Review of Computer Engineering Research
2021 Vol. 8, No. 2, pp. 27-40.
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/journal.76.2021.82.27.40

TWO-STAGE MODEL WITH ROUGH CLUSTER AND SALP OPTIMIZATION TECHNIQUE FOR EPISTASIS DETECTION

S Priya[1+]
R Manavalan[2]

[1]*Research Scholar, Department of Computer Science, Arignar Anna Government Arts College, Villupuram, Tamilnadu, India.*
Email: priyasri.ash@gmail.com
[2]*Assistant Professor, Department of Computer Science, Arignar Anna Government Arts College, Villupuram, Tamilnadu, India.*
Email: manavalan_r@rediffmail.com Tel: 9865125145

(+ Corresponding author)

Article History
Received: 12 July 2021
Revised: 16 August 2021
Accepted: 7 September 2021
Published: 23 September 2021

Keywords
Epistasis
Genetic interactions
Salp optimization
Cluster
Rough k-means
Single nucleotide polymorphism (SNPs).

ABSTRACT

The discovery of gene-gene interactions to identify complex diseases is one of the primary challenges in genome-wide association studies (GWAS). Genetic interactions (Epistasis) are typically seen as interactions between various single nucleotide polymorphisms (SNPs). Genetic interactions discovery can assist the researchers in identifying gene pathways, recognizing gene activity, and discovering potential drug targets. Rough Cluster based Salp Optimization for Epistasis detection (RCSalp-Epi) is a two-stage epistasis model that has been evaluated on a variety of simulated disease models. In the screening stage, the rough clustering algorithm is employed to partition the genotype data into different clusters. The selection stage presents Salp optimization with a single objective function (SalpEpi-SO) and multiple objective functions (SalpEpi-MO) over the clusters to find disease-related SNP combinations. RCSalp-performance Epi's is evaluated in comparison with SalpEpi-SO and SalpEpi-MO. The outcome of the experimental analysis revealed that RCSalp-Epi-MO is superior to SalpEpi-SO and SalpEpi-MO in terms of power and execution time.

Contribution/Originality: The paper's primary contribution is finding the higher order genetic interactions with high detection power and minimal computational effort.

1. INTRODUCTION

The growing interest in medical science is to examine the genetic architecture of diseases, especially non-Mendelian diseases. It plays a crucial role in predicting complex diseases in human beings [1]. With the recent emergence of high-performance genotype technologies, many emphases have been dedicated to identifying the correlation between genes and complex diseases. Both genetic and environmental risk factors increase pathogenicity. Genome-wide association studies (GWAS) researchers aim to find genotype variants of interest for several diseases such as hypertension, rheumatoid arthritis, cancer, chronic illness, cardiovascular disease, diabetes, psoriasis, etc. [2]. The principal aim of GWAS is to associate the related genetic variants to phenotypic traits of interests, especially a disease [3]. GWAS incorporates extensive data collection to trace phenotypes and genetic markers linked with disease [4]. Single Nucleotide Polymorphisms (SNPs) are a typical marker of genetic variations that play a crucial role in many complex disease traits [5]. The SNP is a sequence variation in DNA dependent on the four nucleotides Cytosine (C), Thyamin (T), Adenine (A), and Guanine (G), and also modifies the amino acid sequence [6]. By analyzing the gene regulatory pathways, each SNP is linked to a set of characteristics that can be used to classify the genetic predisposition of diseases [7]. Several approaches for Gene-Gene Interaction

(GGIs) detection have been employed in recent years. Epistasis detection can currently be achieved using stochastic search, exhaustive search, and optimization-based methods. An exhaustive search can return all SNP combinations, but the computational cost is prohibitive [8]. It calculates the score for each SNP interaction and then uses the user-specified threshold to make a disease correlation determination. The epistasis based stategies such as PLINK, MDR, BOOST, GMDR-GPU are assessed using exhaustive analysis [9]. The Multifactor Dimension Reduction (MDR) methodology was employed in 2001 to search for GGIs related biomarkers in breast cancer genotype data [10]. Boolean Operations-Based Screening and Testing (BOOST) method uses screening and testing stages to evaluate GGIs. A non-iterative strategy was widely adopted during the screening process to determine the statistical probability ratio of all SNP pairs, and a distinctive SNP pair was chosen based on the threshold defined by the user. The probability ratio test was performed to quantify the selected SNP combinations during the test process [11].

The stochastic search strategies identify effects of disease correlated epistasis using random sampling. The stochastic search takes far less time to complete than an exhaustive search Since it is influenced by the random seed [12]. BEAM, SNPRuler algorithm uses random sampling techniques to evaluate SNP combinations. SNPRuler adapts a predictive rule inference strategy to describe rules in the SNP subset, and these rules are used to infer epistatic interactions [13]. Bayesian Epistasis Mapping Association (BEAM) examines disease-related markers and the correlations through the Bayesian partitioning model. It estimates the posterior likelihood ratio of each diseased SNP markers via Markov Chain Monte Carlo (MCMC) [4].

The exhaustive and stochastic algorithms lead to a high computational cost and affinity for specific disease models. In recent times, evolutionary methods for GGIs detection have been of great concern to minimize computational costs since they efficiently solve NP-hard issues in polynomial times [14]. The evolutionary strategies minimize search time complexity and the scoring functions used to determine the better SNP combinations. A multi-objective ant colony optimization technique (MACOED) was introduced for the detection of genetic interactions. ACO is practiced to filter SNPs in the screening stage and the filtered SNPs passed into the clean stage to detect significant SNP combinations using the chi-square test [15]. Epistasis based on Ant Colony Optimization Algorithm (epiACO) was introduced to recognize SNP interactions. The different strategies for path selection and a memory-based approach are adapted to improve epiACO [16]. An Epistatic Interaction Multi-Objective Artificial Bee Colony Algorithm Based on Decomposition (EIMOABC/D) model was suggested for epistasis interaction detection. Bayesian score and the Gini score are adopted as objective functions to characterize different epistatic models [17]. The multi-objective bat optimization algorithm (epiBat) was presented for epistasis identification using the Gini score and K2 score as the fitness function. Finally, the G test assesses the significance of the identified disease-related SNP pairs [18]. The primary problem of presently accessible epistasis algorithms is always incurring a huge computational cost and minimal detection power. Compared to the presently available methods, the proposed method aims to classify disease-correlated SNPs with huge detection capacity and minimal runtime.

