Artificial Intelligence Evolution
https://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER
PUBLISHER

Research Article

KMeans-NM-SalpEpi: Genetic Interactions Detection through K-Means Clustering with Nelder-Mead and Salp Optimization Techniques in Genome-Wide Association Studies

S. Priya[*], **R. Manavalan**

Department of Computer Science, Arignar Anna Government Arts College, Villupuram, Affiliated to Thiruvalluvar University, Vellore, Tamilnadu, India
Email: priyasri.ash@gmail.com

Received: 13 August 2021; **Revised:** 14 September 2021; **Accepted:** 26 September 2021

Abstract: Complex diseases identification through Gene-Gene Interactions (GGIs) plays a significant challenge in Genome-Wide Association Studies (GWAS). A typical indicator of genetic variations in many human diseases is Single Nucleotide Polymorphisms (SNPs). SNPs are the most prevalent sort of genetic variation seen in human beings. The interactions between various SNPs are called Epistasis or genetic interactions. This research paper proposes a two-stage epistasis detection approach based on K-Means clustering and optimization techniques to detect epistasis effects responsible for complex human diseases. In the screening stage, K-Means clustering is adapted to partition the genotype dataset into various clusters. Traditional K-Means clustering algorithms have the flaw of arbitrary selection of the initial k centroid, which leads to inconsistent solutions and traps in the local optimum. We present a hybridized technique based on the K-Means algorithm and Nelder-Mead (NM) optimization (KMeans-NM) to avoid local optima, and all the genotype data falls into a unique collection of clusters for different runs. In the search stage, Salp Optimization with single objective functions (Salp-SO) and Salp Optimization with multi-objective functions (Salp-MO) are employed over the clusters obtained from the screening stage to find disease correlated SNP combinations. The performance of the various proposed algorithms is tested over the simulated datasets. Experimental findings indicated that the KMeans-NM-SalpEpi-SO and KMeans-NM-SalpEpi-MO method is superior to other techniques.

***Keywords*:** epistasis, genetic interactions, cluster, Nelder-Mead optimization, Single Nucleotide Polymorphism (SNP)

1. Introduction

Finding and investigating the association between genetic markers and related human genetic diseases is one of the current advanced diagnostic measures by the physician. The general inquiry of genetic patterns and variations within the human genome is significant in GWAS. Both genetic and environmental risk factors increase pathogenicity. Genome-Wide Association Studies (GWAS) researchers tend to find genotype variations of several diseases such as hypertension, rheumatoid arthritis, cancer, chronic illness, cardiovascular disease, diabetes, psoriasis, etc. [1]. GWAS screened for a massive volume of SNPs and biomarkers of phenotypes linked to human disease cases and controls [2].

GWAS incorporates extensive data collection to trace phenotypes and genetic markers as an indicator of various

DOI: https://doi.org/10.37256/aie.2220211099

diseases. Single Nucleotide Polymorphisms (SNPs) are a form of genetic variation marker that plays a critical role in developing various complex disease traits [3]. SNPs are variations in the DNA sequence that are based on the bond between nitrogenous bases as follows Cytosine (C), Thymine (T), Adenine (A), and Guanine (G), and also changes in the amino acid sequence [4]. Each SNP is linked to characteristics that can identify the genetic predisposition to the related diseases by examining the gene regulatory pathways [5].

In past decades, numerous approaches have been established for identifying GGIs. Currently, stochastic search, exhaustive search, statistical-based techniques, and optimization-based strategies could be used to detect epistasis. Statistical methods are typically used in epidemiological research to recognize genetic associations that can be categorized as parametric or nonparametric [6]. An exhaustive search will return all possible combinations, but the computational cost will be prohibitively high. It calculates the score for each SNP interaction, and a user-specified threshold is used to detect disease correlated interactions. The epistasis-based algorithms such as Multifactor Dimensionality Reduction (MDR), Boolean Operation-Based Screening and Testing (BOOST), Generalized Multifactor Dimensionality Reduction (GMDR), Efficient Survival-Multifactor Dimensionality Reduction (ES-MDR), PLINK, Generalized Multifactor Dimensionality Reduction-Graphics Processing Unit (GMDR-GPU) are evaluated based on exhaustive analysis [7]. Yang et al. [8] developed the method Fuzzy Set-based Multiobjective Multifactor Dimensionality Reduction (FSMOMDR) in 2020 to identify genetic interactions in Coronary Artery Disease (CAD).

Random sampling is used to discover statistical correlations between diverse effects of epistasis and disease through stochastic search strategies. The stochastic search takes much less time to complete the prediction task than the exhaustive search since it is influenced by the random seed [9]. BEAM, SNPRuler algorithm uses random sampling techniques to evaluate SNP combinations.

Exhaustive and stochastic algorithms require high computational costs, and it is only a preference for certain disease models. In recent times evolutionary algorithms for epistasis detection have been of great concern to minimize computational costs, as they can solve NP-hard issues in polynomial times efficiently [10]. The evolutionary strategies minimize the search time complexity and use the scoring functions to determine the best SNP combinations. A multi-objective ant colony optimization technique (MACOED) was introduced to detect genetic interactions [11]. Epistasis based on Ant Colony Optimization Algorithm (epiACO) was presented to recognize SNP interactions. The different strategies for path selection and the memory-based approach are adapted to improve epiACO [12]. An Epistatic Interaction Multi-Objective Artificial Bee Colony Algorithm Based on Decomposition (EIMOABC/D) model was suggested for epistasis interaction detection [13]. The multi-objective bat optimization algorithm called epiBat is also presented for epistasis identification [14]. In 2019, Liyan Sun et al. [9] developed a multi-revolutionary method called SEE to recognize epistasis effects. The SEE algorithm consists of Sort, Exploitation, and Exploration procedures to detect GGIs. The multi-objective chaotic atom search optimization is proposed to detect 2-locus associations [15]. The primary problem in available epistasis detection algorithms is always incurring a huge computational cost and minimal detection power. While comparing to the presently available methods, the proposed method aims to discover disease-correlated SNPs with high detection capacity and also detect multi-locus interactions.

