applied sciences

MDPI

Article

Efficient and Intelligent Feature Selection via Maximum Conditional Mutual Information for Microarray Data

Jiangnan Zhang [1], Shaojing Li [2], Huaichuan Yang [2], Jingtao Jiang [3] and Hongtao Shi [2,*]

[1] Network Information Management Division, Qingdao Agricultural University, Qingdao 266109, China; zjn@qau.edu.cn
[2] School of Science and Information Science, Qingdao Agricultural University, Qingdao 266109, China; lishaojing88@qau.edu.cn (S.L.); yanghuaichuan@stu.qau.edu.cn (H.Y.)
[3] College of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao 266109, China; jjtao_2518@163.com
* Correspondence: sht@qau.edu.cn

Abstract: The challenge of analyzing microarray datasets is significantly compounded by the curse of dimensionality and the complexity of feature interactions. Addressing this, we propose a novel feature selection algorithm based on maximum conditional mutual information (MCMI) to identify a minimal feature subset that is maximally relevant and non-redundant. This algorithm leverages a greedy search strategy, prioritizing both feature quality and classification performance. Experimental results on high-dimensional microarray datasets demonstrate our algorithm's superior ability to reduce dimensionality, eliminate redundancy, and enhance classification accuracy. Compared to existing filter feature selection methods, our approach exhibits higher adaptability and intelligence.

Keywords: feature selection; microarray; curse of dimensionality; mutual information; greedy search; filter methods

Citation: Zhang, J.; Li, S.; Yang, H.; Jiang, J.; Shi, H. Efficient and Intelligent Feature Selection via Maximum Conditional Mutual Information for Microarray Data. *Appl. Sci.* **2024**, *14*, 5818. https://doi.org/10.3390/app14135818

Academic Editor: Cosimo Nardi

Received: 15 May 2024
Revised: 24 June 2024
Accepted: 2 July 2024
Published: 3 July 2024

1. Introduction

With the advancement of bioinformatic technology in the domain of animal breeding, biological datasets, especially those derived from microarrays, have become vital analytical tools for animal scientists and breeders. These datasets significantly improve the precision of trait selection and breeding outcomes, enhancing decision-making processes in breeding programs [1]. However, the analysis of microarray datasets faces significant challenges due to the curse of dimensionality and the complexity of feature interactions [2]. Microarray datasets are typically characterized by a limited number of samples yet a vast array of gene features, many of which are redundant or irrelevant. This redundancy and irrelevance can obscure vital features, leading to poor performance of machine learning algorithms [3]. Additionally, an excessive number of features can lead to the curse of dimensionality, an issue characterized by an exponential increase in problem complexity as the number of features grows. This complication can hinder pattern detection and may result in overfitting or poor model generalization.

To solve these issues, feature selection (FS) is employed as a critical preprocessing step, aiming to optimize the feature space by removing irrelevant and redundant features [4]. This process not only reduces the dimensionality of datasets but also enhances the accuracy, interpretability, and efficiency of the machine learning models [5,6]. Importantly, FS assists in uncovering the crucial mechanisms that link gene expression with specific traits in livestock, supports the analysis of data patterns to identify novel genetic markers correlated with specific traits, and contributes to minimizing expenses in breeding programs [7]. This enhancement in dataset management and analysis improves the efficacy of machine learning algorithms, leading to more accurate associations between traits, genes, and their expression levels, thereby facilitating more informed decisions in trait selection and breeding strategies [8,9].