We have recently introduced a two stage approach based on KMeans clustering, pillar algorithm and salp optimization technique (KMeans-Pillar-SalpEpi) for epistasis detection. The KMeans-Pillar-SalpEpi involves high computational complexity since K-Means approach leads to inconsistent cluster group for different runs due to optimal centroids and hence hybrid with pillar optimization to increase the efficiency [19]. This research introduced a novel epistasis detection strategy with a two-stage process called Rough Cluster-based Salp Optimization for Epistasis detection (RCSalp-Epi) to identify multi-locus SNP interactions. Traditional clustering assigns items to non-overlapping groups on the basis of a similarity score. The borders of these groups or clusters may not be precisely defined in the real world. Some of the objects may be nearly equidistant from the cluster's centroid. These objects must be assigned to a single cluster according to traditional set theory. The overlapping clusters can be represented using rough set theory. Compared to the k-means clustering approach, rough sets give a more flexible representation. At the screening phase, the Rough K-means algorithm is employed to partition the genotype data into different clusters. In the search phase, two distinct strategies, such as exhaustive search and Salp optimization with G-test over the clusters to detect the disease-relevant SNP pairs. The main scope of this research is to establish a rapid and efficient epistasis detection model RCSalp-Epi to discover disease-related SNP-SNP interactions from hundreds of SNPs. Here, we have fixed the following primary objectives to accomplish our research goal.

- A Rough Cluster-based Salp Optimization for epistasis detection (RCSalp-Epi) is developed to detect disease-related SNPs in genotype data. The proposed approach is intended to detect high order epistasis interactions with reduced running time.
- The performance of RCSalp-Epi is measured over the disease models with marginal effects (DMEs) and disease models with no marginal effects (DNMEs).

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 RCSalp-Epi algorithm. Section IV explores experimental results and discussion. Finally, Section V summarizes this article with future scope.

2. MATERIALS AND METHODS

In this section, we formally introduce the two components of RCSalp-Epi approach such as Rough K-means cluster and Salp Swarm Algorithm (SSA) for genetic interactions identification. Rough K-Means clustering technique is adapted in the screening stage to group SNPs into three clusters. The SSA is applied in the selection stage to find high order SNP combinations.

2.1. Rough K-Means Clustering Technique

Lingras [20] developed the rough k-means algorithm by hybridizing rough set theory and the k-means technique [20]. It is a heuristic way of depicting each cluster based on the centre of a cluster. To cope with the complexity involved in cluster analysis, the Rough K- means algorithm combines a rough set-theoretic flavor to the traditional k means algorithm. The features of this algorithm are:

(1) A data object can only bound lower approximation of one class.

(2) If a particular object cannot be assigned to a class's lower approximation, it must be assigned to two or more classes' upper approximations.

(3) Each class's lower approximation is a subset of the same class's upper approximation.

Let us consider, $\overline{C_k}$ is the set of upper approximation of class C_k and $\underline{C_k}$ is the set of lower approximation of class C_k. The centroid C_j can be calculated in the following ways

If $\underline{C_k} \neq \emptyset \ and\ \overline{C_k} - \underline{C_k} = \emptyset$

$$C_j = \sum\nolimits_{x \epsilon \overline{C_k}} \frac{x_i}{\left|\underline{C_k}\right|}$$

Else

If $\underline{C_k} = \emptyset \ and\ \underline{C_k} - \overline{C_k} \neq \emptyset$

$$C_j = \sum\nolimits_{x \in \underline{C_k} - \overline{C_k}} \frac{x_i}{\left|\underline{C_k} - \overline{C_k}\right|}$$

Else

$$C_j = w_{lower} \times \sum\nolimits_{x \in \overline{C_k}} \frac{x_i}{\left|\overline{C_k}\right|} + w_{upper} \times \sum\nolimits_{x \in \overline{C_k} - \underline{C_k}} \frac{x_i}{\left|\overline{C_k} - \underline{C_k}\right|}$$

2.2. Salp Swarm Algorithm (SSA)

Salp swarm algorithm (SSA), a population-based optimization [21]. The SSA imitates salps' social actions as they are collected in a chain during their sailing and foraging for foods in the sea. Two kinds of agents present in SSA: The leader leads the salp chain; the remaining salps are followers; the leader controls the population's movement path; 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 that has N rows and D columns. 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:

$$x_k^1 = F_k + c_1\big((ub_k - lb_k)c_2 + lb_k\big)c_3 < 0.5$$

$$= F_k - c_1\big((ub_k - lb_k)c_2 + lb_k\big)c_3 \geq 0.5 \qquad (1)$$

where,

x_k^1 represents the position of the salps in kth dimension.

$F_k$indicates the location of the food in kth dimension.

ub_k represents the upper limit of the kth dimension.

lb_krepresents the upper bound of the kth dimension.

$c_{1,}c_2, and\ c_3$ indicates random numbers.

The convergence factor c1 aid the process of exploration and exploitation, which is calculated using the Equation 2.

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

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

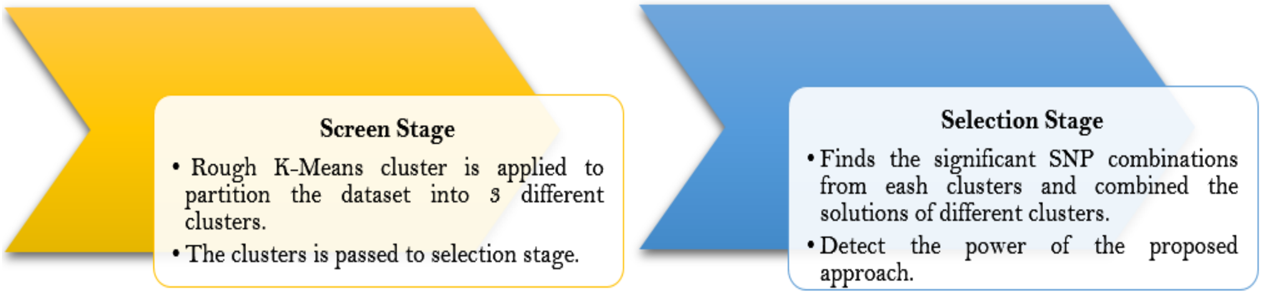
The follower's position is updated as shown in Equation 3.

$$x_d^n = \frac{1}{2}(x_d^n + x_d^{n-1}) \tag{3}$$

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

3. FRAMEWORK OF ROUGH CLUSTER BASED SALP OPTIMIZATION FOR EPISTASIS DETECTION (RCSALP-EPI)

The RCSalp – Epi consists of two stages, such as screen and clean stage. The objective of the screen and the clean stage is exposed in Figure 1. The general structural design of the proposed system is expressed in Figure 2. In the screening stage, the rough K-means clustering technique partitions the SNPs into three clusters. These clusters are passed to the selection stage to detect disease related SNP combinations. A detailed description of the screen and selection stage is exposed in section 3.1 and 3.2.