A novel epistasis detection strategy with a two-stage hybridization method is K-Means Cluster with Nelder-Mead (NM) and Salp Swarm Algorithm (SSA) (KMeans-NM-SalpEpi) is introduced to identify multi-locus SNP interactions. The Nelder-Mead optimization technique is adapted to find the initial centroids for K-Means clusters to partition the genotype data into different clusters at the screening phase. In the clean phase, two variations of the SSA, i.e. Single Objective SSA (SOSSA) and Multi-Objective SSA (MOSSA), are employed over the clusters that emerged from the screening phase to detect the disease-relevant SNP combinations. The main objective of this work is to establish an efficient multi-locus epistasis model to accelerate the identification of disease-related SNP-SNP interactions from hundreds of SNPs. The performance of the proposed approaches is measured over the 2-locus and 3-locus Disease models with Marginal Effects (DMEs) and Disease models with No Marginal Effects (DNMEs) and also compared over MACOED [11].

The structure of the research work is arranged as follows. Section II discusses the material and methods used for epistasis detection. Section III outlines the detailed description of the proposed algorithm. Section IV explores experimental results and discussion. Finally, Section V concludes this article with future scope.

2. Materials and methods

In this section, we formally introduce the components of the proposed approach, such as the K-Means cluster, Nelder-Mead optimization, and Salp optimization strategy for genetic interactions detection. K-Means clustering technique with Nelder-Mead optimization is adapted in the screening stage to group SNPs into three clusters. The Salp Search Algorithm (SSA) is applied in the search stage to find high-order SNP combinations. A detailed description of each component of the proposed approach is hereunder.

2.1 *K-Means clustering technique*

K-Means clustering is one of the most commonly used cluster analysis techniques. The main aim of this algorithm is to divide n number of unlabeled observations into k number of clusters. This algorithm assigns each set of data points to any cluster based on similarity. The degree of similarity between two objects is determined by calculating their distance using the Euclidean distance metric. This iterative approach starts with initial estimates of k number of centroids, which can be taken from the dataset arbitrarily [16]. The steps behind this technique are the assignment of data points to clusters and updating the centroids. The first step involves allocating each data point to its corresponding centroid, which is calculated based on distance measure as shown below.

$$P = \sum_{j=1}^{k}\sum_{i=1}^{m} \| x_i - C_j^2 \|$$

Where $\| x_i - C_j^2 \|$ represents the distance between a data point and cluster center, C_j denotes centroid and x_i denotes data point.

In the second step, the centroid updating is made by computing the mean value of all the data points of a specific cluster. This procedure is repeated until the maximum number of iterations is reached, or the centroids are not changed for the subsequent iterations.

2.2 *Salp Swarm Algorithm (SSA)*

Salp Swarm Algorithm (SSA) is a population-based optimization technique [17]. SSA imitates salps' social actions as they are collectively together in a chain during their sailing and foraging for food in the sea. Two kinds of agents are present in SSA: the leader exits in the chain's head and the other salps are designated as followers. The leader is in charge of guiding the path movement of the population, the supporters follow the leader one by one.

The salp population size is N, which denotes the number of SNPs, and its location is defined in the D dimensional search space. The salps positions are interpreted in a two-dimensional coordinate system of N rows and D columns.

$$X_{N\times D} = lb + rand(N,D)\times(ub-lb) \tag{1}$$

The best global search solution is described as F, which is responsible for the foraging target of the swarm. The leader's position is generated by Equation (1) as follows:

$$\begin{aligned} x_k^1 &= F_k + c_1\left(\left(ub_k - lb_k\right)c_2 + lb_k\right),\ c_3 \geq 0.5 \\ &= F_k - c_1\left(\left(ub_k - lb_k\right)c_2 + lb_k\right),\ c_3 < 0.5 \end{aligned} \tag{2}$$

Where x_k^1 represents the position of the salps in the kth dimension
F_k indicates the location of the food in the kth dimension
ub_k represents the upper limit of the kth dimension
lb_k represents the upper bound of the kth dimension

c_1, c_2 and c_3 indicate random numbers.

The convergence factor c_1 accounts for exploration and exploitation, which is described as follows:

$$c_1 = 2e^{-(4t/T)^2} \tag{3}$$

where t denotes the present iteration count and the maximum iterations are represented by T. c_2 and c_3 are randomly generated numbers within the interval [0, 1].

The follower's position is updated as shown in Equation (4).

$$x_d^n = \frac{1}{2}\left(x_d^n + x_d^{n-1}\right) \tag{4}$$

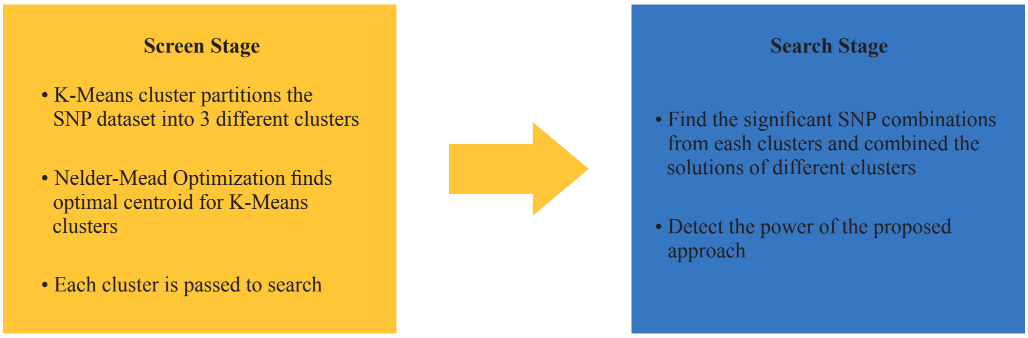
All salps did not determine the location of the target (feed) during the actual iteration. During the iterative process, the fitness values for all the salps are computed. Then, the salps with the best scoring value are updated as the current food position.

2.3 *Nelder-Mead optimization*

The Nelder-Mead method is a multidimensional unconstrained minimization search method formulated based on a simplex algorithm. The key concept behind the NM method is to find the worst and best vertices of the simplex then replacing the worst point with another point, which has a better cost value. As a result, the simplex proceeds from the worst to the best point [18] and measured natural frequencies are obtained by using cracked beam frequency response and modal analysis. A hybrid Particle Swarm-Nelder-Mead (PS-NM). Initially, Nelder-Mead begins with a simplex that is generated at random. It then continues to transform this simplex one vertex at a time towards an optimal region in the search space during each iteration. It iterates through each vertex in the search space and transforms the simplex one vertex at a time towards an optimal region. The Nelder-Mead method comprises five steps such as sorting, reflection, expansion, contraction, and shrinkage. The detailed descriptions of these steps are exposed in [19] an optimal gain tuning method for PID controllers is proposed using a novel combination of a simplified Ant Colony Optimization algorithm and Nelder-Mead method (ACO-NM).