Over the past two decades, FS has been extensively studied and has had a significant impact in many fields [10,11]. There are three main types of FS methods: filter methods, wrapper methods, and embedded methods. Filter methods are based on the measurement of data characteristics that are independent of any machine learning (ML) algorithms. This method evaluates the relevance between features and the target class (as well as the dependence between features) by using different measurement criteria, such as information, distance, dependence, and consistency, and then selects features based on the evaluation results. In contrast, wrapper and embedded methods select features using a predetermined ML algorithm. Wrapper methods select features based on the classification accuracy of the ML algorithm, while embedded methods select features based on the contribution to the learning process of the ML algorithm. However, compared to filter methods, wrapper and embedded methods have several significant drawbacks, especially for high-dimensional datasets. First, they tend to inherit the bias of the predetermined ML algorithm, resulting in poor versatility for other ML algorithms. Second, they focus only on the classification accuracy of the selected features, which may lead to overfitting or the elimination of some potentially important features. Third and most importantly, they are more computationally intensive and time-consuming because they involve iterative processes of feature search, ML model building, and testing. Therefore, the focus of this study is on filter methods, which can provide a more efficient and versatile approach to FS without being tied to specific ML algorithms.

Early filter methods evaluate the relevance of each individual feature with the target class and then remove the irrelevant and redundant features according to their rankings of relevance. Such methods are called individual evaluation methods (IEMs) [12,13]. Their advantage is that they are computationally light, but they neglect the inter-feature dependence. Therefore, the selected feature subset inevitably contains some redundant features. To overcome this problem, filter methods based on subset evaluation, also known as subset evaluation methods (SEMs), have been proposed [4]. However, these methods involve two critical issues: search strategy and evaluation criterion, which are typically encountered in wrapper and embedded methods. Finding the optimal subset from a high-dimensional dataset is an NP-hard problem, which cannot be solved optimally with an exhaustive search in a reasonable amount of time. Hence, some alternative search strategies (e.g., heuristic search [14], complete search [15], and random search [16]) have been used for the FS process. Although these search strategies significantly enhance the efficiency of the FS process, the absence of robust and precise evaluation criteria presents a considerable challenge in accurately identifying the optimal feature subset.

Recently, evaluation criteria based on relevance and redundancy analysis have gained increasing attention. Among these criteria, mutual information (MI) stands out as a powerful measure of relevance, capturing both linear and non-linear relationships between features and the target class. MI quantifies the amount of information obtained about one variable through the other, making it an ideal tool for assessing the relevance of features in the context of FS. Presently, typical SEMs include MIM [17], JMI [18], CIFE [19], CMIM [20], JMIM [21], MIFS [22], mRMR [23], and maxMIFS [24]. Despite these methods possessing certain advantages in unveiling complex relationships between features and the target class, as well as among the features themselves, they also confront significant challenges, such as failing to accurately capture the maximum relevance between features and the target class and the minimum redundancy in the selected feature subset, not being able to automatically determine the optimal number of features, and overly depending on meticulous parameter calibration.

In this paper, we propose an FS algorithm based on maximum conditional MI and a greedy algorithm to search for the optimal feature subset in high-dimensional datasets. The algorithm is evaluated by comparing it to existing classical filter methods on three pig gene microarray datasets. The significant contributions of this paper are as follows:

1. Proposal of an accurate evaluation criterion based on MI for the maximum relevant no-redundant (MRNR) feature subset along with a detailed proof process. This criterion

proves effective in practical contexts to identify whether a feature subset is an MRNR feature subset.
2. Introduction of a greedy search strategy based on maximum conditional MI to efficiently search for the MRNR feature subset with the minimum size. This strategy can automatically determine the number of features in the subset without the need for any wrapper method.
3. Introducing two evaluation metrics, relevance and redundancy, to assess the feature quality of FS methods. These metrics provide a comprehensive analysis of the selected feature subset's quality and its potential impact on classification performance.

For the convenience of the reader, Abbreviations table lists the abbreviations used in this paper.

2. Methods

MI is a measure of the amount of information shared between two random variables. It evaluates the dependence and relevance between the two variables. For two random variables x and y, $MI(x;y)$ can be defined as follows:

$$\begin{gathered} MI(x;y) = H(x) + H(y) - H(x,y) \\ = -\int p(x)\log p(x)dx - \int p(y)\log p(y)dy + \int p(x,y)\log p(x,y)dxdy \end{gathered} \tag{1}$$

where $H(x)$ and $H(y)$ are the entropies of x and y, respectively, and $H(x,y)$ is the joint entropy of x and y, $p(x)$ is the probability of x, $p(y)$ is the probability of y, and $p(x,y)$ is the joint probability of x and y.