Screen Stage

- Rough K-Means cluster is applied to partition the dataset into 3 different clusters.
- The clusters is passed to selection stage.

Selection Stage

- Finds 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 RCSalp-Epi approach.

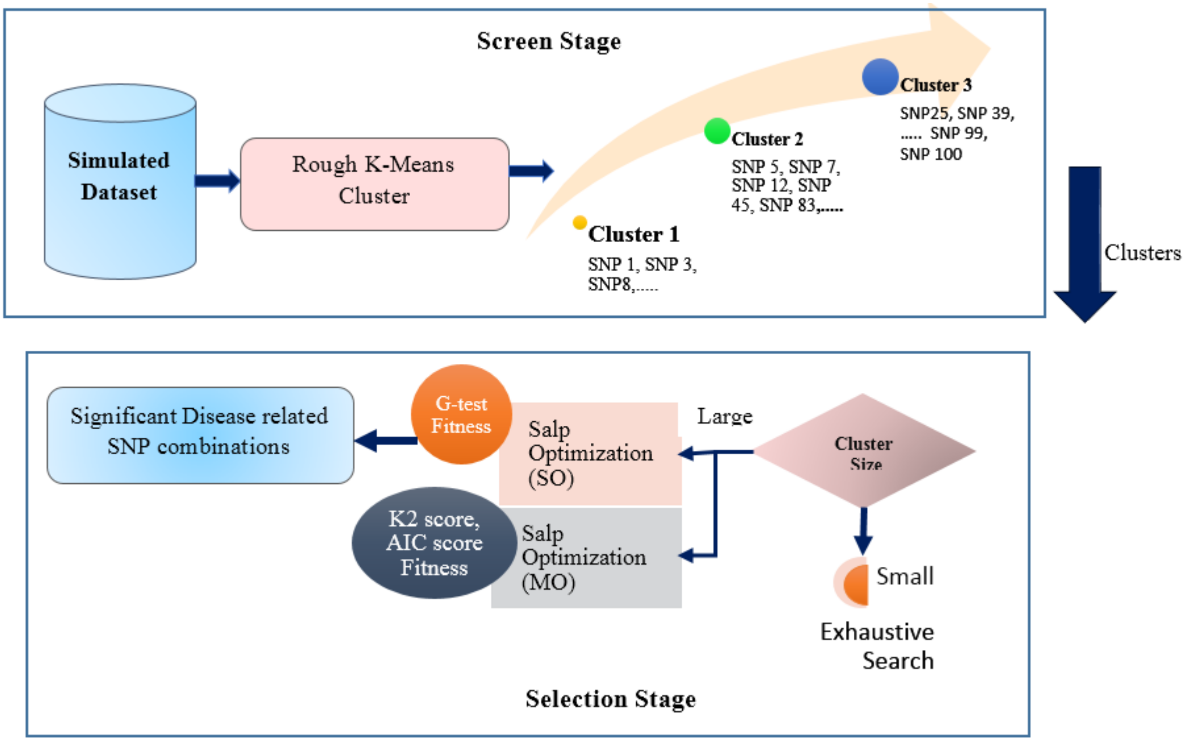


Figure-2. General Structure of RCSalp-Epi.

3.1. Screen Stage – Rough K- Means Clustering Technique

The genotype dataset is separated into three different clusters using the rough clustering technique in the screening stage. Based on a similarity measure, traditional clustering procedures segments a group of items into various non-overlapping clusters. The boundaries of these segments may not be accurately defined in the real world. Some objects can close to the centre of the cluster. These objects are consigned to a single cluster according to traditional set theory. The overlapping clusters can be represented using rough set theory. The main advantage of the rough clustering strategy is that it avoids local optimum and groups all SNPs into a distinct set of clusters for each iteration. The pseudo-code of the screen is exhibited in Figure 3.

```
Stage 1 – Rough K Means Clustering for grouping SNPs for Epistasis Detection
Input
Data: Simulated dataset
k: number of clusters
$W_{lower}$
$W_{upper}$
Output
Three Clusters consists of various SNPs
Screen Stage
Step 1: Randomly assign each SNPs into lower approximization.
Step 2: Calculate new mean value by finding the no. of elements in the lower bound and the upper bound.
Step 3: Find the distance between each SNPs and centroid
Step 4: Find the objects whose difference is less than epsilon and keep the object in upper bounds of  cluster
Step 5: Repeat steps (2) to (4) until convergence.
```

Figure-3. Pseudo Code of RCSalp-Epi in Screen Stage

3.2. Selection Stage - Salp Optimization for Epistasis Detection

In the selection stage, the size of each cluster is checked. Then, an Exhaustive search is adapted for the clusters with less number of SNPs. For large cluster sets, SSA is adapted to identify the epistasis effects. The G-test is used as a fitness function for SSA. The Salp with SO (SSO) and Salp with MO (SMO) optimization are proposed to find the significant disease associated SNPs. The fitness function for SalpEpi-SO is G-test, while SalpEpi-MO utilizes the K2 and AIC score as fitness functions. The pareto optimal approach aid to choose non-dominated SNPs from the large volume of SNPs. Then, non-dominated SNPs are passed into G-test to find disease correlated SNPs for 2-locus and 3-locus models. The pseudo-code of selection stage for RCSalp-Epi is presented in Figure 4.

```
Stage 2 – Salp Optimization for finding Significant SNP Combinations
Input
Data: Simulated dataset based on selected featured indexed SNPs from screen stage
N: number of Salp
m: interaction order
max_iter: Maximum iterations
Output
Optimal SNP pairs
Selection Stage – SalpEpi-SO
Step 1: Initialize the necessary parameters for SSA.
Step 2:  Every salp is assigned a random position in the population based on the SNPs in the feature set.
Step 3: Repeat the step until maximum iterations reached
Choose a combination of SNPs for each salp in the solution space and generate a local solution through G-test.
Choose a source of food from repository: F=SelectFood(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 Compared it with the previously stored solution space and update the current solution.
End for
End while
```

```
Selection Stage – SalpEpi-MO
Steps 1 through 3 in SalpEpi-SO are the same in SalpEpi-MO.
Step 4:  The Non-dominated SNPs are returned by the Pareto optimal technique.
Step 5: For $k$=1 to no. of SNPs in non-dominated solution
For $l$=k+1 to no. of SNPs in non-dominated solution
$GGIs\_pair$ = G-test ($x_k$, $x_l$)
End For
```

Figure-4. Pseudo Code of RCSalp-Epi in Selection Stage

4. EXPERIMENTAL ANALYSIS 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 SalpEpi-MO model. Sections 4.1 and 4.2 describe the simulations models and evaluation metrics, respectively. MATLAB R2018(b) is used to implement the proposed epistasis models. The experimental results of simulated disease models are revealed in Section 4.3.