3. Hybridization of K-Means cluster with Nelder-Mead optimization and Salp optimization for Epistasis detection (KMeans-NM-SalpEpi)

The KMeans-NM-SalpEpi consists of two stages including the screen and clean stage. The objective of the screen and the clean stage is exposed in Figure 1.



Screen Stage

- K-Means cluster partitions the SNP dataset into 3 different clusters
- Nelder-Mead Optimization finds optimal centroid for K-Means clusters
- Each cluster is passed to search

Search Stage

- Find the significant SNP combinations from eash clusters and combined the solutions of different clusters
- Detect the power of the proposed approach

Figure 1. Stages of KMeans-NM-SalpEpi approach

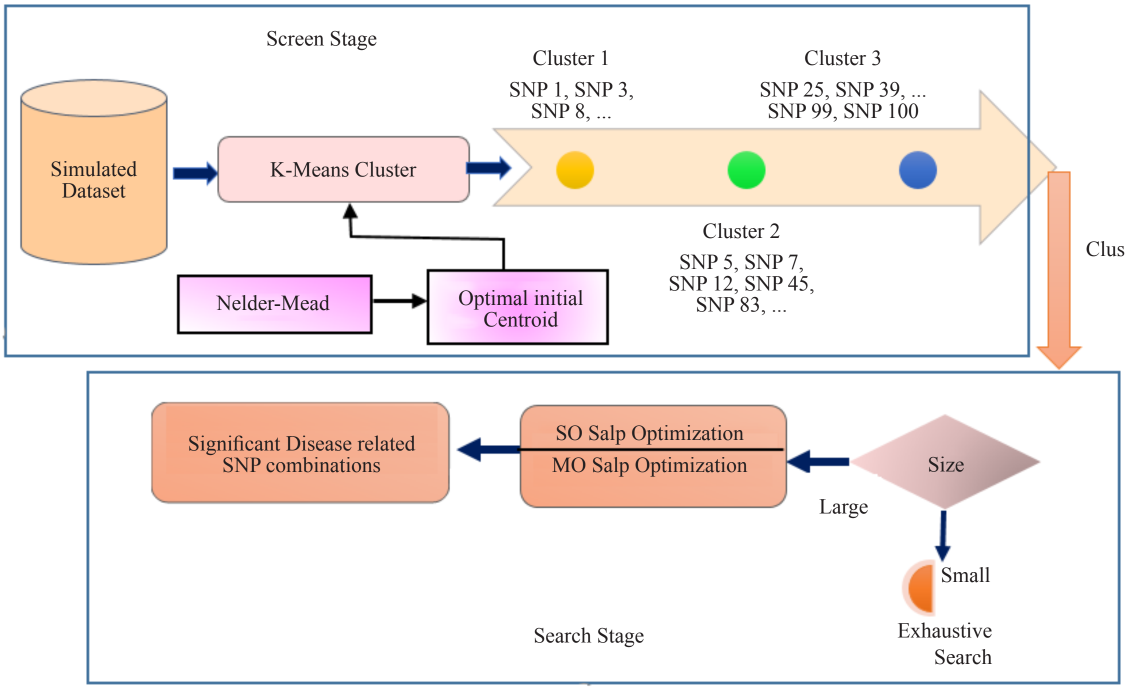


Figure 2. General Architecture of KMeans-NM-SalpEpi

3.1 *Screen stage*

The SNP genotype dataset holds M samples with M1 cases and M0 controls and N SNPs. The phenotype of individuals is denoted as v, v = 1 denotes cases and v = 0 denotes controls; We use S_i (i = 1, 2, 3, ..., N) to denote the ith SNP. This work uses a case-control design and assumes all SNPs are biallelic. We'll call A as major allele and a denotes minor allele. Each SNP has three genotypes: homozygous major (AA), heterozygous (Aa), and homozygous minor (aa). They are generally coded as 0, 1, and 2. The dataset consists of 100 SNPs (N) and 1600 samples (M). Each SNP is considered as a feature set. In the screening stage, the K-Means clustering technique partitions the featured SNPs subset into three clusters. We used cluster size (k) as three since partition the SNPs into more number clusters (above 3 clusters) cause the disease correlated SNPs to fail to fall into the same cluster set. In 2-locus association, SNPs 99 and 100 are considered as disease-causing pairs, similarly, SNPs 98, 99, and 100 are considered as disease-related SNP combinations in a 3-locus dataset.

The K-Means cluster algorithm randomly selects any three SNPs as initial centroid, which produces inconsistent partition of SNPs into different clusters and traps in local minima. These issues are overcome by hybridization of K-Means clustering and Nelder-Mead optimization to find optimal initial centroid to produce unique clusters for different runs [20]. The NM method selects three featured SNPs subset as an optimal centroid. Based on these centroids, all the SNPs are categorized into three different clusters. The computational complexity of searching genetic interactions is significantly reduced while dividing the SNPs into different clusters. For a dataset with 1,000 SNPs, 499,500 2-locus SNP combinations are necessary for analysis. But, if we divide these SNPs into 10 groups, the number of 2-locus SNP combinations required for analysis is 49,500 only. These clusters are passed to the search stage to detect disease correlated SNP combinations. The pseudo-code for the screen stage is presented in Figure 3.

```
Stage 1-Hybridization of K-Means Cluster with Nelder Mead optimization for clustering SNPs to detect epistasis effects

Input
Data: Simulated genotype dataset, k: number of clusters
Output
Three clusters consists of various SNPs

Screen Stage
Step 1: Initialize clusters centroids using Nelder Mead (NM) optimization technique, the NM takes the simulated dataset as input
    and produces optimal centroids as output. Each SNP ($S_i$) is considered as a point (feature) in a feature subset. The NM chose three
    featured SNPs subsets as an optimal centroid.
Step 2: For Each SNP $S_i$ (i = 1, 2, 3, …, 100) in a feature subset and the centroid of cluster $C_m$ (m = 1, 2, 3), calculate the
    Euclidean distance between $S_i$ and $C_m$. Here, centroid $C_m$ is a subset of $S_i$ (i = 1, 2, …, 100). Each SNP is assigned to any one of
    the three cluster group based on the closest center. This means every $S_i$ (i = 1, 2, 3, …, 100) is divided into m ($1 \le m \le k$)
    groups based on Euclidean distance.
Step 3: Update centroids: In each iteration, update each centroid after each SNP has been placed into one of the k clusters.
    For each cluster group, update the clustering centroid.
Step 4: Steps (2) and (3) are repeated until the centroids of k clusters no longer change or the maximum number of iterations
    is reached.
```