2.1. Evaluation Criterion

Maximum relevance between the feature subset and the target class is a key criterion for *FS*. Let D be a dataset that has N features $F = \{f_i, i = 1, 2,\ldots,N\}$ and c be the target class. According to the definition of MI, selecting the feature subset with maximum relevance can be formalized as selecting a feature subset $Fs = \{f_j, j = 1, 2,\ldots,n\}$ $(n < N)$ that satisfies the following equation:

$$\max MI(F_s;c) \tag{2}$$

To determine the feature subset that has the maximum mutual information (MI) with the target class, we introduce the following theorem:

Theorem 1. *For the original feature set F, the joint mutual information between any feature subset Fs and the target class c is less than or equal to the joint mutual information between the entire feature set F and the target class c. The equation form of Theorem 1 is as follows:*

$$\forall F_s \subset F,\ MI(c;F_s) \leq MI(c;F)$$

Proof of Theorem 1. Let $F_s = \{f_j, j = 1, 2, \ldots, n\}(n < N)$ be a feature subset of the original feature set F.

$$\because F_s \subset F, \therefore F = \{\underbrace{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_{N-n}}_{F-F_s}, \underbrace{f_1, f_2, \ldots, f_n}_{F_s}\}$$

where feature $\hat{f}_i$ belongs to $F - Fs$. According to the definition of MI, we have

$$\begin{gathered} MI(c;F) = MI\left(c;\hat{f}_1,\hat{f}_2,\ldots,\hat{f}_{N-n},f_1,f_2,\ldots,f_n\right) \\ = \sum_{i=1}^{N-n} MI\left(c;\hat{f}_i\middle|\hat{f}_{i-1},\ldots,\hat{f}_1,f_1,f_2,\ldots,f_n\right) + MI(c;f_1,f_2,\ldots,f_n) \\ \because \sum_{i=1}^{N-n} MI\left(c;\hat{f}_i\middle|\hat{f}_{i-1},\ldots,\hat{f}_1,f_1,f_2,\ldots,f_n\right) \geq 0, \\ \therefore MI(c;F_s) \leq MI(c;F). \end{gathered} \tag{3}$$

□

According to Theorem 1, we demonstrate that the feature subset F_s has the maximum MI with the target class if and only if F_s satisfies the following equation:

$$MI(c; F_s) = MI(c; F) \tag{4}$$

Although Equation (4) can ensure that the selected feature subset has maximum relevance, it cannot guarantee that the selected feature subset has no redundant features. In fact, a redundant feature is one that is typically deemed irrelevant to the target class given the presence of other features in the dataset. Therefore, the definition of a redundant feature for a feature set is as follows:

Definition 1. *(Redundant feature): For a feature subset F_s, a feature $f_j (f_j \in F_s)$ is a redundant feature if and only if*

$$MI\left(c; f_j | F_s - \{f_j\}\right) = 0 \tag{5}$$

or

$$MI\left(c; F_s - \{f_j\}\right) = MI(c; F_s). \tag{6}$$

By combining Theorem 1 and Definition 1, the FS for the MRNR feature subset can be defined as selecting a feature subset *Fs* that satisfies the following equations:

$$\begin{cases} MI(c; F_s) = MI(c; F) \\ \forall f_j \in F_s,\ MI\left(c; F_s - \{f_j\}\right) < MI(c; F_s) \end{cases}. \tag{7}$$

2.2. Search Strategy

Selecting the optimal MRNR feature subset quickly and accurately poses a challenging task due to the large number of candidate MRNR feature subsets in high-dimensional datasets. In this paper, from the perspective of dimensionality reduction and classification performance, we consider that the MRNR feature subset with the minimum size is the optimal feature subset.

