Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 4471736, 19 pages
https://doi.org/10.1155/2022/4471736

Research Article

A Secure High-Order Gene Interaction Detecting Method for Infectious Diseases

Huanhuan Wang[1,2] **Hongsheng Yin**[1] **and Xiang Wu**[2]

[1]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou221116, China*
[2]*School of Medical Information & Engineering, Xuzhou Medical University, Xuzhou 221000, China*

Correspondence should be addressed to Xiang Wu; wuxiang@xzhmu.edu.cn

Received 26 January 2022; Accepted 1 March 2022; Published 21 April 2022

Academic Editor: Shan Zhong

Infectious diseases pose a serious threat to human life, the Genome Wide Association Studies (GWAS) can analyze susceptibility genes of infectious diseases from the genetic level and carry out targeted prevention and treatment. The susceptibility genes for infectious diseases often act in combination with multiple susceptibility sites; therefore, high-order epistasis detection has become an important means. However, due to intensive computational burden and diversity of disease models, existing methods have drawbacks on low detection power, high computation cost, and preference for some types of disease models. Furthermore, these methods are exposed to repeated query and model inversion attacks in the process of iterative optimization, which may disclose Single Nucleotide Polymorphism (SNP) information associated with individual privacy. Therefore, in order to solve these problems, this paper proposed a safe harmony search algorithm for high-order gene interaction detection, termed as HS-DP. Firstly, the linear weighting method was used to integrate 5 objective functions to screen out high-order SNP sets with high correlation, including K2-Score, JS divergence, logistic regression, mutual information, and Gini. Then, based on the Differential Privacy (DP) theory, the function disturbance mechanism was introduced to protect the security of individual privacy information associated with the objective function, and we proved the rationality of the disturbance mechanism theoretically. Finally, the practicability and superiority of the algorithm were verified by experiments. Experimental results showed that the algorithm proposed in this paper could improve the detection accuracy to the greatest extent while guaranteeing privacy.

1. Introduction

The prevention and treatment of infectious diseases is an important and long-term task for human beings. The Genome Wide Association Studies (GWAS) can analyze susceptibility genes from the whole gene range and carry out targeted prevention and treatment of infectious diseases, which is of great significance for the long-term development of human beings. More and more studies showed that the interaction between genes is the main cause of genetic variation in infectious diseases [1]. Detection of gene interaction refers to the search for multiorder gene site combinations affecting diseases to determine the pathogenic cause and genetic mechanism, which has become an important research direction in Genome Wide Association Studies (GWAS) [2, 3].

Thousands of methods for identifying gene interactions have been studied, and they can be divided into exhaustive [4], random [5], filtering [6], modeling [7], and intelligent methods [8]. The exhaustive method identifies genes that interact by combining all the possibilities. The random method extracts only partial gene combinations from the data to analyze the disease model. The former method has comprehensiveness and completeness, but the calculation burden is too heavy, and the detection accuracy of the latter still needs

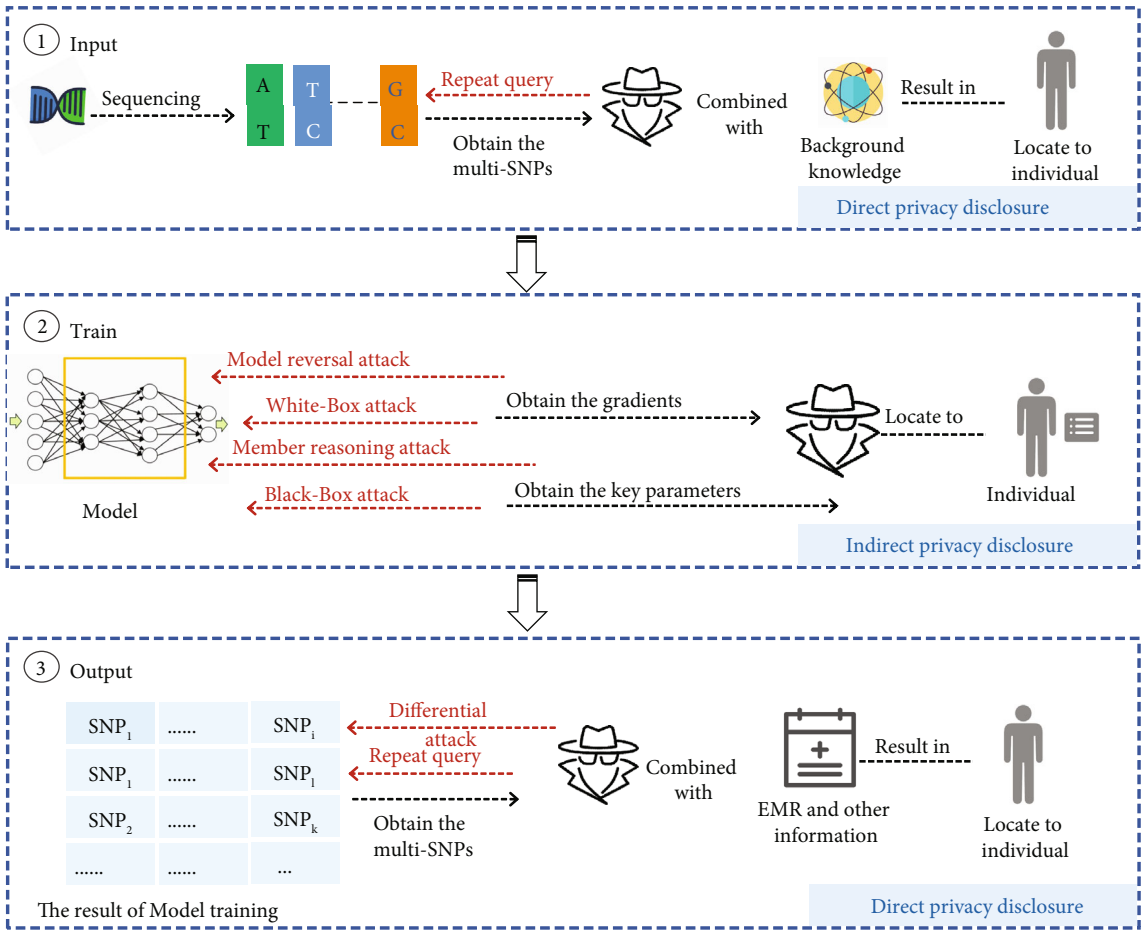


Figure 1: The privacy disclosure of high-order gene interaction detection.

to be improved. The filtering method is a combination of exhaustive and filtering method, while reducing the computational burden, still misses some important combinations of gene interactions. Modeling is based on traditional machine learning methods, which use probabilistic methods to identify gene interactions, but this method still cannot detect gene combinations with marginal effects.

The latest breakthrough for identifying gene interactions in GWAS is swarm intelligence search method, which makes full use of the information contained in current optimization parameters to generate final new search results, and has become one of the most popular methods, for example, AntEpiSeeker [9], IEACO [10], FHSA-SED [11], IOBLPSO [12], DECMDR [13], and other methods.

Although swarm intelligence search method has the advantages of controllable time complexity compared with other methods, it still faces many serious problems. First, the evaluation function design of swarm intelligence method is unreasonable [14]. The evaluation function is not selected from multiple dimensions, so the auxiliary optimization search strategy has defects. Second, the accuracy of identifying the results of gene interactions still needs to be improved [15]. Last but not least, there are privacy risks associated with identifying multiorder genetic interactions [16, 17]. As shown in Figure 1, there are risks of privacy disclosure in the input, training, and output stages of genetic data. In the data input stage, the untrusted third party launches repeated query attack to obtain the original data information for many times and locate the individuals by combining these information with the background knowledge. In the model training stage, the attacker can obtain the gradient and some key parameters directly related to the original data through various means such as model inversion, so as to mine more personal information based on the background knowledge. This process is an indirect privacy breach. In the data output stage, the untrusted third party obtains more genetic detection result information by differential privacy budget attack. Combining this information with a genetic history of medical visits for certain diseases can target individuals. Therefore, the privacy security of genetic data needs to be solved urgently.

Therefore, in order to solve the above problems, this paper proposed a safe Harmony Search (HS) algorithm for high-order gene interaction detection. Firstly, the linear weighting method was used to integrate various objective functions, including K2-Score, JS divergence, logistic regression, mutual information, and Gini, to screen out the Single Nucleotide Polymorphism (SNP) solution set with high correlation. Then, to protect the privacy security of sensitive information, we introduced the function disturbance mechanism and analyzed the rationality of this mechanism. Specific contributions are as follows:

(i) Select fitness functions from different dimensions to overcome the difficulty of poor detection results caused by gene interaction among functions in the same gradient. Use the linear weighted sum method to combine these functions in the same gradient direction for comprehensive evaluation of the identified multiorder gene combinations

(ii) Propose a privacy protection mechanism to solve the problem of privacy disclosure in high-order gene interaction identification. Specifically, this mechanism interferes with the polynomial coefficients of the objective optimization function rather than simply adding noise to the result, achieving the balance between privacy and utility

(iii) Theoretical and experimental results show that HSDP can not only identify the accuracy of gene interaction but also protect the security of SNP information associated with individual privacy

The remainder of the paper is organized as follows. Section 2 overviews the related works. List the preliminaries of this paper in Section 3. In Section 4, we introduce our algorithm in detail. The experimental evaluations and results are discussed in Section 5. Finally, Section 6 summarizes the paper.

2. Related Work

Gene interaction refers to the influence of the interaction between two or more single nucleotide polymorphisms (or genes) on the phenotype. Epistasis is one of the important genetic factors affecting complex diseases. The swarm intelligence search algorithm is widely used in the interactive recognition of SNPs because it can efficiently and quickly search for feasible solutions within a given range. Commonly used swarm intelligence search algorithms include particle swarm optimization algorithm, differential evolution algorithm, artificial bee colony algorithm, and ant colony optimization algorithm.