Figure 3. Pseudo-code for KMeans-NM-SalpEpi in Screen stage

3.2 *Search stage*

In the search stage, two search techniques are adapted. An exhaustive search technique is applied in the cluster subset for a small cluster size (Less than 10). For clusters with more number of SNPs, the salp optimization technique is applied. Consider S = {S_1, S_2, …, S_N, N <= No. of SNPs} is a set with N SNPs. ε(S, H) is a score function to examine the association between S and phenotype H. The k-way SNP combination S is said to be a strong association with H if ε(S, H) > α (α is threshold value).

The mathematical model for optimization to detect a k-way disease-causing combination model can be denoted as max

$$\underbrace{maxf}_{X}(X,Y), X = (X_{s_1}, X_{s_2}, X_{s_3}......X_{s_k})$$

where S_i (i = 1, 2, …, k) is the index SNP locus X_{s_i} and f(X, Y) denotes the objective function for evaluating the association between genotype X and phenotype Y. Each agent is randomly assigned with a combination of SNPs. Two variations of the Salp optimization technique, such as Single Objective (SO) Salp optimization and Multi-Objective (MO) salp optimization, are suggested to find the significant disease correlated SNP combinations. The fitness function for SalpEpi-SO is G-test. SalpEpi-MO utilizes K2 score and AIC score as fitness functions, then Pareto optimal front approach is used to select non-dominated SNPs from these two fitness functions. These non-dominated SNPs are passed into G-test to find significant disease correlated SNPs for 2-locus and 3-locus models. Finally, the performance of KMeans-NM-SalpEpi-SO and KMeans-NM-SalpEpi-MO is analyzed and compared with KMeans-Salp-Epi-SO and KMeans-Salp-Epi-MO and also contrast without applying clustering techniques such as SalpEpi-SO and SalpEpi-MO.

The pseudo-code of KMeans-NM-SalpEpi for the search stage is presented in Figure 4.

```
Stage 2-Salp Optimization for finding Significant SNP Combinations

Input
Data: Genotyped Simulated dataset divided into clusters
N: number of Salp, m: interaction order
max_iter: Maximum iterations
Output
Optimal SNP combination

Search Stage-SalpEpi-SO
Step 1: Initialize the salp locations.
Step 2: Each salp is initialized on the random position of 2 columns (for 2-locus combinations) or 3 columns (for three locus associations), based on the combinations of SNPs in the clusters
Step 3: while t < max_iter do
        Select a Salp (SNP combinations) and evaluate scoring functions based on the G-test statistic (SO) or
        K2 score and AIC score for MO optimization.
        Choose a source of food from repository: F = Select Food (repository)
        For each salp (xi)
          if (i = 1)
              Update the position of the leader salp
          else
              Update the position of the follower salp
          end
        end
          The Salp evaluates new combinations of SNPs and compares them with the previously stored solution space and updates the current solution space.
        End for
        End while
Search Stage-SalpEpi-MO
Follow the steps Step 1 to Step 3 in SalpEpi-SO
Step 4: Pareto optimal front return the Non-dominated SNPs
Step 5: For $i$ = 1 to size (non-dominated SNPs)
          For $j = i + 1$ to size (non-dominated SNPs)
          $Epistasic\_pair$ = G-test ($x_i$, $x_j$)
End For
```

Figure 4. Pseudo-code for KMeans-NM-SalpEpi in Search stage

4. Experimental result and discussion

Two simulation models, such as the Disease loci without Marginal Effects (DNME), and Marginal Effect Disease (DME) models, are considered to evaluate the robustness of the proposed methods. Sections 4.1 and 4.2 provide a detailed description of these simulations models and evaluation metrics, respectively. The proposed epistasis models are implemented using MATLAB R2018(b) software in a single CPU system with Intel(R) Core(TM) i5-7200U processor @ 2.50GHz speed. Section 4.3 exposes the experimental outcome of proposed epistasis detection techniques.

4.1 *Simulated datasets*

The efficacy of proposed algorithms is measured over simulated datasets of various disease models. A disease model is characterized as the likelihood of being affected by the disease given a mixture of SNPs [21]. Two distinct types of epistatic models, such as Disease loci with Marginal effect (DME) models and Disease loci without Marginal Effects (DNME) models are generated for two-locus and multi-locus disease analysis. DME model characterizes the interactive and marginal effects of the disease. Three gene disease models such as additive, multiplicative, and threshold models are chosen for three-locus and two-locus analysis [22]. DNME model reveals only interactive effects without marginal effects. The data sets for the research are generated using Gametes [21] software. The description of DME and DNME models chosen for experimental analysis is exposed in Table 1.

Table 1. Simulated dataset details

Dataset name	Disease Model	No. of Models	SNP Details	Description
3-locus dataset	DME Models-Additive, Multiplicative, Threshold Models	5 Models	3 Pathogenic SNPs 97 Non-Pathogenic SNPs	No. of Datasets-100 No. of Samples-1600 with 800 cases and 800 controls
	DNME Models	10 Models		
2-locus dataset	DME Models-Additive Model, Multiplicative, Threshold models	4 Models	2 Pathogenic SNPs 98 Non-Pathogenic SNPs	
	DNME Models	10 Models		

4.2 *Performance metrics*

The efficacy of the proposed epistasis detection models is evaluated using evaluation metrics power. Power is a statistical measure of detecting true disease loci by rejecting the null hypothesis, and the same is expressed as

$$Power = \frac{\#Dcount}{TDS}$$

where #Dcount represents the number of successful detection of datasets containing disease-related SNPs among the Total number of Datasets (TDS) produced by the same criteria and penetrance table.

4.3 *Simulation results and interpretation*

The primary focus of GWAS is to identify associations between SNP and phenotype. The epistasis identification is essential for determining human genetic disease susceptibility. In this section, the performance of KMeans with Nelder optimization and Salp optimization technique for Epistasis detection (KMeans-NM-SalpEpi) is compared with epistasis detection ability of MACOED [11], SSA with G-test fitness function using DNME and DME models.