Let the feature subset $F_{min} = \{f_k, k = 1, 2, \ldots, m\}$ $(m < N)$ be the MRNR feature subset with the minimum size. When $m = 1$, the solution for f_1 is the feature that maximizes $MI(c; f_i)(1 \le i \le N)$. When $m > 1$, $MI(c; F_{min})$ can be expressed in the form of (8):

$$\begin{gathered} MI(c; F_{min}) = MI(c; f_1, f_2, \ldots, f_m) \\ = MI(c; f_1, f_2, \ldots, f_{m-1}) + MI(c; f_m | f_{m-1}, \ldots, f_1) \\ = MI(c; f_1, f_2, \ldots, f_{m-2}) + \sum_{k=m-1}^{m} MI(c; f_k | f_{k-1}, \ldots, f_1) \\ = MI(c; f_1) + \sum_{k=2}^{m} MI(c; f_k | f_{k-1}, \ldots, f_1) \\ = \sum_{k=1}^{m} MI(c; f_k | f_{k-1}, \ldots, f_1) \end{gathered} \tag{8}$$

the *kth* feature f_k can be determined as the one that maximizes the conditional MI $MI(c; f_k | f_{k-1}, \ldots, f_1)$, where $f_{k-1}, \ldots, f_1$ are previously selected features.

Based on the above analysis, a greedy search strategy can be used to iteratively select the feature with the maximum conditional mutual information (MCMI) until the MI between the feature subset and the target class is equal to that of the original feature set and the target class. The evaluation criterion of MCMI can be formulated as follows:

$$\max_{f_i \in F - Fs} MI(c; f_i | F_s) \tag{9}$$

where F is the original feature set, F_s is the previously selected feature subset, and feature f_i belongs to $F - F_s$.

Furthermore, it should be noted that in specific situations, the later selected features may cause one or more previously selected features to become redundant. Therefore, it is necessary to remove any potential redundant features from the selected feature subset after the FS process.

2.3. Feature Selection Algorithm

This section describes the design of the FS algorithm based on MCMI. According to the analysis in Section 2.2, the algorithm consists of two processes: a greedy search and redundancy elimination, the framework of which is shown in Figure 1. In the first phase, a greedy search strategy is used to search for the candidate feature subset. In the second phase, a redundancy elimination strategy is utilized to eliminate redundant features from the candidate feature subset.

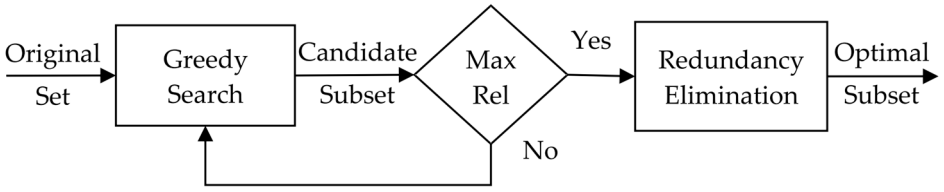

Figure 1. Framework of FS.

Algorithm 1 presents the pseudocode for the FS algorithm based on MCMI. Below, a simplified description is provided to illustrate the process, along with explanations for each step's purpose:

1. Initialization (line 1): Calculate the maximal mutual information (maxR) between the original feature set F and the target class c, serving as a benchmark for the best possible mutual information. Initialize two sets: F_{list} to hold candidate features, and F_{opt} to store the optimal features selected through the algorithm.

2. Candidate Selection (lines 2–4): Calculate the MI between each feature f_i in F and the target class c, add features with MI > 0 to F_{list}, and sort F_{list} by descending MI values. This ensures only informative features are considered.

3. Greedy Search (lines 5–12): Iteratively select the feature with the highest conditional MI $MI(c; f_i|F_{opt})$ from F_{list}, add it to F_{opt}, and remove features from F_{list} with conditional MI = 0. Repeat until $MI(c; F_{opt})$ equals maxR. This ensures the selected feature set is the minimal subset with maximal relevance to the target class c.