In the study of Particle Swarm Optimization (PSO), Yang et al. [18] proposed an Odds Ratio-Based binary Particle Swarm Optimization (OR-BPSO) method to evaluate the risk of breast cancer. BPSO provides the combinational SNPs with their corresponding genotype, called SNP barcodes, with the maximal difference of occurrence between the control and breast cancer groups. Chuang et al. [19] proposed IPSO algorithm to improve the reliability of PSO for the identification of the best protective SNP barcodes associated with breast cancer. The top five SNP barcode results are retained for computing the next SNP barcode with a one-SNP increase for each processing step. Shang et al. [12] proposed an improved opposition-based learning particle swarm optimization for the detection of SNP-SNP interactions. The proposed algorithm enhances the global exploration capability and also avoids premature convergence. The particles cover a wider search space and perform in-depth search on highly suspicious SNP sets. Chuang et al. [20] applied an evolutionary algorithm to facilitate statistical methods in the analysis of associated variations for disease susceptibility. The Gauss particle swarm optimization algorithm was used to detect and identify the best protective association (i.e., combinations of SNPs with genotypes) with breast cancer. Yang et al. [21] proposed a PSO-Based Multifactor Dimensionality Reduction approach (PBMDR). MDR was used to detect multilocus interactions based on the PSO algorithm. Chuang et al. [22] proposed to combine chaotic graphs with PSO methods to detect the interaction of SNPs in high-dimensional datasets, where chaotic graphs help the PSO algorithm to avoid falling into local optima.

In the study of Differential Evolution (DE). Yang et al. [13] proposed a new algorithm which combines the DE algorithm with a Classification-based Multifactor Dimensionality Reduction (CMDR), termed DECMDR. Yang et al. [23] proposed a catfish Taguchi-based binary differential evolution (CT-BDE) algorithm for identifying SNP-SNP interactions. In the search space, the catfish effect prevents the premature convergence of the population, and the Taguchi method improves the search ability of the BDE algorithm. Guan et al. [24] proposed a fast evolutionary optimization method named search-history-guided differential evolution. This method applies the search history memorized in a binary space partitioning tree to enhance its power for selecting feature combinations. Guan et al. [25] proposed a two-stage algorithm DEseeker to detect epistatic effects. This scheme can identify hidden SNPs, but it takes too much time to execute in large-scale datasets.

In the study of Artificial Bee Colony (ABC). Yang et al. [26] proposed a method of superiority mining based on the artificial bee colony algorithm to optimize the Bayesian network. The algorithm is applied to the Bayesian network heuristic search strategy. Li et al. [27] proposed and formulated a decomposition-based upper interactive multiobjective artificial bee colony algorithm. Two objective functions are formulated to characterize various upper models and a rank probability model based on the fast nondominated ranking method is proposed. After that, a local search algorithm based on mutual information was proposed.

In the study of Ant Colony Optimization (ACO). Wang et al. [9] proposed a new tool for discovering apparent interactions in large-scale case control studies, which uses a two-stage optimization program. Moreover, Wang et al. [28] developed AntEpiSeeker2.0. By looking at pheromone distribution across pathways, epistasis-associated pathways can be easily identified. Sinnott-Armstrong et al. [29] implemented ACO MDR on the GPU. The performance advantages of GPUs combined with the computational efficiency of heuristic evolutionary algorithms can solve larger-scale problems. Li et al. [30] proposed a novel approach which could find a gene-gene interaction model consists of a flexible number of susceptible loci based on ACO strategy. The proposed method becomes a potential solution for finding the complex association rules between susceptible SNP subsets and common human diseases in the future. Shang et al. [31] introduced an algorithm based on ACO, which by incorporating heuristic information into ant decision rules. Introduce heuristic information in the search process, and perform a chi-square test during the iteration. When the iterative process is completed, sort and use postprocedures to filter. Jing and Shen [32] proposed a multiobjective

ACO algorithm to detect genetic interactions, which combines the standard logistic regression and Bayesian network methods, and also design a memory-based multiobjective ACO algorithm. Liu et al. [33] proposed a flexible two-stage method (called HiSeeker) to detect high-level interactions. In the screening phase, HiSeeker uses chi-square test and logistic regression model. In the search phase, exhaustive search and search based on ant colony optimization are used. HiSeeker can detect high-level interactions more efficiently and effectively. Sapin et al. [34] introduced an ACO-based algorithm called to identify all possible binding sites of transcription factors from upstream of coexpressed genes, this algorithm uses the powerful optimization capabilities of ACO, which can not only improve the accuracy of the results but also achieve very high speeds. Sun et al. [35] proposed an algorithm based on ACO and a new fitness function value, which combines Bayesian networks and mutual information to detect SNP-SNP interactions. Guan et al. [36] proposed a new ACO algorithm based on automatic adjustment mechanism to solve the problem of combinatorial explosion of stratum by mining apparent interaction from large-scale data. The mechanism automatically adjusts the behavior of artificial ants based on real-time feedback information, so that the algorithm can run to the best state. Guan improved and proposed SEPACO [37], a self-evolved parameter based on ACO algorithm.

3. Preliminaries and Backgrounds

In this section, we will introduce the related concepts and background knowledge of the HS algorithm and the Differential Privacy (DP) mechanism.

3.1. HS Algorithm. The HS algorithm is inspired by the music-making process of jazz musicians, who improvise the pitches of their instruments in search of perfect harmony [38]. It is a group-based metaheuristic algorithm, whose idea is to realize the cognition of unknown complex problems through information exchange and learning among individuals in a group. More precisely, harmony and its tonal set are analogous to the candidate solution and its decision variable set $X = (x_1, x_2, \cdots x_N)$, respectively. In addition, the measurement of harmony's pleasurable state by the audience's aesthetic evaluation corresponds to an objective function $f(\cdot)$. Each attempt by a musician to progressively improve harmony by producing some new pitch corresponds to the application of search operators that change the value of some decision variable during each iteration of the harmony search. The musician's memory is where good harmony is stored, similar to the solution, called *HM* in the harmony algorithm.

$$HM = \left[\begin{array}{c|c} X^1 & f(X^1) \\ X^2 & f(X^2) \\ \cdots & \cdots \\ X^{HMS} & f(X^{HMS}) \end{array}\right] = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_N^1 & f(X^1) \\ x_1^2 & x_2^2 & \cdots & x_N^2 & f(X^2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^{HMS} & x_2^{HMS} & \cdots & x_N^{HMS} & f(X^{HMS}) \end{array}\right]. \tag{1}$$

HS algorithm has been used in gene interaction recognition. Tuo's team pioneered the FHSA-SED approach [11], which is an integrated HS and local search approach for identifying 2-order gene interaction recognition studies. In this method, K2 score and Gini score were selected to represent SNP loci, and local search algorithm with two-dimensional Tabu table was used to screen some disease models with strong epistatic effects. Although FHSA-SED has higher detection accuracy than other intelligent algorithms, its rationality still needs to be verified. In 2017, this team proposed validation in a AMD, to demonstrate the rationality and research ability of HS algorithms for identifying gene interactions [39]. However, the experimental results show that the accuracy of this method still needs to be improved. Recently, the team expanded the research content and proposed MP-HS-DHSI algorithm [40]. A new evaluation standard of harmony target is designed and G-test is integrated as verification method. In general, this method greatly improves the detection accuracy but ignores the data security problem, and the comprehensiveness of the objective evaluation function still needs to be considered. On this basis, we propose a safe HS algorithm for high-order gene interaction detection.

3.2. Differential Privacy. DP is a set of mechanisms developed for data analysis on sensitive data. By obfuscating database query results, the privacy of data at the personal level is realized and the query results are approximately correct. Before we introduce the definition of differential privacy, let us first agree on some symbols.

Define dataset as the D, each row of the data takes a value in this set. The number of rows in our fixed database is n, then, a database x is an element in the power set D^n (the set composed of all subsets of D. The query is defined as a function q, enters the database x, and outputs certain values. $M(x)$ is the random mechanism attached to the query q, Enter a database x to get a randomized query result. x is the database random variable, X_i represents the content of the i-th row of the database. p is the mapping defined on the database row. Pr is the probability distribution.

Definition 1. (ε-Differential Privacy) [41]. If for each pair of databases x and $x^{'}$ with only one row that are not the same, and the output y of each possible $M(x)$, all satisfy

$$\Pr\left[M(x) = y\right] \leq e^{\varepsilon} \Pr\left[M\left(X'\right) = y\right], \tag{2}$$

where $\varepsilon > 0$, and δ represents the event that the ratio of the probabilities for two adjacent datasets x, y cannot be bounded by e^{ε} after adding a privacy preserving mechanism. With an arbitrarily given δ, a privacy preserving mechanism with a larger ε gives a clearer distinguish ability of neighboring datasets and hence a higher risk of privacy violation.

After introducing the definition of difference, we will introduce three different noising mechanisms for differential privacy.

3.2.1. Laplace Mechanism. The Laplace mechanism is a mechanism that satisfies differential privacy and is mainly used in counting queries. And, the query returns a vector of nonnegative integers, and only the case where the query returns a nonnegative integer is considered here. We now define a quantity called l_1 sensitivity.

Definition 2. (Sensitivity) [42]. The l_1 sensitivity of a function $f : D^n \longrightarrow R^k$.

$$\Delta f = \max_{\substack{x,y \in D^n \\ \|x-y\|_1=1}} \|f(x) - f(y)\|_{1,}. \tag{3}$$

where Δf means single record changes the output of f, how much can be changed at most. Δf can be used to control the amplitude of noise.