4.1. Simulated Datasets

The efficacy of the RCSalp-Epi algorithm 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. For a disease model, these probabilities are gathered in a penetrance table. A penetrance is denoted by P(D|Gi), where D represents someone affected by the disease, and Gi denotes the ith genotype combinations of SNPs. GAMETES 2.0 is commonly used to build genotype simulation datasets. We created two-locus disease models in this study [22]. Two unique types of epistatic models are produced for two-locus analysis in order to identify diseases: DME and DNME models. DME model characterizes the interactive and marginal effects of the disease. Three gene models such as additive, multiplicative and threshold models are chosen for three-locus and two-locus analysis [23]. DNME model reveals only interactive effects without marginal effects. Gametes were used to construct the data sets for the study, which were varied in terms of heritability h^2 and Minor Allele Frequency (MAF), and disease prevalence rate P(D). Table 1 lists the DME and DNME models that were chosen for experimentation.

Table-1. Simulated epistasis dataset details.

Dataset	Model	Number of Models	SNP Particulars	Description
3-Locus model	DME epistasis models - Additive, Multiplicative, Threshold	5 Models	A total of 97 non-pathogenic SNPs and 3 disease related SNPs	Datasets size - 100 Samples size – 1600 with 800 cases and controls
	DNME Models	10 Models		
2-Locus model	DME epistasis models - Additive Model, Multiplicative, Threshold models	4 Models	2 disease correlated SNPs including 98 Non-Pathogenic SNPs	
	DNME Models	10 Models		

4.2. Performance Metrics

The suggested epistasis detection model's efficacy is assessed utilizing evaluation criteria such as power. The statistical procedure of discovering true disease locus by neglecting the null hypothesis is known as power, and it is expressed as.

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

where #DC denotes the number of data sets of are successful in the detection of disease-associated SNPs is reported among Total Data Sets (TDS) [24].

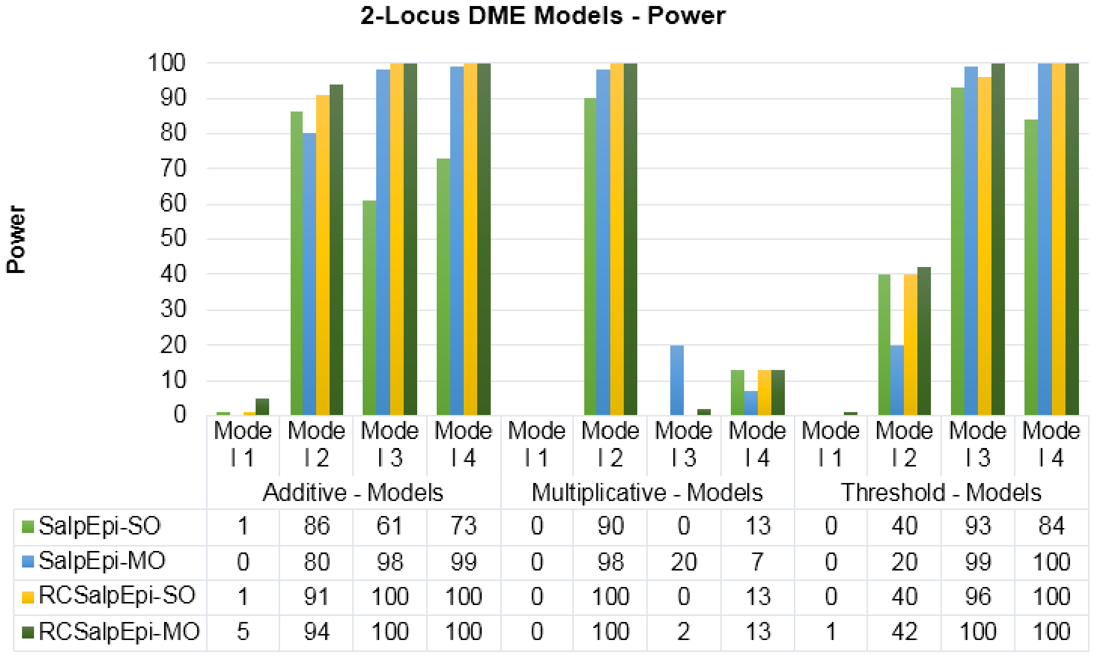
4.3. Simulation Results and Interpretation

The goal of GWAS is to find relationships between SNPs and phenotypes. The epistasis identification is essential for determining human genetic disease susceptibility. RCSalp-Epi's is contrasted to SalpEpi-SO [19] and SalpEpi-MO [19] techniques for epistatic identification.

4.3.1. Experimental Analysis of 2-Locus DME Models

The power of SalpEpi-SO [19] SalpEpi-MO [19] RCSalpEpi-SO and RCSalpEpi-MO for 12 DME models is exhibited in Figure 5. In the additive model, RCsalpEpi-SO and RCSalpEpi-MO achieved 100% power for model 3 and 4. For additive model 1, RCsalpEpi-MO yielded the power of 5%, whereas SalpEpi-SO and RCSalp-Epi gained the power of 1. SalpEpi-MO did not detect any disease causative SNPs. RCsalpEpi-MO obtains the highest power of 94% for additive model 3, which is superior to others.