4.3.1 *Experimental results of 2-locus DME models*

Table 2 and Figure 5 expose the detection power of epistasis approaches of the 2-locus DNME models. MACOED [11] and SalpEpi-MO did not find even a single disease causative SNP pair among the 100 datasets in additive model 1. In additive model 2, KMeans-SalpEpi-SO obtained the highest power of 99%, whereas SalpEpi-MO obtained the lowest power of 80%. In multiplicative model 1, none of the methods found any disease causative SNP pairs. In multiplicative model 2, KMeans-Epi-SO and KMeans-Epi-MO achieved 100% power. In multiplicative model 3, SalpEpi-MO yielded 20% of power, whereas the remaining method did not find any SNP pairs correlated to diseases.

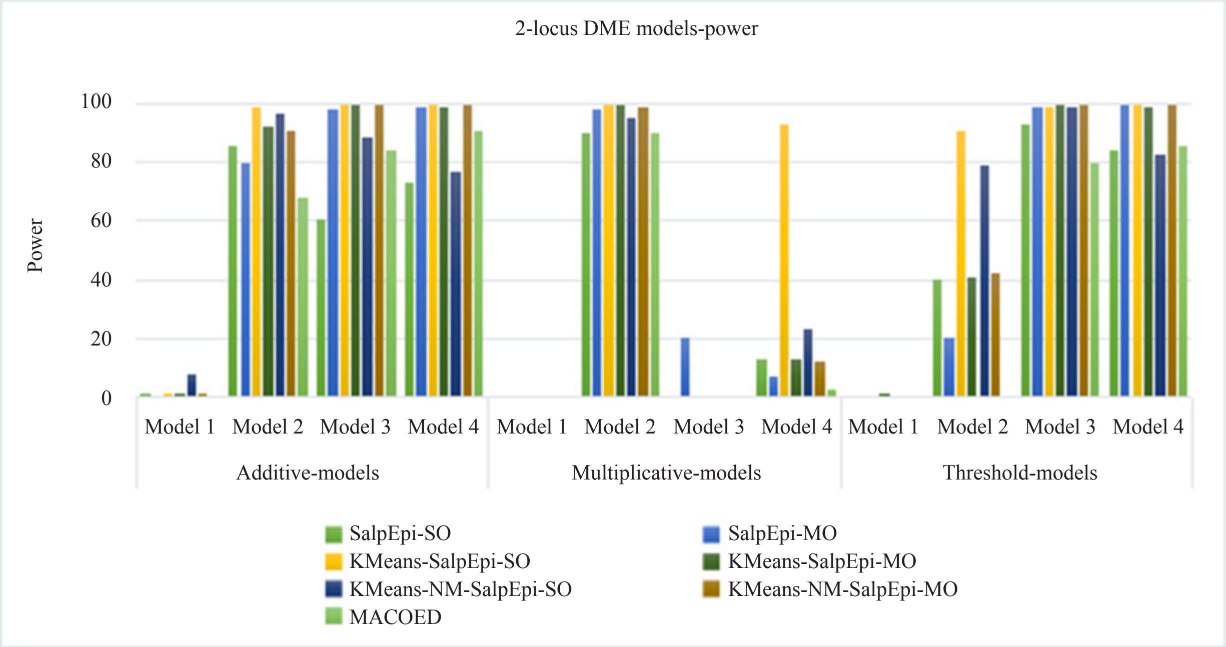


Figure 5. Detection power comparison of 2-locus DME models

In multiplicative model 4, KMeans-SalpEpi-SO achieved the highest power of 93%, whereas SalpEpi-MO obtains the lowest detection power of 7%. In threshold model 3, KMeans-SalpEpi-MO and KMeans-NM-SalpEpi-MO yielded 100% power. In threshold model 4, KMeans-SalpEpi-SO, SalpEpi-MO, and KMeans-NM-SalpEpi-MO achieved 100% power. KMeans-SalpEpi-SO gained the highest detection power of 93%, and SalpEpi-MO yielded the lowest detection power of 20% for threshold model 2. In threshold model 1, all the methods lost detection power of 0 except KMeans-SalpEpi-MO, which has gained the power of 1%. Among the 12 DME models, KMeans-NM-SalpEpi-MO obtained 100% detection power in 3 models, whereas MACOED [11] and SalEpi-SO didn't find 100% detection power for even a model. The experimental finding proved that KMeans-NM-SalpEpi-MO yielded superior detection power compared to others.

Table 2. Detection power of 2-locus DME models

Model	Additive-Models				Multiplicative-Models				Threshold-Models			
Methods	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
MACOED [11]	0	68	84	91	0	90	0	3	0	0	80	86
SalpEpi-SO	1	86	61	73	0	90	0	13	0	40	93	84
SalpEpi-MO	0	80	98	99	0	98	20	7	0	20	99	100
KMeans-SalpEpi-SO	1	99	100	100	0	100	0	93	0	91	99	100
KMeans-SalpEpi-MO	1	92	100	99	0	100	0	13	1	41	100	99
KMeans-NM-SalpEpi-SO	8	97	89	77	0	95	0	23	0	79	99	83
KMeans-NM-SalpEpi-MO	1	91	100	100	0	99	0	12	0	42	100	100

4.3.2 *Experimental results of 2-locus DNME models*

Table 3 and Figure 6 expose the detection power of epistasis approaches of the 2-locus DNME models. Among the 10 DNME models, KMeans-NM-SalpEpi-MO achieved 100% power for all the models except model 5, and it is superior to others. The KMeans-NM-SalpEpi-SO achieved 100% detection power for the models M1, M2, M5, and M8. MACOED [11] and SalpEpi-SO didn't find 100% power in any DNME models. KMeans-SalpEpi-SO achieved 100% power for four models: model 1, model 2, model 5, and model 6. KMeans-SalpEpi-MO obtained 100% power for 6 DNME models such as model 1, model 4-model 6, model 8, and model 9. The SalpEpi-MO yielded 100% power for two models: model 8 and model 9. The highest detection power of SalpEpi-SO and MACOED [11] is 97% and 93%, respectively. Even though SalpEpi-SO is inferior to SalpEpi-MO and all the K-Means clustering-based approaches, its performance is superior to MACOED [11] in all DNME models. The result designates that one of the proposed approaches KMeans-NM-SalpEpi-MO is superior compared to others.