Define the Laplace mechanism for any function $f : D^n \longrightarrow R^k$ as

$$M_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \cdots, Y)_k, \tag{4}$$

where Y_i is an independent and identically distributed (iid) random variable sampled from $Y_i \sim Lap(\Delta f/\varepsilon)$.

3.2.2. Gaussian Mechanism. The Gaussian mechanism mainly provides differential privacy protection for numerical data. The Laplace mechanism provides a strict $(\varepsilon, 0) - DP$, while the Gaussian mechanism provides a relaxed $(\varepsilon, \delta) - DP$ mechanism.

For any $\delta \in (0, 1), \sigma > \sqrt{2 \ln (1.25/\delta)}\Delta f/\varepsilon$, noisy $Y \sim N(0, \sigma^2)$ satisfies $(\varepsilon, \delta) - DP$.

$$P(M(x) \in S) \leq e^{\varepsilon} P(M(y) \in S) + \delta, \tag{5}$$

where $M(x) = f(x) + Y$, there are three main parameters here. The standard deviation of the Gaussian distribution σ, which determines the scale of the noise. ε indicates the privacy budget, which is negatively correlated with noise. δ represents the relaxation term. For example, if it is set to 10^{-5}, it means that the probability of 10^{-5} can only be tolerated, which violates strict differential privacy.

3.2.3. Exponential Mechanism. Exponential mechanism is used for differential privacy protection of nonnumerical data. The overall idea of the exponential mechanism is that when a query is received, it does not output a O_i result exactly but returns the result with a certain probability value, thereby achieving differential privacy. And this probability value is determined by the scoring function, the output probability of the high score is high, and the output probability of the low score is low.

Definition 3. (Utility function sensitivity). Utility function $u : D^n \times O \longrightarrow \mathbb{R}$, mapping (database-query output) to utility score, we define the sensitivity of the utility function u as

$$\Delta u \equiv \max_{r \in R} \max_{x,y:\|x-y\|_1 \leq 1} |u(x, o) - u(y, o)|. \tag{6}$$

Exponential mechanism $M_E(x, u, O)$. The probability of selecting and outputting a result $o \in O$ is proportional to $\exp(\varepsilon u(x, o)/2\Delta u)$. But $\exp(\varepsilon u(x, o)/2\Delta u)$ does not express the probability value, so it is necessary to normalize all possible values to get the corresponding probability value.

$$\Pr[O_i] = \frac{\exp(\varepsilon(D, O_i)/2\Delta u)}{\sum_j \exp(\varepsilon(D, O_j)/2\Delta u)}. \tag{7}$$

Finally, choose an output with a higher utility score with a higher probability.

4. Our Proposed HS-DP Scheme

Current studies on gene interactions are still focused on identifying 2-order gene interactions, but lack of higher-order gene interactions. In addition, few researchers have focused on the security of gene interaction GWAS based. In order to solve the current problems, a framework for high-order gene interaction detection HS-DP based on secure harmony search is proposed in this paper. This framework provides privacy guarantee for high-order gene interactions, not only effectively preserves privacy information in training data but also ensures the availability of the framework through adaptive functional perturbation mechanism. As shown in Figure 2, HS-DP mainly consists of 7 steps, including standardized data input, data quality control, multiobjective memory seize design, linear combination, differential perturbation, verification, and outputting results. Some of the key steps are detailed below. In addition, the definition of high-order gene interaction and specific problems will also be introduced in the following.

4.1. High-Order Gene Interaction. Studies have shown that almost no phenotypic characteristics of an individual are determined by a single gene, so gene-gene (or gene-environment interaction) has important theoretical and practical significance in explaining individual characteristics. High-order gene interaction is defined as the combinations of at least K SNPs affecting phenotype or disease genes. We expressed the gene interactions process as $R = \{S, G, A\}$, where $S = \{S_1, S_2 \cdots, S_i\}$ represented SNP typing, $G = \{G_{11}, G_{12}, \cdots, G_{ij}\}$ represented interaction between G_i and G_j corresponding genes, and $A = \{A_1, A_2 \cdots, A_i\}$ represented association results. The K-order gene interaction represents the recognition of SNP interaction results of the order of 3^n. Among them, when $G_{mn} > \theta$, G_{mn} is called the result with the main effect, and when $G_{ab} < \theta$, G_{ab} is the result of the edge effect.

4.2. Problem Statement. Let the set of gene variables $X = \{X_1, X_2, \cdots, X_i\}$ includes $S = \{S_1, S_2, \cdots, S_j\}$ SNP marker for N individuals. For high-order gene interaction detection algorithms, the temporal $O(f(n))$ and spatial $S(n)$ complexity of the algorithm increases exponentially in 3^n detection demand. There are three ways in which neural network training data may reveal genetic privacy. In the data input phase, one attacker A initiates $AK = \{AK_1, AK_2, \cdots, AK_n\}$ attacks, including repeated query, to obtain the original SNP data information $I_1, I_2, \cdots I_n$ for several times and locate individuals based on the background information $KN = \{$

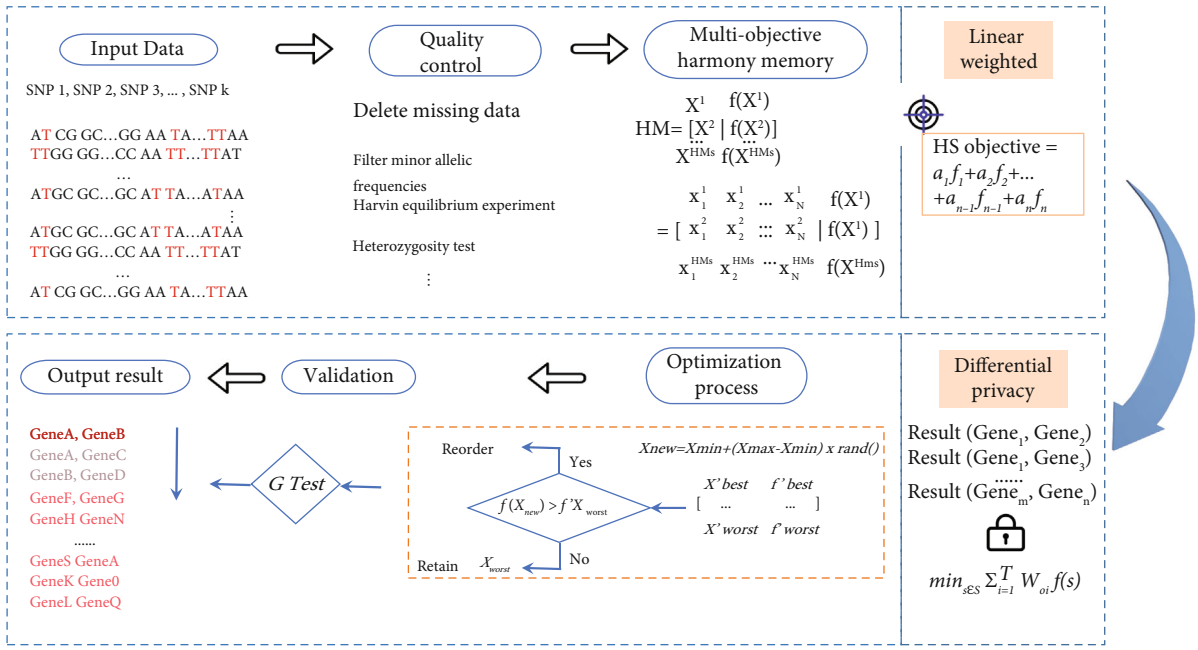


Figure 2: HS-DP overview.

$KN_1, KN_2, \cdots, KN_n\}$. In the model training stage, A can obtain gradients and some key parameters $\{\alpha, \theta, \cdots, \mu\}$ directly related to the original data through $AK = \{AK_1, AK_2, \cdots, AK_n\}$ such as model inversion, so as to mine more personal information based on $KN = \{KN_1, KN_2, \cdots, KN_n\}$. In the data output stage, A obtains more genetic detecting results information through $AK = \{AK_1, AK_2, \cdots, AK_n\}$ such as differential privacy budget. Combining this information with the genetic history of visits for certain diseases can identify individuals.

4.2.1. K2-Score. Initially, the *Bayesian network* (*BN*) is a graphical statistical model that represents the dependence of some random variables, and its reasoning model by directed acyclic graph, where nodes denote random variables and edges denote dependence between two link nodes. *BN* is also used to identify gene interaction among SNP, which calculates the association of variants and genetic genotype. In general, there are directed links from SNP M_i to genotype status T when M_i associates with T. From more than 20 kinds of *BN* models [40], HS-DP selects *K2-score* to assess the effects of SNP combinations and genotype. More details about *K2-score* based on *Bayesian Network* are introduced in literature [40]. The *K2-score* function can be written as Equation (8). In a word, lower *K2-score* value shows more stronger association between SNP combinations and the genetic phenotype.

$$k2 - Score = \prod_{n=1}^{N}\left(\frac{(I-1)!}{(M_i+I-1)!}\prod_{i=1}^{i} M_{\text{in}}!\right). \tag{8}$$

4.2.2. JS Divergence. JS divergence, derived from Kullback-Leibler (KL) divergence [43], refers to the metric of symmetry between two probability distributions.