For multiplicative model 1, none of the four techniques discovered any disease-related SNPs. The RCSalpEpi-SO and RCSalpEpi-MO obtained 100% of power for model 2, which is 2% and 10% higher than SalpEpi-SO and SalpEpi-MO, respectively. In multiplicative model 3, SalpEpi-MO [19] obtained the maximum power of 20%, which is 18% superior to RCSalpEpi-MO. In multiplicative model 4, SalpEpi-SO [19] RCSalpEpi-SO and RCSalpEpi-MO yielded the power of 13%, while SalpEpi-MO attained the power of 7%. RCSalpEpi-MO discovers a single illness causal SNP pair across 100 datasets in Threshold model 1, but the other three techniques have a power of 0. The power of 42% is yielded by RCSalpEpi-MO in Threshold model 2, whereas the SalpEpi-MO produced 20% power. In threshold model 3, RCSalpEpi-MO achieved 100% accuracy, 7%, 1%, and 4% superior to SalpEpi-SO [19] SalpEpi-MO [19] and RCSalpEpi-SO, respectively. According to threshold model 4, the power produced by all three methods hit 100%, while SalpEpi-SO [19] obtained 84% of its power.



	Mode l 1	Mode l 2	Mode l 3	Mode l 4	Mode l 1	Mode l 2	Mode l 3	Mode l 4	Mode l 1	Mode l 2	Mode l 3	Mode l 4
	Additive - Models				Multiplicative - Models				Threshold - Models			
🟩 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
🟨 RCSalpEpi-SO	1	91	100	100	0	100	0	13	0	40	96	100
🟩 RCSalpEpi-MO	5	94	100	100	0	100	2	13	1	42	100	100

Figure-5. Detection Power of 2-Locus DME Models.

Figure 6 shows the runtime of 12 DME models in 2-locus category. The proposed RCSalpEpi-SO technique consumes the lowest runtime compared to all 12 DME models. For majority of the 12 DME models the SalpEpi-MO [19] technique takes the highest time. The running time of SalpEpi-SO [19] is superior to SalpEpi-MO [19].

Compared to SalpEpi-MO [19] Rough clustering with SalpEpi-MO requires the lowest runtime. The runtime of additive model 3 and additive model 4, Multiplicative model 3 and 4 and threshold model 4 pay less than one minute in rough clustering-based approaches such as RCSalpEpi-SO and RCSalpEpi-MO. Since the rough cluster split these datasets into below 10 SNPs, these models were directly entered into exhaustive search instead of searching through Salp optimization. It is concluded that RCSalpEpi-SO and RCSalpEpi-MO is superior to SalpEpi-SO [19] and SalpEpi-MO [19] in connection with runtime for all 12 DME models.

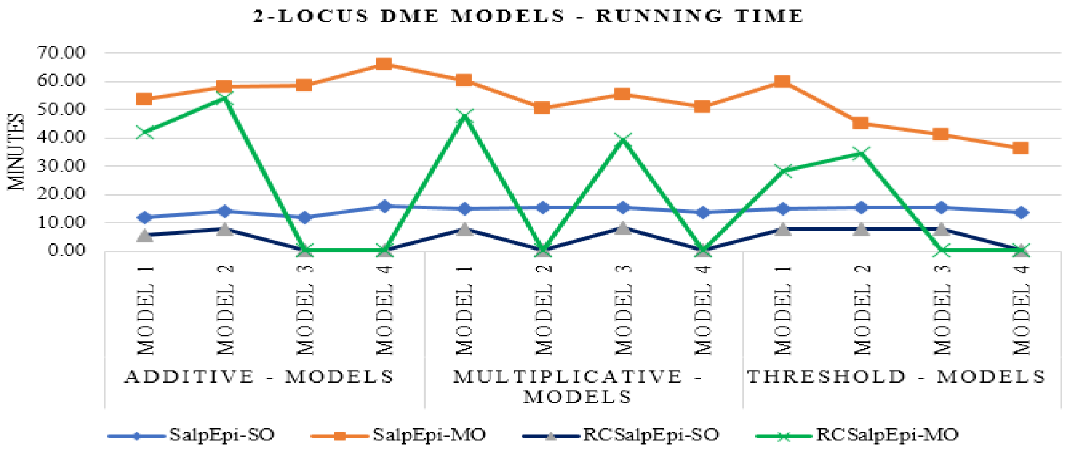
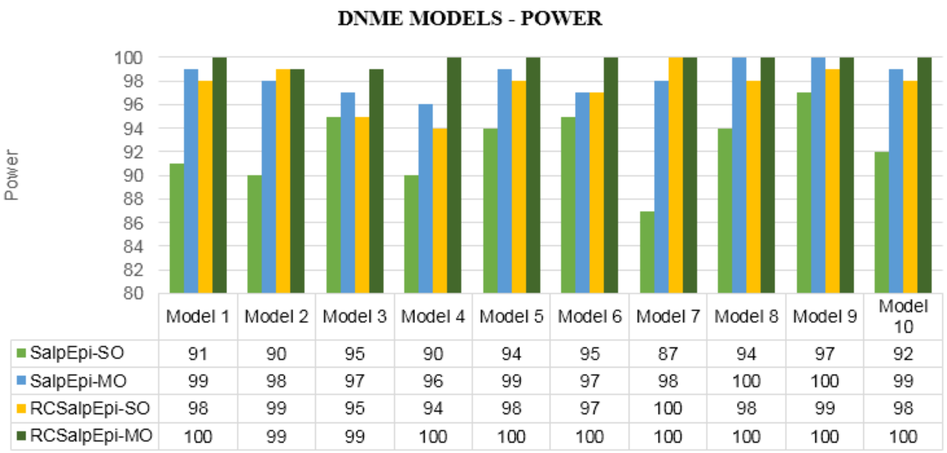


Figure-6. Running time comparison of 2-Locus DME Models.