Table 3. Detection power of 2-locus DNME models

Model	DNME models									
Methods	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
MACOED [11]	92	88	88	86	93	93	86	93	92	93
SalpEpi-SO	91	90	95	90	94	95	87	94	97	92
SalpEpi-MO	99	98	97	96	99	97	98	100	100	99
KMeans-SalpEpi-SO	100	100	93	97	100	100	99	99	99	99
KMeans-SalpEpi-MO	100	99	99	100	100	100	99	100	100	99
KMeans-NM-SalpEpi-SO	100	100	91	96	100	99	99	100	99	99
KMeans-NM-SalpEpi-MO	100	100	100	100	99	100	100	100	100	100

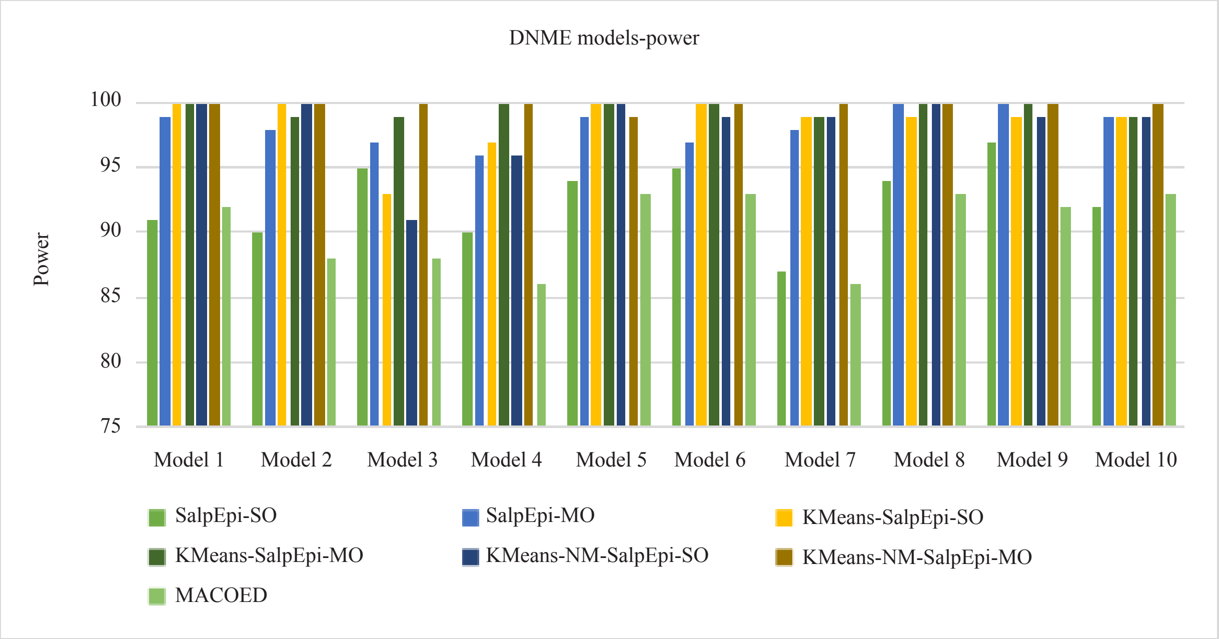


Figure 6. Performance comparison of 2-locus DNME models

4.3.3 *Experimental results of 3-locus DME models*

The power of the discussed epistasis approaches for fifteen 3-locus DME models is exhibited in Figure 7, and the same is presented in Table 4. Due to the computational complexity, the state-of-the-art method MACOED was only evaluated on 2-locus interactions. As a result, MACOED was not considered comparative methodologies for the analysis of three-locus disease models in this research. For the additive model, KMeans-SalpEpi-MO and KMeans-NM-SalpEpi-MO achieved 56% and 55% of power, respectively. These two models are superior to other approaches for model 1. In additive model 2, KMeans-SalpEpi-MO achieved the highest detection power of 28%, whereas the SalpEpi-SO achieved the lowest detection power of 6%. The KMeans-NM-SalpEpi-MO yielded 74% of power for additive model 3. In additive model 4, KMeans-SalpEpi-MO obtained the highest power of 83%, which is superior to the others. KMeans-NM-SalpEpi-MO obtained power of 90% for model 5, which is 86%, 20%, 71%, 2%, 71% superior to SalpEpi-SO, SalpEpi-MO, KMeans-SalpEpi-SO, KMeans-SalpEpi-MO, KMeans-NM-SalpEpi-SO, respectively.

For the multiplicative model 5, KMeans-NM-SalpEpi-MO produced the highest detection power of 80%. For multiplicative model 1 and model 2, KMeans-NM-SalpEpi-MO yielded 9% and 6% detection power, respectively, which is superior to others. KMeans-SalpEpi-MO yielded the highest detection power of 10% for model 4. For model 5, KMeans-SalpEpi-MO and KMeans-NM-SalpEpi-MO have arrived same detection power of 10%, which is superior to others. The highest detection power of 81% is yielded for the threshold model 5, whereas Salp-SO has earned the lowest detection power of 5%. In threshold model 4, KMeans-SalpEpi-MO yielded the power of 80%, which is 1% higher than KMeans-NM-SalpEpi-MO. KMeans-NM-SalpEpi-MO gained the highest detection power of 80%, whereas KMeans-SalpEpi-SO earned the lowest detection power of 48%. In threshold model 1 and model 2, KMeans-SalpEpi-MO is superior to others. The experimental result proved that the performance of KMeans-SalpEpi-MO and Kmeans-NM-SalpEpi-MO is superior to others over 15 DME models.