In GWAS, JS divergence can be used to measure SNP genotype deviation between case data and control data. For a SNP combination, the genotype distribution of case and control was set as ρ_{case} and ρ_{control}, respectively. JS divergence between ρ_{case} and ρ_{control} can be expressed by Equation (9).

$$JS = 0.5\left(\sum_{i=1}^{I}\sum_{j=1}^{2}\frac{n_{ij}}{n_i}\log\frac{2n_{ij}}{ni}\right), \tag{9}$$

where ρ_{case} and ρ_{control} represent the ratio of the i-th genotype combination in the case and control samples, respectively. In general, when looking for gene interaction, the larger the JS divergence value is, the greater the difference between the genotype of the case group and the control group is, and the stronger the association between SNP combination and disease status is, that is, the gene pair has epistatic effect.

4.2.3. Logistic Regression. Logistic regression method is often used to identify the interactions with SNPs with strong epistatic effects [44]. Let M_1 and M_2 denote two SNPs and Y is the result of gene interactions. In two-order SNP M_i and M_j, HS-DP adopts a logistic regression model to identify the association between (M_i and M_j) and disease status D (1 for yes and 0 for no) as follows:

$$\log\left(\frac{P(D=1\mid(M_i,M_j))}{P(D=0\mid(M_i,M_j))}\right) = \alpha_0 + \alpha_i M_i + \alpha_j M_j + dM_iM_j, \tag{10}$$

where α_i and α_j are the main effects for SNP M_i and M_j, respectively. Using the Newton-Raphson method to search the optimization value $\hat{LF}$ of the maximum likelihood of Equation (10) iteratively.

4.2.4. Mutual Information. HS-DP fits mutual information to identify which combinations are true gene interactions. Mutual information has become one of the widely used functions for measuring the correlation of two variables [45]; thus, the formula can be written as

$$MI(M;T) = H(M) + H(T) - H(M,T), \tag{11}$$

in which $H(M)$ is the entropy of M, $H(T)$ is the entropy of T. $H(M,T)$ is the joint entropy of M and T. M is the position of a variant that is a SNP combination, and T is the genetic phenotype.

The definition of entropy and the joint entropy can be written as

$$H(M) = -\sum_{i_1=1}^{3} \cdots \sum_{i_n=1}^{3} \left(p\left(M_{i_1}, \cdots, M_{i_n}\right) \cdot \log p\left(M_{i_1}, \cdots, M_{i_n}\right)\right), \tag{12}$$

$$H(T) = -\sum_{i=0}^{1} \left(p(t_i) \cdot \log p(t_i)\right), \tag{13}$$

$$H(M,T) = -\sum_{i_1=1}^{3} \cdots \sum_{i_n=1}^{3} \sum_{i=0}^{1} \left(p\left(M_{i_1}, \cdots, M_{i_n}, t_i\right) \cdot \log p\left(M_{i_1}, \cdots, M_{i_n}, t_i\right)\right), \tag{14}$$

where n is the number of SNPs in SNP combinations, and t is the label of samples, and then, p represents the probability distribution function. In general, higher mutual information value shows more stronger association between SNP combinations and the genetic phenotype.

4.2.5. Gini Score. The Gini index is a measure of dispersion that can be used to measure the impure nature of data partitions or the inequality between values of frequency distributions [46]. The correlation problem of gene interactions is essentially a dichotomous problem and can therefore be measured by the Gini coefficient. The Gini index is a diversity index, specifically defined as

$$\text{Gini} = \sum_{i=1}^{I} \rho_i \cdot \left(1 - \sum_{j=1}^{J} \rho_{i,j}^2\right), \tag{15}$$

where $\rho_{i,j}(\rho_{i,j} = n_{ij}/n_i)$ is the estimated probability that the i-th genotype combination is actually associated with phenotype y_j. $(1 - \sum_{j=1}^{J} \rho_{i,j}^2)$ represents the estimated probability of genotype combinations being misclassified as phenotypic y_j. $\rho_i(\rho_i = n_i/L)$ is the percentage of the i-th genotype combination in the sample set. The smaller the Gini coefficient is, the stronger the correlation between SNP combination and phenotype is, that is, the gene has epistasis.

4.3. Functional Differential Perturbation. HS-DP utilized the linear weighted sum method to perform the above multiple fitness assessment functions. Weight allocation is one of the most important steps in the composition process. In this study, formula (16) is used to assign the appropriate weight, and the calculation process is as follows:

$$\text{Power} = \frac{\text{True SNPs}}{\sum_{n=1} \text{SNPs}}, \tag{16}$$

$$\min_{s \in S} \sum_{i=1}^{T} W_{o_i} f(s), \tag{17}$$

$$W_i = \frac{\text{Power}_i}{\sum_{i=1} \text{Power}}, \ \sum_{i}^{m} W_i = 1, \tag{18}$$

where equation (17) is the objective optimization function of HS-DP; however, literature [47] had confirmed that the objective function directly related to the original data in optimization problems will leak data privacy information. The objective function of HS-DP is not only directly related to the original data, but also its results are directly related to privacy information. On this basis, this paper proposed a general differential privacy function perturbation framework for the study of high-order gene interactions in GWAS to solve the privacy leakage problem.

Before introducing the specific content of this method, we would first introduce the preliminaries. Let O be a set of n objectives $\{O_1, O_2, \cdots, O_n\}$ and i genes. For each goal $O = (O_{i1}, O_{i2}, \cdots, O_{id}, f_i)$, we assume no loss of generality, $f_i \geq 0$. Our goal is to build an O-based regression model (also known as the objective function F) that takes the predictions of F as inputs and outputs $S_1, S_2, \cdots, S_M$. HS-DP shows that F is a linear regression model parameterized by W, and W is an N-dimensional vector, where the number of j-th $(j \in \{1, 2, \cdots, n\})$ is equal to the weight of f_i in F. To evaluate the accuracy of W, we define a cost function W^* with O_i and W as inputs. According to the definition of linear regression cost function, the equation of parameter W^* is as follows:

$$W^* = \arg \min_{W} \sum_{n=1}^{i} F(f_i, W). \tag{19}$$

Definition 4. (ϵ-differential privacy). The randomized algorithm A satisfies ϵ-differential privacy, if for any output R of A and for any two neighbor databases, we have

$$\Pr\left[A(D_1) = R\right] \leq e^{\int} \cdot \Pr\left[A(D_2) = R\right]. \tag{20}$$

The regression task for genetic data returns the parameter W^* that minimizes the objective optimization function $F = \min_{s \in S} \sum_{i=1}^{T} W_{o_i} f(s)$. Releasing W^* directly would compromise the privacy of information that reveals gene data. In general, our method perturbs and optimizes the function objective to protect the analysis results rather than directly perturb the regression results. However, the key issue is how to protect differentiated private information. According to the Stone-Weierstrass Theorem [47], we use the polynomial of F as follows.

$$\psi_i = \left\{ W_1^{S_1} W_2^{S_2} \cdots W_n^{S_n} \mid \sum_{m}^{n} S_m = i \right\}, \tag{21}$$

Input: D: SNP dataset
k: the number of SNPs in the combinations.
ϵ: Privacy budget
F: Objective function
HMCR:
PAR:
MaxFEs: the number of iterations
Step 1. Initialize the key parameters of HS algorithm.
Step 2. Calculate the five value of each harmony X by linear weighted sum method.
Step 3. Apply the functional differential perturbation.
for each $0 \le i$**do**
for each $\varphi \in \psi_i$**do**
set $\eta_\varphi = \eta_\varphi + Laplace$
End for
End for
Set $\Delta = 2 \max_S \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \| \eta_{\varphi_{f_i}} \|_1$
Let $F = \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \eta_{\varphi_{f_i}} \psi(W)$
Compute $W^* = \arg \min_W \sum_{n=1}^{i} F$
Step 4. Update the value of harmonies by ranking and iterating.
Step 5. G-test stage.
Calculate G test for each SNP subset left in step 1.
Output: the set of k-order gene combinations that have a strong association with disease model.

Algorithm 1: HS-DP.

where ψ_i denotes the set of all products of $W_1, \cdots, W_n$. Therefore, the optimization function F is formula (22).

$$\sum_{i=0}^{I} \sum_{\varphi \in \psi_i} \eta_{\varphi_{f_i}} \psi(W), \tag{22}$$

where $\eta_{\varphi_{f_i}}$ is the coefficient of $\psi(W)$. The function perturbation mechanism proposed by HS-DP injects noise into this polynomial coefficient and then obtains the model parameter W of the optimized function F, as shown in algorithm 1.

Theorem 1. *Let G and G' be any two adjacent datasets of genes (Assume G and G' differ in the last tuple), and F and F' be the objective functions of regression analysis of G and G'. The polynomial are $F = \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \sum_{G_{i \in G}} \eta_{\varphi_{f_i}} \psi(W)$ and $F' = \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \sum_{G_{i \in G'}} \eta_{\varphi_{f'_i}} \psi(W)$, respectively. Then, the inequality is $\sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \| \sum_{G_{i \in D}} \eta_{\varphi_{f_i}} - \sum_{G'_{i \in D}} \eta_{\varphi_{f'i}} \|_1 \le 2 \max_f \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \| \eta_{\varphi_{f_i}} \|_1$.*

Proof. Algorithm 1 satisfies ϵ-differential privacy.