4. Redundancy Elimination (lines 13–15): Evaluate each feature in F_{opt} for redundancy and remove features that do not significantly contribute to MI when conditioned on the remaining features. This ensures the final feature set is maximum relevant and non-redundant.

5. Return the Optimal Feature Set (line 16): Return F_{opt} as the optimal feature set.

As shown in Algorithm 1, the greedy search strategy involves a major part of computational effort, and its time complexity has a non-linear relationship with the number of original features (dimensionality N). In the best case, the time complexity of the greedy search is $O(N)$ when only one feature is selected as a candidate feature. In the worst case, it has a time complexity of $O(N^2)$ when all features are selected as candidate features. However, in general, the time complexity of the greedy search strategy is much lower than the sequential forward selection (SFS) strategy because, in each iteration, irrelevant or conditionally irrelevant features are removed and not passed to the next iteration. This characteristic significantly reduces the computational burden compared to SFS, especially for high-dimensional datasets. In addition, the time complexity of the greedy search strategy is also influenced by the interdependence of the features in the original feature set. The more features are deleted in earlier iterations, the faster the greedy search strategy becomes. On the other hand, the redundancy elimination strategy has a linear time complexity in terms of the number of candidate features selected by the greedy search strategy.

Algorithm 1. FS algorithm based on MCMI.

Input: dataset D with N features $F = \{f_i, i = 1, 2, \ldots, N\}$ and the target class c
Output: the optimal feature subset F_{opt}
1. Initialize maxR=$MI(c; F)$, F_{list} = {}, F_{opt} = {}
2. For i = 1 to N do
3. If $MI(c; f_i) > 0$ then insert f_i into F_{list}
4. End for
5. Sort F_{list} by descending $MI(c; f_i)$
6. While F_{list} is not empty do
7. f_{temp} = getFirstElement(F_{list})
8. Insert f_{temp} into F_{opt}, remove f_{temp} from F_{list}
9. If $MI(c; F_{opt}) ==$ maxR then break
10. Remove features from F_{list} with $MI(c; f_i | F_{opt}) == 0$
11. Sort F_{list} by descending $MI(c; f_i | F_{opt})$
12. End while
13. For each feature f_i in F_{opt} (in reverse order) do
14. If $MI(c; f_i | F_{opt}) == 0$ then remove f_i from F_{opt}
15. End for
16. Return F_{opt}

3. Experiment Results

We empirically evaluated our FS algorithm by comparing it with representative FS algorithms on three high-dimensional gene microarray datasets. This section is organized as follows: After a brief introduction of the datasets, we describe the experimental setup. We then compare our FS algorithm with the representative filter-based FS algorithms in terms of feature quality, classification accuracy, and computational complexity.

3.1. Datasets

In this study, we conducted an evaluation of our proposed algorithm using three microarray datasets with high dimensionality from the breed Shandong black pig, a popular pig breed in Laiwu, Shandong, China, known for its tender and fine meat. These datasets were obtained from the Life Science Research Institute of Qingdao Agricultural University and were used to analyze intricate biological data and identify relevant genetic information. Remarkably, all datasets contained tens of thousands of features, surpassing the number of samples by a substantial margin, as shown in Table 1.

Table 1. Microarray datasets of Shandong black pig.

Datasets	Acronym	Raw Data Type	Feature Number	Sample Number	Class Number
BlackPic Expression	BPEx	Continuous	24,368	181	4
BlackPic SNP	BPSnp	Discrete	27,268	236	14
BlackPic PRRS	BPPRRS	Continuous	29,768	151	2

The BPEx dataset consists of continuous gene expression data for Shandong black pigs. It comprises 181 samples, with each having 24,368 expression features. The target class of the dataset is pork quality, categorized into four levels: excellent (22 samples), good (49 samples), average (76 samples), and inferior (34 samples). Analyzing this dataset can help identify genes associated with high-quality pork and provide valuable insights for improving pork quality through breeding and production strategies.