4.3.2. Experimental Results of 2-Locus DNME Models

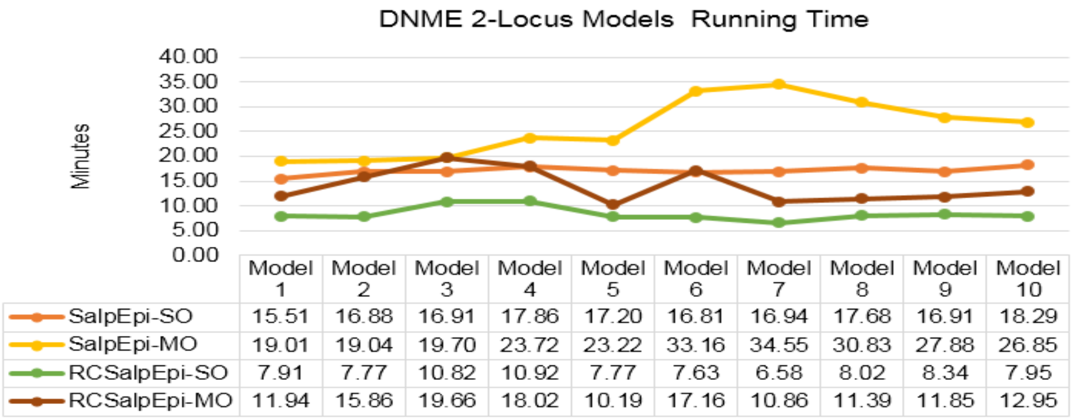
Figure 7 exposes the power of four approaches of the 2-Locus DNME models. With the exception of model 3 and model 4, RCSalpEpi-MO obtained 100% power for all ten DNME variants. The RCSalpEpi-SO only attained 100% power for model 7. In the case of model 8 and model 9, the SalpEpi-MO [19] provided 100 percent power. For model 9, the SalpEpi-SO [19] has the maximum detection capability of 97 percent. The result designates that the proposed approach RCSalpEpi-MO is superior to others.

Figure 8 presents the runtime of 2-locus DNME models. The RCSalpEpi-SO requires less time duration than the other approaches in 10 DNME variants. For all 10 models, the SalpEpi-MO [19] approach requires the most time to run. In model 4 and model 6, the execution time of RCSalpEpi-MO and SalpEpi-SO [19] are almost identical. Similarly, SalpEpi-MO [19] and RCSalpEpi-MO take the same execution time in model 3. The RCSalpEpi-SO requires minimal execution time in comparison with SalpEpi-MO and SalpEpi-SO approaches. It is concluded that the execution time of Rough clustering-based approaches such as RCSalpEpi-SO and RCSalpEpi-MO requires less duration for execution compared to SalpEpi-SO [19] and SalpEpi-MO [19].



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
SalpEpi-SO	91	90	95	90	94	95	87	94	97	92
SalpEpi-MO	99	98	97	96	99	97	98	100	100	99
RCSalpEpi-SO	98	99	95	94	98	97	100	98	99	98
RCSalpEpi-MO	100	99	99	100	100	100	100	100	100	100

Figure-7. Performance Comparison of 2-Locus DNME Models.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
SalpEpi-SO	15.51	16.88	16.91	17.86	17.20	16.81	16.94	17.68	16.91	18.29
SalpEpi-MO	19.01	19.04	19.70	23.72	23.22	33.16	34.55	30.83	27.88	26.85
RCSalpEpi-SO	7.91	7.77	10.82	10.92	7.77	7.63	6.58	8.02	8.34	7.95
RCSalpEpi-MO	11.94	15.86	19.66	18.02	10.19	17.16	10.86	11.39	11.85	12.95

Figure-8. Runtime Assessment of 2-Locus DNME Models.

4.3.3. Experimental Results of 3-Locus DME Models

The power of SalpEpi-SO [19] SalpEpi-MO [19] RCSalpEpi-SO and RCSalpEpi-MO for 15 3-Locus DME models is presented in Table 2. RCsalpEpi-MO achieved 57% of power for additive model 1, which is superior to other approaches. When it came to additive model 2, the SalpEpi-SO [19] had the lowest power of 6 percent, but the RCSalpEpi-MO had the highest power of 30 percent. The RCSalpEpi-MO and SalpEpi-MO [19] yielded 69% and 62% power, respectively, for additive model 3. RCSalpEpi-MO outperformed the other three techniques with the highest power of 72 percent for additive model 4. RCSalpEpi-MO obtained power of 82% for model 5, which is 78%, 12%, 66% superior to SalpEpi-SO [19] SalpEpi-MO [19] RCSalpEpi-SO, respectively. RCSalpEpi-MO in multiplicative model 5 has the better detection accuracy of 76 percent. In threshold models, the performance of RCSalpEpi-MO is superior to others for all models. Threshold model 5 has the having maximum power of 82 percent. SalpEpi-SO [19] on the other hand, has the minimum detection performance of 1% in model 2. For threshold model 4, SalpEpi-MO yielded a power of 1% higher than RCSalpEpi-MO. Except for threshold model 4, the experimental results showed that RCSalpEpi-MO outperforms others for fourteen DME models.

Table-2. Performance examination of 3-Locus DME Method.

Model / Methods	Additive Models					Multiplicative Models					Threshold Models				
	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
RCSalpEpi-SO	27	10	53	59	16	10	6	4	3	73	10	7	67	19	18
RCSalpEpi-MO	56	30	69	72	82	21	16	8	14	76	20	11	79	74	82

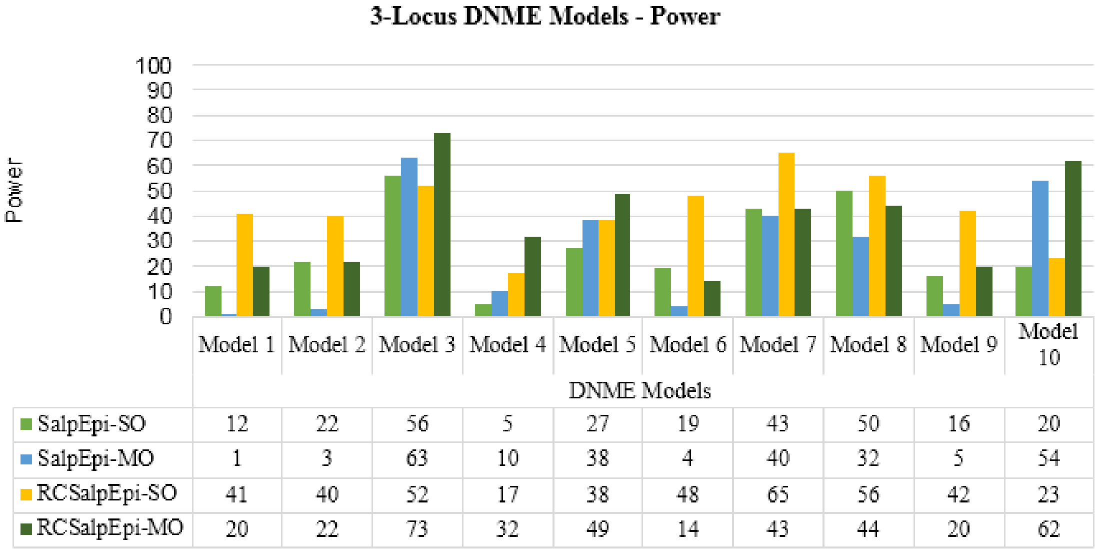
The runtime examination of 3-Locus DME models is exposed in Table 3. When compared to SalpEpi-SO [19] and SalpEpi-MO [19] the RCSalpEpi-SO and RCSalpEpi-MO had the shortest running time among the 15 DME models. For all 15 DME models, the SalpEpi-MO method has the longest execution time. The execution time of RCSalpEpi-MO is superior to SalpEpi-SO [19] SalpEpi-MO [19]. In additive model 4, SalpEpi-MO [19] requires 161 minutes, whereas RCSalpEpi-MO and RCSalpEpi-MO require 32 and 30 minutes, respectively. RCSalpEpi-MO needs 114 minutes to run the additive model 3, which is 13 minutes higher than SalpEpi-SO [19]. In multiplicative model 1, the running time of SalpEpi-MO [19] is 125 minutes which is 91 minutes, 83 minutes higher than RCSalpEpi-SO and RCSalpEpi-MO, respectively. In all the 15 3-locus DME models, SalpEpi-MO [19] and SalpEpi-SO [19] requires more execution duration compared to others. Hence, it is concluded that Rough clustering-based approaches outperform others in terms of running time.