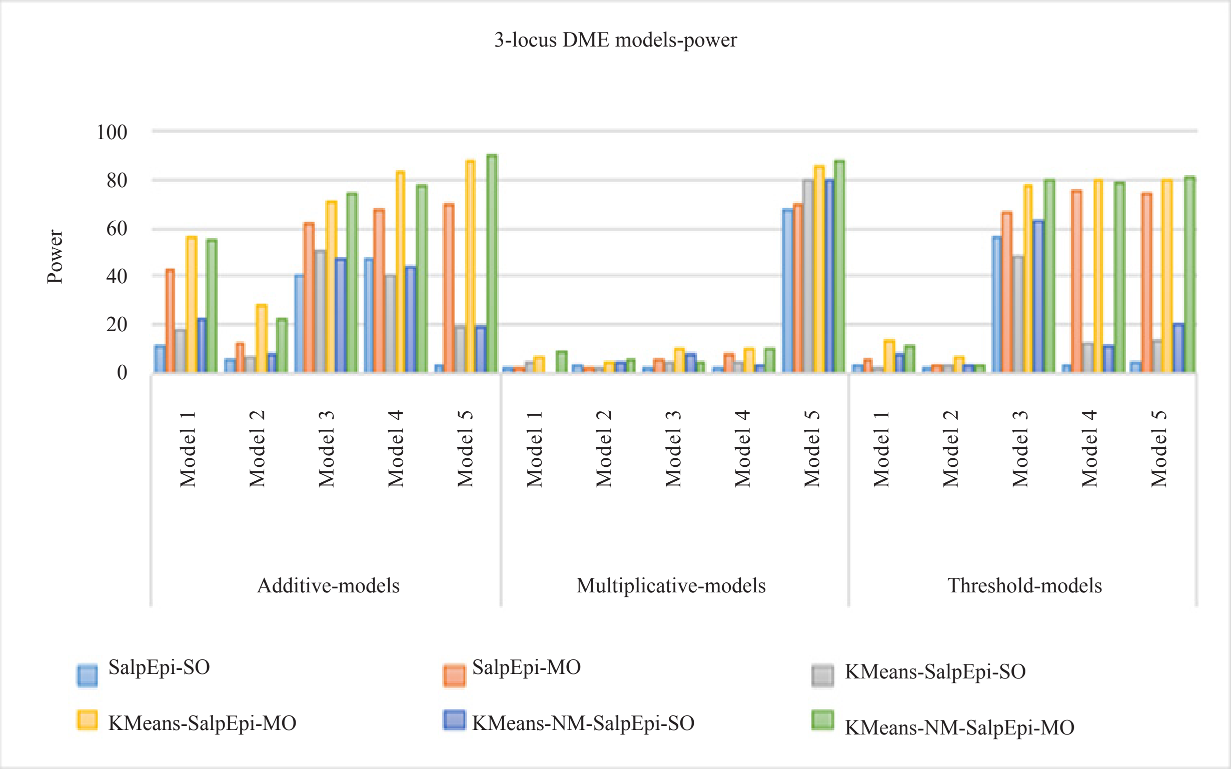


Figure 7. Performance evaluation of 3-locus DME models

Table 4. Performance evaluation of 3-locus DME model

Model	Additive models					Multiplicative models					Threshold models				
Methods	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
SalpEpi-SO	11	6	40	47	4	2	3	1	2	67	3	1	56	4	5
SalpEpi-MO	43	12	62	67	70	2	2	6	8	70	6	4	66	75	74
KMeans-SalpEpi-SO	18	7	51	41	19	5	2	5	5	80	2	4	48	13	14
KMeans-SalpEpi-MO	56	28	71	83	88	7	5	10	10	86	14	7	78	80	80
KMeans-NM-SalpEpi-SO	23	8	47	44	19	0	5	8	3	80	8	3	63	11	20
KMeans-NM-SalpEpi-MO	55	23	74	78	90	9	6	5	10	88	11	4	80	79	81

4.3.4 *Experimental results of 3-locus DNME models*

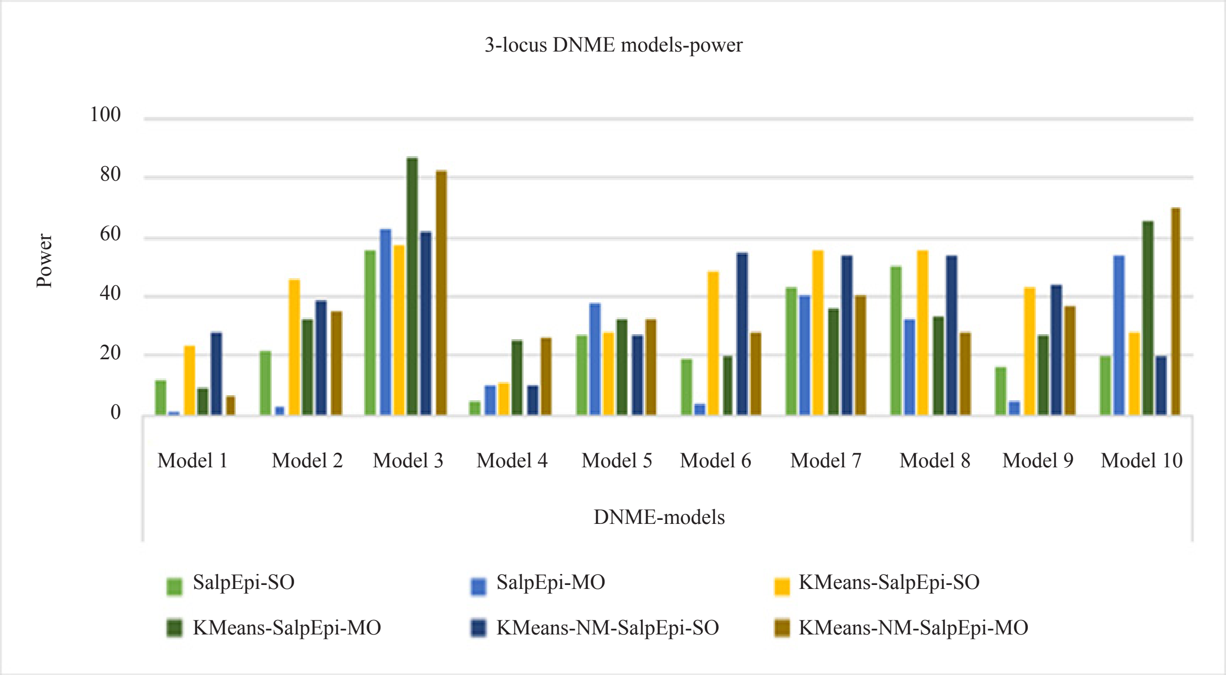


Figure 8. Performance Comparison of 3-locus DNME Models

The power of ten 3-locus DNME models is exhibited in Figure 8, and the same is presented in Table 5. KMeans-SalpEpi-MO obtained the highest accuracy of 87% for Model 3, while SalpEpi-SO gained the lowest detection power of 1% for Model 1. The experimental outcome revealed that the KMeans clustering-based approaches are superior to Salp-MO and Salp-MO for all the 3-locus DNME models.

