missing rate and high sensitivity and specificity. The process completion time of the proposed model also found as much less and linear that compared to the FFS&SVM. The future research can extend this work to discover the possibilities of using other ANOVA standards like Wilcoxon Signed rank, Entropy test to reduce the dimensionality of the feature set.

REFERENCES

[1] T. Thom, *et al*., "Heart disease and stroke statistics--2006 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee" *Circulation*, vol. 113, no. 6, pp. e85, 2006.

[2] K. Thygesen *et al*., "Universal definition of myocardial infarction: Kristian Thygesen, Joseph S. Alpert and Harvey D. White on behalf of the Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction," *European Heart Journal*. vol. 28, no. 20, pp. 2525-2538, 2007.

[3] K. M. Eggers, L. Lind, P. Venge, and B. Lindahl, "Will the universal definition of myocardial infarction criteria result in an overdiagnosis of myocardial infarction," The *American Journal of Cardiology*, vol. 103, no. 5, pp. 588-591, 2009.

[4] Z. Wang, X. Luo, Y. Lu, and B. Yang, "miRNAs at the heart of the matter," *Journal of Molecular Medicine*, vol. 86, no. 7, pp. 771-783, 2008.

[5] M. P. Saguer and M. C. Rodicio, "Detection methods for microRNAs in clinic practice," *Clinical Biochemistry*, vol. 46, no. 10, pp. 869-878, 2013.

[6] O. Melander *et al*., "Novel and conventional biomarkers for prediction of incident cardiovascular events in the community," *JAMA*, vol. 302, no. 1, pp. 49-57, 2009.

[7] T. Shah *et al*., "Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts," *International Journal of Epidemiology*, vol. 38, no. 1, pp. 217-231, 2009.

[8] P.W.F. Wilson, M. Pencina, P. Jacques, J. Selhub, R. D'AgostinoSr, and C. J. O'Donnell, "C-reactive protein and Reclassification of Cardiovascular Risk in the Framingham Heart Study Clinical Perspective," *Circulation: Cardiovascular Quality and Outcomes*, vol. 1, no. 2, pp. 92-97, 2008.

[9] D. M. Pedrotty, M. P. Morley, and T. P. Cappola, "Transcriptomic biomarkers of cardiovascular disease," *Progress in Cardiovascular Diseases*, vol. 55, no. 1, pp. 64-69, 2012.

[10] A. M. Randi *et al*., "Identification of differentially expressed genes in coronary atherosclerotic plaques from patients with stable or unstable angina by cDNA array analysis," *Journal of Thrombosis and Haemostasis*, vol. 1, no. 4, pp. 829-835, 2003.

[11] S. R. Archacki, "Identification of new genes differentially expressed in coronary artery disease by expression profiling," *Physiological genomics*, vol. 15, no. 1, pp. 65-74, 2003.

[12] M. R. Elashoff, "Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients," *BMC Medical Genomics*, vol. 4, no. 1, p. 26, 2011.

[13] M.M. Kittleson, "Identification of a gene expression profile that differentiates between ischemic and no ischemic cardiomyopathy," *Circulation*, vol. 110, no. 22, pp. 3444-3451, 2004.

[14] M. M. Kittleson, "Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure," *Physiological Genomics*, vol. 21, no. 3, pp. 299-307, 2005.

[15] K. D. Min *et al*., "Identification of genes related to heart failure using global gene expression profiling of human failing myocardium," *Bioch and Biophy Res Comm*, vol. 393, no. 1, pp. 55-60, 2010.

[16] R. Suresh, "Transcript me from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction," *Journal of Molecular and Cellular Cardiology*, vol. 74, pp. 13-21, 2014.

[17] C.C. Liew, J. Ma, H.C. Tang, R. Zheng, and A.A. Dempsey, "The peripheral blood transcript me dynamically reflects system wide biology: a potential diagnostic tool," *Journal of Laboratory and Clinical Medicine*, vol. 147, no. 3, pp. 126-132, 2006.

[18] N. Kazmi and T.R. Gaunt, "Diagnosis of coronary heart diseases using gene expression profiling; stable coronary artery disease, cardiac ischemia with and without myocardial necrosis," *PloS one*, vol. 11, no. 3, p. e0149475, 2016.

[19] H. Yu, J. Ni, and J. Zhao, "ACOSampling: An ant colony optimization-based under sampling method for classifying imbalanced DNA microarray data," *Neurocomputing,* vol. 101, no. 2, pp. 309–318, 2013.

[20] M. S. Uzer, N. Yilmaz, and O. Inan, "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification," *The Scientific World Journal*, vol. 2013, no. 11, p. 419187, 2013.

[21] H. Arafat, R. M. Elawady, S. Barakat, and N. M. Elrashidy, "Using rough set and ant colony optimization in feature selection," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 2, no. 1, 2013.

[22] A. M. Taha and A. Y. Tang, "Bat algorithm for rough set attribute reduction," *Journal of Theoretical and Applied Information Technology,* vol. 51, no. 1, pp. 1-8, 2013.

[23] P. G. Kumar, S. A. Vijay, and D. Devaraj, "A hybrid colony fuzzy system for analyzing diabetes microarray data," 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 104-111, 2013.

[24] K. Sathishkumar, V. Thiagarasu, and M. Ramalingam, "An efficient artificial bee colony and fuzzy c means based clustering gene expression data," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 1, no. 5, 2013.

[25] R. Y. Nakamura, L. A. Pereira, D. Rodrigues, K. A. Costa, J. P. Papa, and X. S. Yang, "Swarm Intelligence and Bio-Inspired Computation: 9. Binary Bat Algorithm for Feature Selection," Elsevier Inc. Chapters. 2013.

[26] H. Alshamlan, G. Badr, and Y. Alohali, "MRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed research international*, vol. 2015, no. 9, pp. 1-15, 2015.

[27] C. H. Chen *et al.*, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC bioinformatics,* vol. 15, no. 1, p. 49, 2014.

[28] S. Shilaskar S and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4146-4153, 2013.

[29] H. Liu *et al.*, "CADgene: a comprehensive database for coronary artery disease genes," *Nucleic Acids Research*, vol. 39, no. 1, p. D991-6, 2011.

[30] H. Budak, S. E. Taşabat, "A Modified T-Score for Feature Selection," *Applied Sciences and Engineering,* vol. 17, pp. 845-845, 2016.

[31] O. Kummer, J. Savoy, R. E. Argand, "Feature selection in sentiment analysis," *Proceedings of the 9th Conference on Information Retrieval and Applications,* PP. 273-284, 2012.

[32] P. R. Sahoo and T. M. Theorems, "Functional Equations," World Scientific, 1998.

[33] T-table, [Online]. Available: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf, 2017.

[34] X. S. Yang and S. Deb, "Cuckoo search via Lévy flights," In *Nature & Biologically Inspired Computing*, NaBIC 2009, World Congress on IEEE, pp. 210-214, 2009.

[35] T. Barrett et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic acids research*, vol. 41, no. D1, p. D991-5, 2013.

[36] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, unforcedness, nakedness and correlation," *Journal of Machine Learning Technologies,* vol. 2, no. 1, pp. 37-63, 2011.