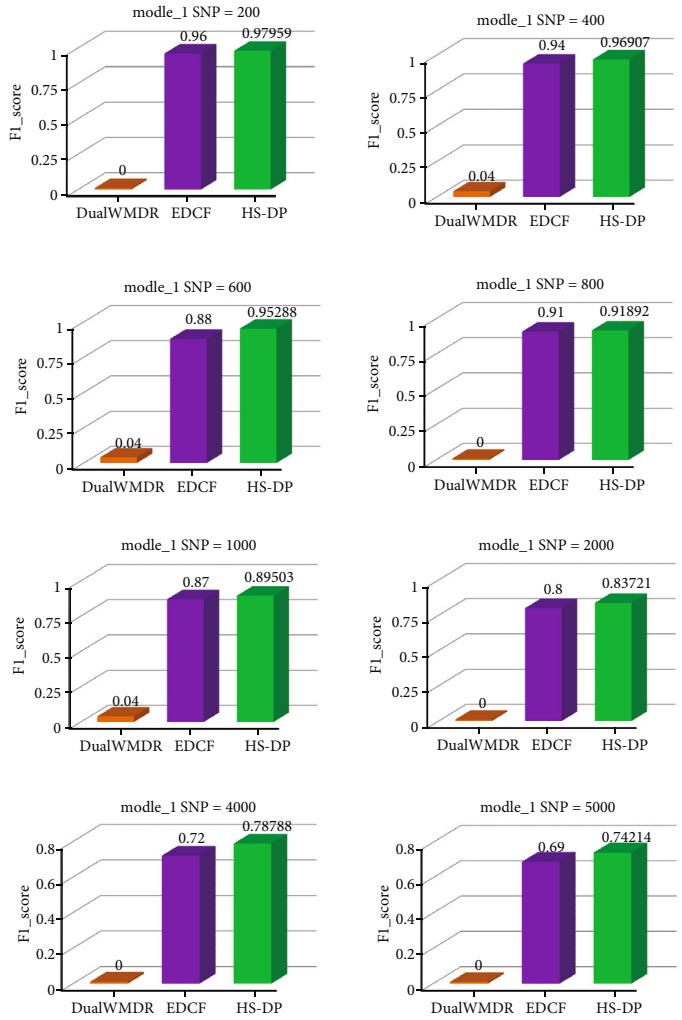
$$\begin{aligned}
\frac{\mathrm{pr}[F \mid G]}{\mathrm{pr}\left[F \mid G'\right]} &= \frac{\prod_{i=1}^{I} \prod_{\varphi \in \psi_i} \exp\left(\varepsilon \cdot \left\| \sum_{f_{i \in G}} \eta_{\varphi_{f_i}} - \eta_\varphi \right\|_1 / \Delta \right.}{\prod_{i=1}^{I} \prod_{\varphi \in \psi_i} \exp\left(\varepsilon \cdot \left\| \sum_{f'_{i \in G'}} \eta_{\varphi_{f'_i}} - \eta_\varphi \right\|_1 / \Delta \right.} \\
&\le \prod_{i=1}^{I} \prod_{\varphi \in \psi_i} \exp\left(\frac{\varepsilon}{\Delta} \cdot \left\| \sum_{f_i \in G} \eta_{\varphi_{f_i}} - \sum_{f'_i \in G'} \eta_{\varphi_{f'_i}} \right\|_1\right), \\
&= \prod_{i=1}^{I} \prod_{\varphi \in \psi_i} \exp\left(\frac{\varepsilon}{\Delta} \cdot \left\| \eta_{\varphi_{G_n}} - \eta_{\varphi_{G'_n}} \right\|_1\right), \\
&= \exp\left(\frac{\varepsilon}{\Delta} \cdot \prod_{i=1}^{I} \prod_{\varphi \in \psi_i} \left\| \eta_{\varphi_{G_n}} - \eta_{\varphi_{G'_n}} \right\|_1\right), \\
&\le \exp\left(\frac{\varepsilon}{\Delta} \cdot 2 \max_G \sum_{i=1}^{I} \sum_{\varphi \in \psi_i} \left\| \eta_{\varphi f_i} \right\|_1\right), \\
&= \exp(\varepsilon).
\end{aligned} \tag{23}$$

□

4.4. G-Test. At the above stage, multiorder gene combinations strongly associated with disease or phenotypic status have been screened out. At this stage, G-test statistical method [48] was used to verify the significance level of candidate high-order gene combinations.

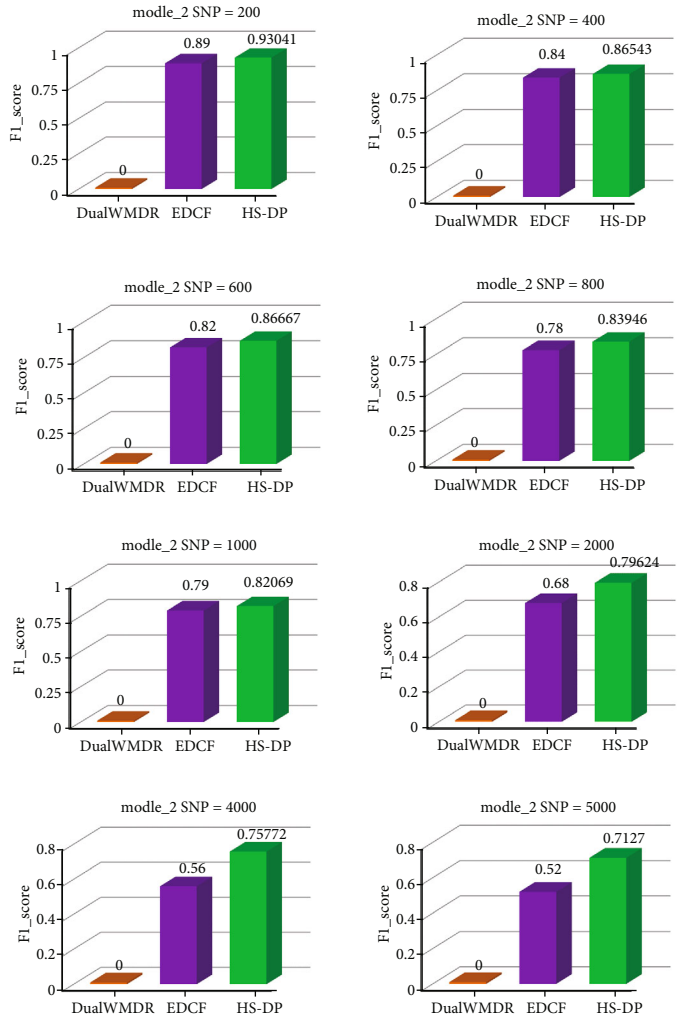
G-test is a logarithmic likelihood ratio test, and X^2 test is the approximation of the second order Taylor expansion of logarithmic likelihood ratio test. It can be understood that the G-test is more accurate than the X^2 test in some scenarios. Logarithmic likelihood ratio statistics are difficult to calculate, so the X^2 test is widely used. But the G-test is now more widely used when computational power is sufficient. In this paper, we redefined the calculation process of G-test for gene interaction in GWAS, as follows:

$$G = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} P_{ij}, \tag{24}$$



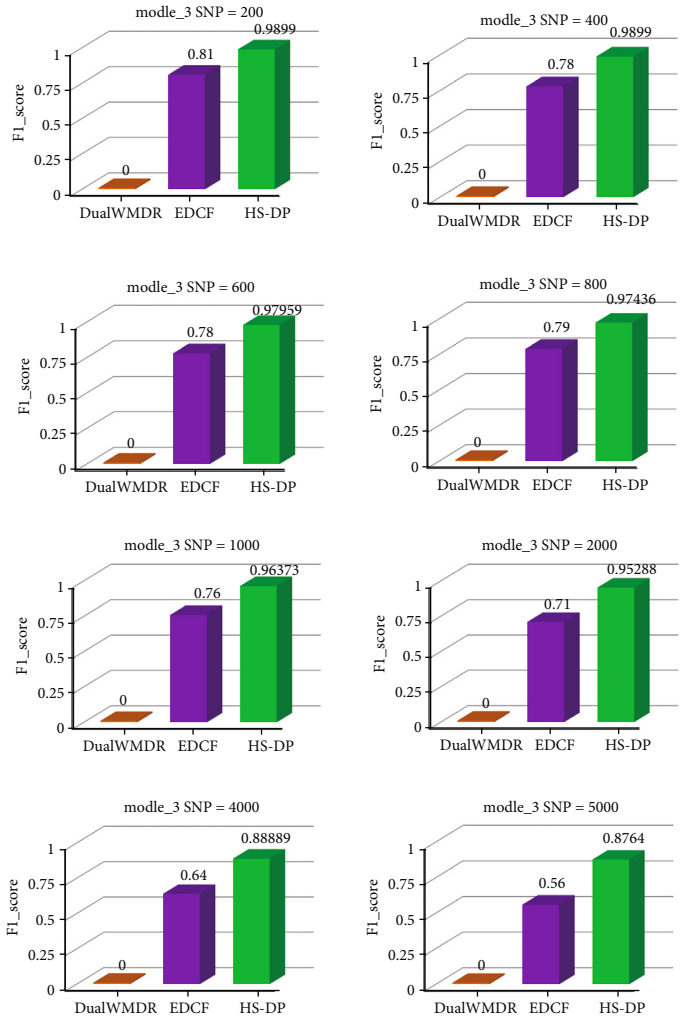
(a)

Figure 3: Continued.