Table-3. Running time of 3-locus DME method.

Model Methods	Additive Models					Multiplicative Models					Threshold Models				
	M1	**M2**	**M3**	**M4**	**M5**	**M1**	**M2**	**M3**	**M4**	**M5**	**M1**	**M2**	**M3**	**M4**	**M5**
SalpEpi-SO	67.3	77.9	100.7	79.1	108.8	69.7	89.1	68.3	108.6	72.8	86.6	66.4	79.9	81.4	80.0
SalpEpi-MO	107.8	102.3	135.1	161.2	113.7	125.3	95.7	101.9	111.4	110.5	94.5	106.5	113.2	153.0	146.7
RCSalpEpi-SO	36.9	33.1	34.9	30.4	31.2	34.3	34.3	39.1	43.4	29.9	34.4	33.9	36.3	27.4	35.6
RCSalpEpi-MO	66.4	47.2	42.1	31.8	30.4	42.1	52.8	42.7	46.5	31.8	36.5	44.7	49.2	22.6	36.6

4.3.4. Experimental Results of 3-Locus DNME Models

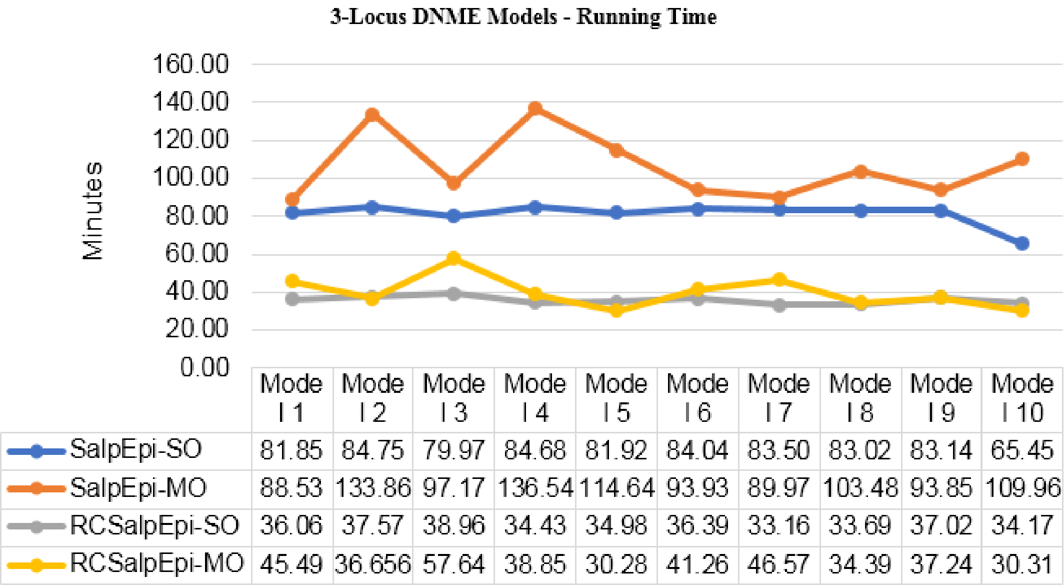
The power of SalpEpi-SO [19] SalpEpi-MO [19] RCSalpEpi-SO and RCSalpEpi-MO for 10 3-Locus DNME models is shown in Figure 9. RCSalpEpi-SO achieves the best accuracy of 73 percent in Model 3. SalpEpi-SO in Model 1 yields the minimum power of 1 percent. In Model 9, RCSalpEpi-MO yielded the power of 20% which is 22% lower than RCSalpEpi-SO. RCSalpEpi-SO outperforms the competitors in six models: Models 1-2, and Models 6-9. Similarly, RCSalpEpi-MO approach is superior to other approaches in 4 models such as Model 3, Model 4, Model 5, Model 10. The experimental outcome clearly revealed that Rough clustering based approaches is superior to SalpEpi-SO [19] and SalpEpi-MO [19] in 3-Locus DNME models.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
🟩 SalpEpi-SO	12	22	56	5	27	19	43	50	16	20
🟦 SalpEpi-MO	1	3	63	10	38	4	40	32	5	54
🟨 RCSalpEpi-SO	41	40	52	17	38	48	65	56	42	23
🟩 RCSalpEpi-MO	20	22	73	32	49	14	43	44	20	62

Figure-9. Performance Analysis of 3-Locus DNME Models.

The execution time of ten 3-Locus DNME epistasis models is presented in Figure 10. The line chart proved that the RCSalpEpi-SO and RCSalpEpi-MO requires the lowest runtime for ten DNME models. SalpEpi-MO [19] approach avail the high runtime in all the ten DNME models. The experimental analysis proved that Rough clustering-based approaches perform superior than others in connection with running time.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
SalpEpi-SO	81.85	84.75	79.97	84.68	81.92	84.04	83.50	83.02	83.14	65.45
SalpEpi-MO	88.53	133.86	97.17	136.54	114.64	93.93	89.97	103.48	93.85	109.96
RCSalpEpi-SO	36.06	37.57	38.96	34.43	34.98	36.39	33.16	33.69	37.02	34.17
RCSalpEpi-MO	45.49	36.656	57.64	38.85	30.28	41.26	46.57	34.39	37.24	30.31

Figure-10. Runtime evaluation of 3-Locus DNME Models.