The BPSnp dataset comprises discrete gene single nucleotide polymorphisms (SNPs) for Shandong black pigs. It includes 236 samples, each having 27,268 SNP features. This dataset has three target classes:

1. Body size and weight are categorized into five levels: very large (49 samples), large (61 samples), medium (59 samples), short (46 samples), and very short (21 samples).

2. The ratio of muscle to fat is categorized into five levels: AAA (46 samples), AA (53 samples), A (61 samples), B (47 samples), and C (29 samples).
3. The growth rate is categorized into four levels: rapid growth (57 samples), fast growth (82 samples), moderate growth (62 samples), and slow growth (35 samples).

Analyzing the BPSnp dataset can provide insights into the genetic factors influencing body size, weight, muscle-to-fat ratio, and growth rate in Shandong black pigs, potentially contributing to improved breeding and production strategies for these traits.

The BPPRRS dataset consists of gene expression data specifically focused on porcine reproductive and respiratory syndrome (PRRS), a highly prevalent viral disease in pigs worldwide that causes significant losses in the swine industry. This dataset includes 29,768 expression features. The dataset is divided into two categories: infected (90 samples) and normal (61 samples). With its detailed gene expression profiles, the BPPRRS dataset serves as a critical resource for researchers aiming to understand the molecular mechanisms underlying PRRS, identify potential genetic markers for resistance or susceptibility to the disease, and develop effective strategies for breeding and managing pigs to minimize the impact of PRRS on pork production.

Due to the excessively large number of features in the datasets, it is imperative that FS algorithms are highly efficient to handle the computational complexity and resource requirements effectively.

3.2. Experimental Setting

The focus of this study is to analyze the effectiveness of the proposed algorithm primarily from a data science perspective. To better evaluate the algorithm, we consider the following three issues in the experimental setup:

1. Evaluation metrics

In this paper, common metrics such as overall accuracy (OA), feature number (FN), and execution time are used to evaluate feature selection (FS) algorithms. Additionally, we propose two novel metrics: relevance (Rel) and number of redundant features (NRF). The Rel metric represents the ratio of the relevance between the selected feature subset F_s and the target class c to that between the original feature set F and the target class c. It is calculated using the following equation:

$$Rel = MI(c; F_s)/MI(c; F) \tag{10}$$

The Rel value ranges from 0 to 1. The NRF metric represents the number of redundant features contained in the selected feature subset. These two metrics serve to evaluate the quality of the selected feature subset and conduct a more in-depth analysis of the relationship between feature quality and classification accuracy.

In practical terms, OA and FN are crucial for gauging the performance and complexity of models in classification tasks, while Rel and NRF are essential for elucidating and interpreting the significance and effectiveness of the selected features.

2. Comparison with existing methods

Since the proposed FS algorithm is typically a filter-based model, we introduce representative filter-based FS methods for comparison. We consider two types of methods: IEMs and SEMs. For IEMs, we choose InfoGain [25], GainRatio [13], SU [26], ChiSquare [12], and Fisher [27] as the representative methods. For SEMs, we choose CIFE [19], CMIM [20], JMIM [21], mRMR [23], and maxMIFS [24] as the representative methods. These selected comparison methods cover a wide range of filter-based techniques to assess the performance of the proposed algorithm effectively.

3. ML algorithms for classification

To demonstrate that our FS algorithm is not biased towards specific ML algorithms, we used four widely used ML algorithms: Naive Bayes (NB) [28], SVM [29], C4.5 [30],

and Random Forest (RF) [31], to evaluate the classification performance of our FS algorithm. These four ML algorithms represent different ways of solving supervisory learning problems. Their parameter settings are shown in Table 2.

Table 2. Parameter settings of ML algorithms.