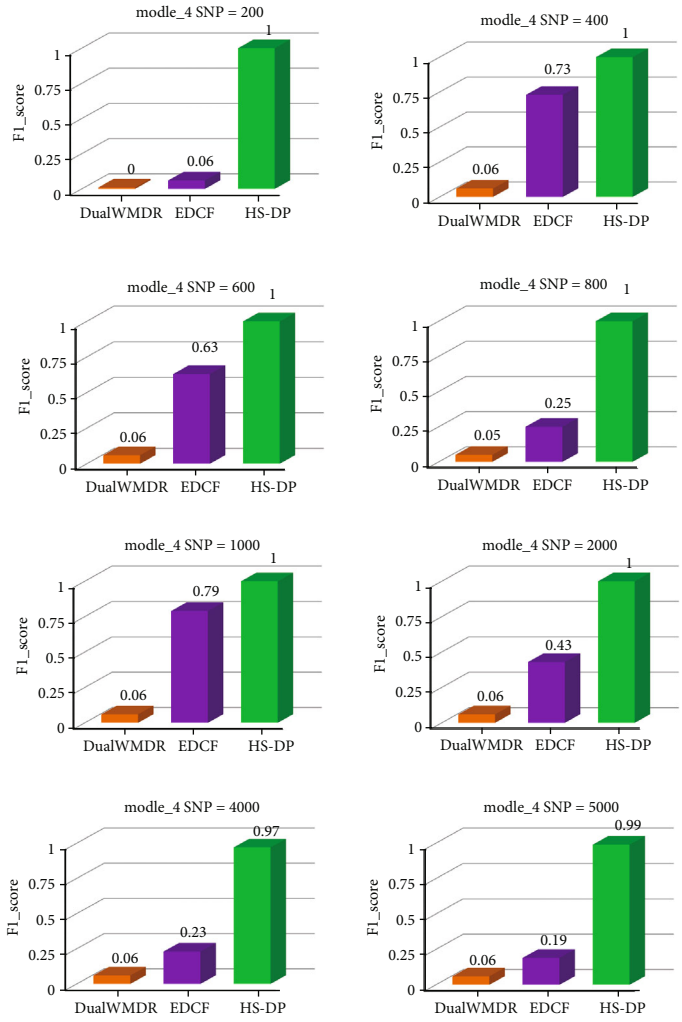
(b)

Figure 3: Continued.



(c)

Figure 3: Continued.



(d)

Figure 3: The accuracy of with-marginal effect models.

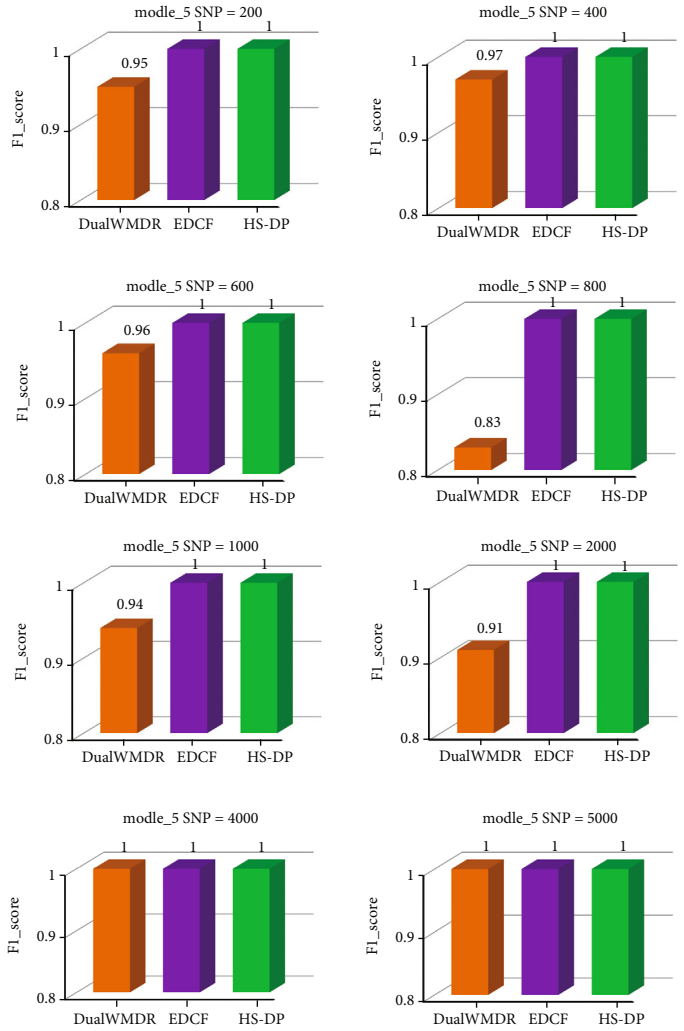
$$P_{ij} = \begin{cases} a \ln \dfrac{O_{ij}}{e_{ij}}, \sum_{j=1}^{J} O_{ij} > \zeta \\ 0, otherwise \end{cases}, \tag{25}$$

where O_{ij} is the observed number of genotype I when the disease state is y_j, e_{ij} is the corresponding expected number of genotype I when the disease state is y_j, which can be calculated according to the Hardy-Weinberg principle [49].

5. Experimental Analysis

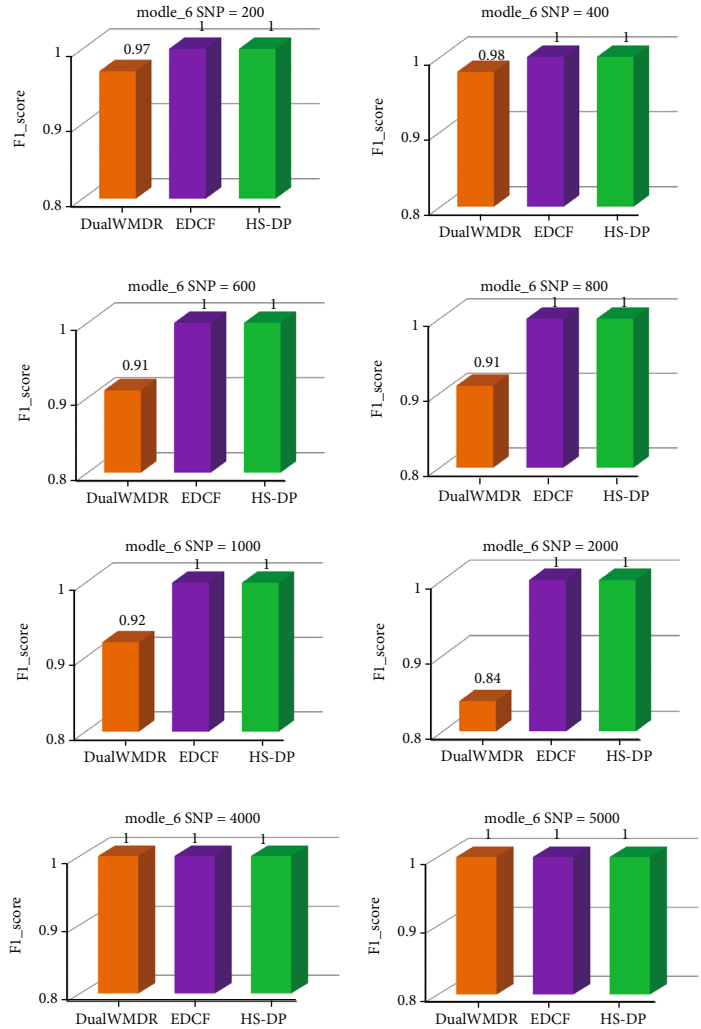
In order to protect the privacy of high-order gene interactions and improve the power, a HS-DP framework was proposed in this paper. The above sections have described the main content of HS-DP. In this section, we will verify the performance of this framework through comparing the different algorithms by virtual simulation experiment, still including the source of dataset and the experimental operating environment.

5.1. Experimental Setup. There are two types of datasets, simulated and real, of which the simulated dataset was generated by the GAMETES 2.0 software [50]. The sample size of case and control was 4000, respectively, and the SNP number changed within 5000. There were 8 disease models in total, among which models 1-4 were marginal effect models (reference literature [51]. Models 5-8 are generated from the penetrance table with no marginal effect. In addition, we selected age-related macular degeneration (AMD) [52] datasets to judge the practical performance of HS-DP. The framework was trained in a 64-bit Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz processor and 32GB RAM simulation environment. And, we used Python 3.6 as the primary programming language in Windows 10. In addition, since the interaction results of the simulated dataset are the last three SNPs, in order to ensure the actual effect of the framework, we distorted this order. Taking the identification of third-order gene interactions as an example, we selected DualWMDR [53] and EDCF [54] algorithms as the comparison algorithms to test the performance of HS-DP proposed in this paper.



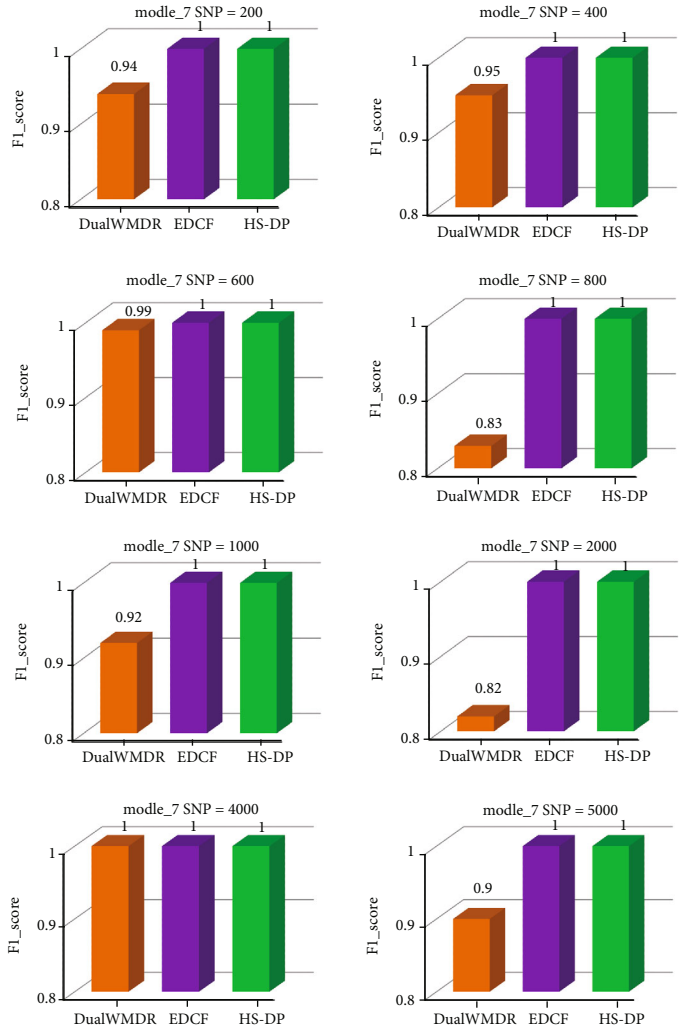
(a)

Figure 4: Continued.