5. CONCLUSIONS

The identification of possible genetic interactions is a critical and difficult problem in GWAS. In this article, we suggest a two-step approach called RCSalpEpi-SO and RCSalpEpi-MO to detect epistasis effects. RCSalp-Epi employs a rough clustering technique to partition the large genotype dataset into three clusters. During the selection step, the suggested methods are more suited for detecting higher-order SNP interactions, and they can use either of two distinct strategies: exhaustive or optimization-based search. Exhaustive search is employed on a narrow clustered dataset, whereas salp-based search is used on a huge candidate set. The proposed approach discovers high-order genetic interactions with minimal computational effort and high detection power. For both DNME and DME models, experimental findings showed that RCSalpEpi-SO and RCSalpEpi-MO are superior to SalpEpi-MO and SalpEpi-SO. The future scope of this research works may be extended to assess the real datasets for diagnosing complex diseases in human.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Acknowledgement: Both authors contributed equally to the conception and design of the study.

REFERENCES

[1] J. Shang, J. Zhang, Y. Sun, D. Liu, D. Ye, and Y. Yin, "Performance analysis of novel methods for detecting epistasis," *BMC Bioinformatics*, vol. 12, pp. 1-17, 2011. Available at: https://doi.org/10.1186/1471-2105-12-475.

[2] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, pp. 714-721, 2009. Available at: https://doi.org/10.1093/bioinformatics/btp041.

[3] D. B. Blumenthal, L. Viola, M. List, J. Baumbach, P. Tieri, and T. Kacprowski, "EpiGEN: An epistasis simulation pipeline," *Bioinformatics*, vol. 36, pp. 4957-4959, 2020. Available at: https://doi.org/10.1093/bioinformatics/btaa245.

[4] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, pp. 1167-1173, 2007. Available at: https://doi.org/10.1038/ng2110.

[5] P. M. Visscher, "Five years of GWAS discovery," *The American Journal of Human Genetics*, vol. 90, pp. 7-24, 2012. Available at: https://doi.org/10.1016/j.ajhg.2011.11.029.

[6] Genetics Home Reference, "What are single nucleotide polymorphisms (SNPs)? Retrieved from ghr.nlm.nih.gov/primer/genomicresearch/snp," 2019.

[7] T. F. Mackay and J. H. Moore, "Why epistasis is important for tackling complex human disease genetics," *Genome Medicine*, vol. 6, pp. 1-3, 2014. Available at: https://doi.org/10.1186/gm561.

[8] R. Manavalan and S. Priya, "Genetic interactions effects for cancer disease identification using computational models: A review," *Medical & Biological Engineering & Computing*, vol. 59, pp. 733-758, 2021. Available at: https://doi.org/10.1007/s11517-021-02343-9.

[9] S. Priya and R. K. Manavalan, "Genetic interactions effects of cardiovascular disorder using computa-tional models: A review," *Current Biotechnology*, vol. 9, pp. 177-191, 2020. Available at: https://doi.org/10.2174/2211550109999201008125800.

[10] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics*, vol. 69, pp. 138-147, 2001. Available at: https://doi.org/10.1086/321276.

[11] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang, and W. Yu, "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *The American Journal of Human Genetics*, vol. 87, pp. 325-340, 2010. Available at: https://doi.org/10.1016/j.ajhg.2010.07.021.

[12] L. Sun, G. Liu, L. Su, and R. Wang, "See: A novel multi-objective evolutionary algorithm for identifying SNP epistasis in genome-wide association studies," *Biotechnology & Biotechnological Equipment*, vol. 33, pp. 529–547, 2019. Available at: https://doi.org/10.1080/13102818.2019.1593052.

[13] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics*, vol. 26, pp. 30-37, 2010. Available at: https://doi.org/10.1093/bioinformatics/btp622.

[14] S. Tuo, J. Zhang, X. Yuan, Z. He, Y. Liu, and Z. Liu, "Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations," *Scientific Reports*, vol. 7, pp. 1-18, 2017. Available at: https://doi.org/10.1038/s41598-017-11064-9.

[15] P.-J. Jing and H.-B. Shen, "MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, pp. 634-641, 2012. Available at: https://doi.org/10.1093/bioinformatics/btu702.

[16] Y. Sun, J. Shang, J.-X. Liu, S. Li, and C.-H. Zheng, "epiACO-a method for identifying epistasis based on ant Colony optimization algorithm," *BioData Mining*, vol. 10, pp. 1-17, 2017. Available at: https://doi.org/10.1186/s13040-017-0143-7.

[17] X. Li, S. Zhang, and K.-C. Wong, "Nature-inspired multiobjective epistasis elucidation from genome-wide association studies," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, pp. 226-237, 2018. Available at: https://doi.org/10.1109/tcbb.2018.2849759.

[18] J. Sitarčík and M. Lucká, "EpiBAT: Multi-objective bat algorithm for detection of epistatic interactions," presented at the 2019 IEEE 15th International Scientific Conference on Informatics. IEEE, 2019.

[19] S. Priya and R. Manavalan, "Kmeans-pillar-Salpepi: Genetic interactions detection through K-means clustering with pillar and salp optimization techniques in genome-wide association studies," *Bioscience Biotechnology Research Communications*, vol. 14, 2021 (In press).

[20] P. Lingras, "Unsupervised rough set classification using GAs," *Journal of Intelligent Information Systems*, vol. 16, pp. 215–228, 2001.

[21] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Advances in Engineering Software*, vol. 114, pp. 163-191, 2017. Available at: https://doi.org/10.1016/j.advengsoft.2017.07.002.

[22] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData Mining*, vol. 5, pp. 1-14, 2012. Available at: https://doi.org/10.1186/1756-0381-5-16.

[23] Q. Chen, X. Zhang, and R. Zhang, "Privacy-preserving decision tree for epistasis detection," *Cybersecurity*, vol. 2, pp. 1-12, 2019. Available at: https://doi.org/10.1186/s42400-019-0025-z.

[24] M. Aflakparast, H. Salimi, A. Gerami, M. Dubé, S. Visweswaran, and A. Masoudi-Nejad, "Cuckoo search epistasis: A new method for exploring significant genetic interactions," *Heredity*, vol. 112, pp. 666-674, 2014. Available at: https://doi.org/10.1038/hdy.2014.4.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.