Table 5. Performance Evaluation of 3-locus DNME Models

Model	DNME Models									
Methods	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
SalpEpi-SO	12	22	56	5	27	19	43	50	16	20
SalpEpi-MO	1	3	63	10	38	4	40	32	5	54
KMeans-SalpEpi-SO	23	46	57	11	28	48	56	56	43	28
KMeans-SalpEpi-MO	9	32	87	25	32	20	36	33	27	65
KMeans-NM-SalpEpi-SO	28	39	62	10	27	55	54	54	44	20
KMeans-NM-SalpEpi-MO	6	35	82	26	32	28	40	28	37	70

5. Conclusion

Genetic interactions play a vital role in the identification of complex human diseases. In this article, we suggest a two-stage approach called KMeans-NM-SalpEpi-SO and KMeans-NM-SalpEpi-MO. In the screening stage, the hybrid algorithm based on K-Means clustering algorithm and Nelder-Mead optimization technique was used to split the genotype dataset into various clusters. In the search stage, these clustered dataset is passed to the salp optimization to find the epistasis effects. The experiments are conducted over the simulated dataset. Experimental findings proved that KMeans-NM-SalpEpi-SO and KMeans-NM-SalpEpi-MO are superior to MACOED, SalpEpi-SO, and SalpEpi-MO for all 2-locus and 3-locus DNME and DME models. The future scope of this research work may be extended to assess the real datasets for diagnosing complex diseases in humans and also use some other clustering technique to group the SNPs.

References

[1] Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genomewide association analysis by Lasso Penalized Logistic Regression. *Bioinformatics*. 2009; 25: 714-721. Available from: doi: 10.1093/bioinformatics/btp041.

[2] Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*. 2007; 39: 1167-1173. Available from: doi: 10.1038/ng2110.

[3] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American Journal of Human Genetics*. 2012; 90: 7-24. Available from: doi: 10.1016/j.ajhg.2011.11.029.

[4] Genetics Home Reference. *What Are Single Nucleotide Polymorphisms (SNPs)?* 2019. Available from: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp [Accessed 19th August 2021].

[5] Mackay TFC, Moore JH. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*. 2014; 6: 42. Available from: doi: 10.1186/gm561.

[6] Musani SK, Shriner D, Liu NJ, Feng R, Coffey CS, Yi NJ, et al. Detection of Gene × Gene Interactions in Genome-Wide Association Studies of Human Population Data. *Human Heredity*. 2007; 63: 67-84. Available from: doi: 10.1159/000099179.

[7] Priya S, Manavalan RK. Genetic interactions effects of cardiovascular disorder using computa-tional models: A review. *Current Biotechnology*. 2020; 9. Available from: doi: 10.2174/2211550109999201008125800.

[8] Yang CH, Chuang LY, Lin YD. Fuzzy logic system application for detecting SNP-SNP interaction. *IEEE Access*. 2020; 8: 49951-49960. Available from: doi: 10.1109/ACCESS.2020.2977108.

[9] Sun LY, Liu GX, Su LT, Wang RQ. SEE: A novel multi-objective evolutionary algorithm for identifying SNP epistasis in genome-wide association studies. *Biotechnology & Biotechnological Equipment*. 2019; 33: 529-547.

Available from: doi: 10.1080/13102818.2019.1593052.

[10] Tuo SH, Zhang JY, Yuan XG, He ZZ, Liu YJ, Liu ZW. Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations. *Scientific Report*. 2017; **7**: 1-18. Available from: doi: 10.1038/s41598-017-11064-9.

[11] Jing PJ, Shen HB. MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*. 2014; 31. Available from: doi: 10.1093/bioinformatics/btu702.

[12] Sun YX, Shang JL, Liu JX, Li SJ, Zheng CH. EpiACO-A method for identifying epistasis based on ant Colony optimization algorithm. *BioData Mining*. 2017; 10: 1-17. Available from: doi: 10.1186/s13040-017-0143-7.

[13] Li XT, Zhang SX, Wong KC. Nature-inspired multiobjective epistasis elucidation from genome-wide association studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2018. p. 1. Available from: doi: 10.1109/TCBB.2018.2849759.

[14] Sitarcik J, Lucka M. EpiBAT: Multi-objective bat algorithm for detection of epistatic interactions. In: *2019 IEEE 15th International Scientific Conference on Informatics*. 2019. Available from: doi: 10.1109/Informatics47936.2019.9119310.

[15] Priya S, Manavalan RK. Multi-objective chaotic atom search optimization for epistasis detection in genome-wide association studies. *BT-Proceedings of International Conference on Scientific and Natural Computing*. Singapore: Springer Singapore; 2021. p. 11-22.

[16] Su T, Dy J. A deterministic method for initializing K-means clustering. *Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence ICTAI*. 2004. p. 784-786. Available from: doi: 10.1109/ICTAI.2004.7.

[17] Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. Salp swarm algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software*. 2017; 114. Available from: doi: 10.1016/j.advengsoft.2017.07.002.

[18] Vakil Baghmisheh MT, Peimani M, Sadeghi MH, Ettefagh MM, Tabrizi AF. A hybrid particle swarm-Nelder-Mead optimization method for crack detection in cantilever beams. *Applied Soft Computing*. 2012; 12: 2217-2226. Available from: doi: 10.1016/j.asoc.2012.03.030.

[19] Blondin MJ, Sanchis J, Sicard P, Herrero JM. New optimal controller tuning method for an AVR system using a simplified Ant Colony Optimization with a new constrained Nelder-Mead algorithm. *Applied Soft Computing*. 2018; 62: 216-229. Available from: doi: 10.1016/j.asoc.2017.10.007.

[20] Li J, Qiao BM. A Novel differential evolution algorithm with K-Means and simplex search method for absolute value equations. In: *2015 11th International Conference on Computational Intelligence and Security (CIS)*. 2015. p. 266-229. Available from: doi: 10.1109/CIS.2015.72.

[21] Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH. GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*. 2012; 5. Available from: doi: 10.1186/1756-0381-5-16.

[22] Chen QF, Zhang X, Zhang RC. Privacy-preserving decision tree for epistasis detection. *Cybersecurity*. 2019; 2. Available from: doi: 10.1186/s42400-019-0025-z.