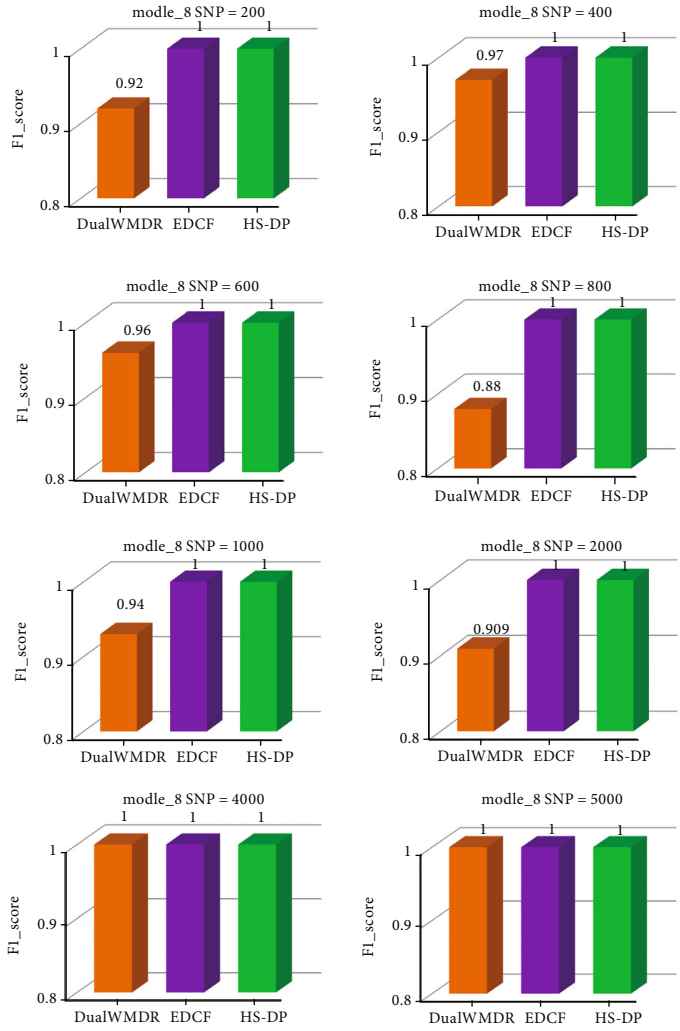
(b)

Figure 4: Continued.



(c)

Figure 4: Continued.



(d)

Figure 4: The accuracy of without-marginal effect models.

5.2. Accuracy Comparison Based on Simulated Datasets

5.2.1. With Marginal Effect. For models 1 to 4 with marginal effects, we conducted experiments on 9 types of SNP datasets of different sizes. The experimental results are shown in Figure 3.

As can be seen from Figure 3, DualWMDR can identify gene combinations when the number of SNPS in model 1 is 400, 600, and 1000, and model 4 is 400, 600, 800, 1000, and 2000. In other cases, all of the accuracies are 0. It can be concluded that DualWMDR cannot identify most disease models with marginal effects. Although EDCF detects the epistatic combination of genes in most models with higher accuracy than DualWMDR, the accuracy of the algorithm gradually decreases with the increase of data size. The HS-DP algorithm proposed in this paper identifies epistatic gene interactions in 4 disease models and 9 datasets, and the accuracy of the algorithm does not significantly decrease with the increase of SNP size. This is because the harmony search algorithm introduced by HS-DP enhances the search ability of HS-DP for high-order combinations, and five fitness functions integrated that vary on different gradients, which can retain the candidate solution set to the maximum.

5.2.2. Without Marginal Effect. For models 5 to 8 without marginal effect, we conducted experiments on 9 datasets of different sizes. The experimental results are shown in Figure 4. From Figure 4, experimental results concluded that DualWMDR can find the epistatic genes of without-marginal effect models and 9 kinds of datasets. And the accuracy of DualWMDR does not decrease with the amount of data, it showed that this algorithm is available and has a certain ability to deal with large-scale data. Although EDCF algorithm has high accuracy, it ignores the privacy protection in the research of gene interaction and still has the problem of privacy disclosure. The HS-DP algorithm introduces a differential privacy protection mechanism to solve this problem, meanwhile the detection accuracy is still as high as over 99%.

Based on Figures 3 and 4, it can be concluded that the HS-DP algorithm proposed in this paper not only meets the accuracy requirements of multiple disease models and

Table 1: Three-order epistatic result of AMD.

Gene	SNP	Location	P value
CFH, NPAT, PCDH9	rs380390, rs3781868, rs1036995	11q22, 13q21	8×10^{-18}
NRG3	rs1458402, rs2207768, rs4901408	11p15	8×10^{-18}
NXPH1, PTPRD	rs1476623, rs6967345, rs1408120	7p22, 9p23-p24	3.2×10^{-24}
KANK1	rs595113, rs1569651, rs2031175	9p24	4.9×10^{-24}
CFH, NPAT	rs132948, rs3781868, rs3781868	1p32, 11p22-23	6.78×10^{-10}
NAMPT, KCNH7	rs10487833, rs10495593, rs1740752	10p13	3.24×10^{-18}

datasets of different sizes but also protects the privacy and security.

5.3. Real Datasets. We tested the accuracy of HS-DP by being applied on age-related macular degeneration (AMD) real datasets. AMD is the leading cause of blindness in middle-aged and elderly people and is a common eye disease. We downloaded AMD data from the official website of WTCCC, which contained 96 case individuals and 50 control individuals with 103611 SNPS. Through quality control, the number of SNP is 96607. Klein et. al [55] reported two interaction results most relevant to AMD, rs380390 and rs1329428. After the initialization parameters, the HS-DP framework took these two results as the main effect SNPS to search for the corresponding third-order gene interaction results in AMD. The results are shown in Table 1.

These are the results of three-order gene interactions on AMD datasets. These SNPs are located in a number of important genes and perform important functions. For example, the CFH gene on chromosome 1 encodes a protein that plays a key role in regulating complement activation. The PCDH9 gene encodes cadherin-associated neuronal receptors and we hypothesized that it is involved in specific neuronal connections and signal transduction. In addition, other combinations of SNPS associated with AMD have been found, but their biological explanation requires further research.

6. Conclusion and the Future Work

In order to solve the problem of privacy leakage, improve detection performance, and reduce the detection burden of high-order gene interaction, a secure high-order gene interaction detection framework is proposed in this paper. The framework designed objective function perturbation mechanisms for intelligent algorithms to identify high-order gene interaction combinations. This mechanism added noise according to the distribution characteristics of polynomial data of objective function. In addition, we optimized the process of detecting epistasis by swarm intelligence algorithm and proposed a harmony search algorithm suitable for identification of high-order gene interactions. Experimental evaluations built on simulated and real datasets confirm the accuracy of our framework. In the future, our work will be expanded in the following areas. On the one hand, training convergence is accelerated to improve model accuracy. On the other hand, other noise mechanisms based on differential privacy need to be studied to protect the security of sensitive information from multiple perspectives. Finally, the study of HS-DP on large-scale datasets is also our future research direction.

Data Availability

Experimental data is divided into two types, one is simulated data, one is real data (download address: http://www.ncbi.nlm.nih.gov/SNP/).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Science and Technology Foundation Project under the Grant No. 2019FY100103, National Natural Science Foundation of China under the Grant No. 62003291, and Xuzhou Science and Technology Project under the Grant No. KC20112.

References

[1] J. Labate, J. Glaubitz, and M. Havey, "Genotyping by sequencing for SNP marker development in onion," *Genome*, vol. 63, no. 12, pp. 607–613, 2020.

[2] M. C. Mills and C. Rahal, "The GWAS diversity monitor tracks diversity by disease in real time," *Nature Genetics*, vol. 52, no. 3, pp. 242-243, 2020.

[3] N. Hao, Y. Feng, and H. Zhang, "Model selection for high-dimensional quadratic regression via regularization," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 615–625, 2018.

[4] A. Jacobsen, O. Ivanova, and S. Amini, "A framework for exhaustive modelling of genetic interaction patterns using Petri nets," *Bioinformatics*, vol. 36, no. 7, pp. 2142–2149, 2020.

[5] S. Lee, Y. Pawitan, E. Ingelsson, and Y. Lee, "Sparse estimation of gene-gene interactions in prediction models," *Statistical Methods in Medical Research*, vol. 26, no. 5, pp. 2319–2332, 2017.

[6] P. Liu, Y. Zhao, G. Liu et al., "Hybrid performance of an immortalized F2 rapeseed population is driven by additive, dominance, and epistatic effects," *Frontiers in Plant Science*, vol. 8, pp. 1–9, 2017.

[7] T. Nguyen, J. Huang, Z. X. Wu, and Y. Qing, "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests," *BMC Genomics*, vol. 16, no. S2, 2015.

[8] Y. Sun, X. Wang, and J. Shang, "Introducing heuristic information into ant Colony optimization algorithm for identifying epistasis," *IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, vol. 17, no. 4, pp. 1253–1261, 2020.

[9] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Research Notes*, vol. 3, no. 1, pp. 117–129, 2010.

[10] B. Guan and Y. Zhao, "Self-adjusting ant Colony optimization based on information entropy for detecting epistatic interactions," *Genes*, vol. 10, no. 2, pp. 114–126, 2019.

[11] S. Tuo, J. Zhang, X. Yuan, Y. Zhang, and Z. Liu, "FHSA-SED: two-locus model detection for genome-wide association study with harmony search algorithm," *PLoS One*, vol. 11, no. 3, p. e0150669, 2016.

[12] J. Shang, Y. Sun, S. Li, J. X. Liu, C. H. Zheng, and J. Zhang, "An improved opposition-based learning particle swarm optimization for the detection of SNP-SNP interactions," *BioMed Research International*, vol. 2015, Article ID 524821, 12 pages, 2015.

[13] C. Yang, L. Chuang, and Y. Lin, "CMDR based differential evolution identifies the epistatic interaction in genome-wide association studies," *Bioinformatics*, vol. 33, no. 15, pp. 2354–2362, 2017.

[14] Y. Lin, A. Chang, and D. Shuang, "FAACOSE: a fast adaptive ant Colony optimization algorithm for detecting SNP epistasis," *Complexity*, vol. 2017, Article ID 5024867, 10 pages, 2017.

[15] Y. Sun, J. Shang, J. Liu, S. Li, and C. H. Zheng, "epiACO - a method for identifying epistasis based on ant Colony optimization algorithm," *BIODATA MINING*, vol. 10, no. 1, pp. 1–13, 2017.

[16] N. Homer, S. Szelinger, M. Redman et al., "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, pp. 1–12, 2008.

[17] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," *Science*, vol. 305, no. 5681, p. 183, 2004.

[18] C. Yang, H. Chang, Y. Cheng, and L. Y. Chuang, "Novel generating protective single nucleotide polymorphism barcode for breast cancer using particle swarm optimization," *Cancer Epidemiology*, vol. 33, no. 2, pp. 147–154, 2009.

[19] L. Chuang, Y. Lin, H. Chang, and C. H. Yang, "An improved PSO algorithm for generating protective SNP barcodes in breast cancer," *PLoS One*, vol. 7, no. 5, pp. 1–9, 2012.

[20] L. Chuang, Y. Lin, and Y. Cheng, "SNP-SNP interaction using Gauss chaotic map particle swarm optimization to detect susceptibility to breast cancer," in *47th Annual Hawaii International Conference on System Sciences*, pp. 2548–2554, Waikoloa, HI, USA, 2014.

[21] C. Yang, H. Yang, and L. Chuang, "PBMDR: a particle swarm optimization-based multifactor dimensionality reduction for the detection of multilocus interactions," *Journal of Theoretical Biology*, vol. 461, pp. 68–75, 2019.

[22] L. Chuang, S. Moi, Y. Lin, and C. H. Yang, "A comparative analysis of chaotic particle swarm optimizations for detecting single nucleotide polymorphism barcodes," *Artificial Intelligence in Medicine*, vol. 73, pp. 23–33, 2016.

[23] C. Yang, Y. Kao, L. Chuang, and Y. D. Lin, "Catfish Taguchi-based binary differential evolution algorithm for analyzing single nucleotide polymorphism interactions in chronic dialysis," *IEEE Transactions on Nanobioscience*, vol. 17, no. 3, pp. 291–299, 2018.

[24] B. Guan, Y. Zhao, Y. Yin, and Y. Li, "A differential evolution based feature combination selection algorithm for high-dimensional data," *Information Sciences*, vol. 547, pp. 870–886, 2021.

[25] B. Guan, Y. Zhao, and Y. Li, "DESeeker: detecting epistatic interactions using a two-stage differential evolution algorithm," *IEEE ACCESS*, vol. 7, pp. 69604–69613, 2019.

[26] C. Yang, H. Gao, and X. Yang, "BnBeeEpi: an approach of epistasis mining based on artificial bee colony algorithm optimizing Bayesian network," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 232–239, San Diego, CA, USA, 2019.

[27] X. Li, S. Zhang, and K. Wong, "Nature-inspired multiobjective epistasis elucidation from genome-wide association studies," *IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, vol. 17, no. 1, pp. 1–237, 2018.

[28] Y. Wang, X. Liu, and R. Rekaya, "AntEpiSeeker2.0: extending epistasis detection to epistasis-associated pathway inference using ant colony optimization," *Nature Precedings*, vol. 7, 2012.

[29] N. Sinnott-Armstrong, C. Greene, and J. Moore, "Fast genome-wide epistasis analysis using ant colony optimization for multifactor dimensionality reduction analysis on graphics processing units," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation - GECCO '10*, pp. 215-216, 2010.

[30] S. Li, J. Chen, and Q. Jiao, "A flexible novel approach to learn epistasis based on ant colony optimization," in *31st Chinese Control Conference*, pp. 7370–7375, 2012.

[31] J. Shang, J. Zhang, X. Lei, Y. Zhang, and B. Chen, "Incorporating heuristic information into ant colony optimization for epistasis detection," *GENES & GENOMICS*, vol. 34, no. 3, pp. 321–327, 2012.

[32] P. Jing and H. Shen, "MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, no. 5, pp. 634–641, 2015.

[33] J. Liu, G. Yu, Y. Jiang, and J. Wang, "HiSeeker: detecting high-order SNP interactions based on pairwise SNP combinations," *Genes*, vol. 8, no. 6, pp. 153–219, 2017.

[34] E. Sapin, E. Keedwell, and T. Frayling, "An ant Colony optimization and Tabu list approach to the detection of gene-gene interactions in genome-wide association studies [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 54–65, 2015.

[35] Y. Sun, J. Shang, and J. Liu, "An improved ant Colony optimization algorithm for the detection of SNP-SNP interactions," in *12th International Conference on Intelligent Computing (ICIC)*, pp. 21–32, 2016.

[36] B. Guan, Y. Zhao, and W. Sun, "Ant colony optimization with an automatic adjustment mechanism for detecting epistatic interactions," *Computational Biology and Chemistry*, vol. 77, pp. 354–362, 2018.

[37] B. Guan, Y. Zhao, and Y. Li, "Ant Colony optimization with self-evolving parameter for detecting epistatic interactions,"

in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 249–254, San Diego, CA, USA, 2019.

[38] I. Abu Doush and E. Santos, "A sensitivity analysis for harmony search with multi-parent crossover algorithm," *INTELLIGENT SYSTEMS AND APPLICATIONS*, vol. 1, no. 1037, pp. 276–284, 2020.

[39] 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, no. 1, pp. 115–129, 2017.

[40] S. Tuo, H. Liu, and H. Chen, "Multipopulation harmony search algorithm for the detection of high-order SNP interactions," *Bioinformatics*, vol. 36, no. 16, pp. 4389–4398, 2020.

[41] X. Wu, Y. Wei, Y. Mao, and L. Wang, "A differential privacy DNA motif finding method based on closed frequent patterns," *Cluster Computing*, vol. 21, 2019.

[42] J. He, L. Cai, and X. Guan, "Differential private noise adding mechanism and its application on consensus algorithm," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4069–4082, 2020.

[43] T. Kim, G. Lee, and B. Youn, "PHM experimental design for effective state separation using Jensen-Shannon divergence," *Reliability Engineering & System Safety*, vol. 190, pp. 106503–106516, 2019.

[44] J. Mielniczuk and P. Teisseyre, "A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited," *Genetic Epidemiology*, vol. 42, no. 2, pp. 187–200, 2018.

[45] X. Cao, G. Yu, J. Liu, L. Jia, and J. Wang, "ClusterMI: detecting high-order SNP interactions based on clustering and mutual information," *International Journal of Molecular Sciences*, vol. 19, no. 8, pp. 2267–2313, 2018.

[46] Y. Chen, F. Xu, and C. Pian, "EpiMOGA: an epistasis detection method based on a multi-objective genetic algorithm," *Genes*, vol. 12, no. 2, pp. 191–216, 2021.

[47] D. Chen, "A note on Machado-Bishop theorem in weighted spaces with applications," *JOURNAL OF APPROXIMATION THEORY*, vol. 247, pp. 1–19, 2019.

[48] N. Ahad, F. Alipiah, and F. Azhari, "Applicability of G-test in analyzing categorical variables," in *THE 4TH INNOVATION AND ANALYTICS CONFERENCE & EXHIBITION (IACE 2019)*, pp. 1–9, 2019.

[49] J. Graffelman and L. Ortoleva, "A network algorithm for the X chromosomal exact test for Hardy–Weinberg equilibrium with multiple alleles," *Molecular Ecology Resources*, vol. 21, no. 5, pp. 1547–1557, 2021.

[50] R. Urbanowicz, J. Kiralis, A. Sinnott, T. Heberling, J. Fisher, and J. Moore, "Gametes: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *Bio Data Mining*, vol. 5, no. 1, pp. 5–16, 2012.

[51] Y. Zhang and J. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.

[52] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, and W. Yu, "Detecting two-locus associations allowing for interactions in genome-wide association studies," *Bioinformatics*, vol. 26, no. 20, pp. 2517–2525, 2010.

[53] X. Cao, G. Yu, W. Ren, M. Guo, and J. Wang, "DualWMDR: detecting epistatic interaction with dual screening and multifactor dimensionality reduction," *Human Mutation*, vol. 41, no. 3, pp. 719–734, 2020.

[54] M. Xie and J. Li, "Detecting genome-wide epistasis based on the clustering of relatively frequent items," *Bioinformatics*, vol. 28, no. 1, pp. 5–12, 2012.

[55] R. Klein, C. Zeiss, E. Y. Chew et al., "Complement factor h polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.