Model	Hyperparameters	Values
NB	-	No additional tuning
SVM	C tol kernel	1 0.001 rbf
C4.5	Min_samples_leaf Confidence_factor_for_pruning	5 0.25
RF	n_estimators max_depth min_samples_split random_state	100 None 2 42

A 10-fold cross-validation was applied for performance evaluation. In this process, the dataset was split into ten parts, with 90% used for training and 10% for testing in each iteration. Each subset was used once as the test set, mitigating data variability and providing a reliable estimate of the model's performance. Finally, all algorithms are executed on a Windows computer with an Intel Core i7 3.6 GHz 8-thread, 8 GB RAM, and the Matlab R2018a implementation.

3.3. Analysis of Results

This section aims to demonstrate the effectiveness of the proposed algorithm by comparing it with existing filter methods. Firstly, we analyze and evaluate the results of the proposed algorithm for the five target classes on the three datasets. Table 3 presents the average results obtained using 10-fold cross-validation for the proposed algorithm. It should be noted that datasets BPSnp_1, BPSnp_2, and BPSnp_3 refer to the BPSnp dataset, with body size and weight, muscle–fat ratio, and growth rate as target classes, respectively. The results in Table 3 reflect the performance of the classification models with the proposed FS algorithm applied.

Table 3. Experimental results of the proposed algorithm.

Dataset	Rel	NRF	FN	Execution Time (s)	OA			
					NB	SVM	C4.5	RF
BPEx	1	0	8	44.74	0.887	0.904	0.844	**0.910**
BPSnp_1	1	0	9	62.21	0.925	0.923	0.918	**0.946**
BPSnp_2	1	0	9	68.16	0.922	0.933	0.931	**0.942**
BPSnp_3	1	0	8	62.49	0.928	0.940	0.932	**0.946**
BPPRRS	1	0	6	38.47	0.954	0.968	0.959	**0.970**

As shown in Table 3, the feature subsets selected by the proposed algorithm exhibit excellent data quality, as they demonstrate maximum relevance to the target class and lack any redundant features. This implies that all of the feature subsets belong to MRNR feature subsets. Notably, the algorithm's effectiveness is highlighted by its ability to select a small number of features while still achieving relatively high classification accuracy across all four ML algorithms. This success in accurately classifying the data further attests to the quality of the selected features.

Further analysis of classification accuracy reveals that the accuracy of RF is higher than that of other ML algorithms, primarily because RF is a more complex ML algorithm

that utilizes multiple classifiers compared to others. However, it is noteworthy that NB, SVM, and C4.5 have also achieved high classification accuracy, which is close to that of RF. This observation highlights that the features selected by the proposed FS algorithm possess very high quality and effectively mitigate the impact of classifiers on accuracy.

Additionally, the algorithm's performance is measured positively, as evidenced by its good results in terms of feature stability and running time. In conclusion, these evaluations affirm the effectiveness of the proposed algorithm in identifying high-quality feature subsets for data classification. In the subsequent subsections, we will compare the results of the proposed algorithm with those of existing filter methods.

3.3.1. Comparison with IEMs

IEMs have been proven to be computationally light and have shown good performance on certain datasets [10,11]. However, IEMs only rank features based on specific criteria, and the selection of feature subsets ultimately depends on the thresholds preset by researchers. In contrast, the proposed algorithm does not require preset parameters as it automatically determines the optimal number of features. To compare the two types of FS methods, we present the average results for selecting between 5 and 40 features using IEMs. In this study, InfoGain, GainRatio, SU, ChiSquare, and Fisher are used as representative IEMs for comparison. Due to space limitations, we will only present the results of these IEMs on the dataset BPEx. However, similar results were obtained on other datasets.

Figure 2 shows the average classification accuracy of the IEMs using four ML algorithms on the BPEx dataset. As seen in Figure 2, for the four ML algorithms, the classification accuracy of all the IEMs rapidly increases in the range of feature numbers between 5 and 20 and only exhibits slight changes when the feature number is greater than 20. Among the IEMs, GainRatio consistently achieves the highest classification accuracy in most cases. However, SU consistently yields significantly lower classification accuracy compared to other IEMs.

To compare the IEMs with the proposed algorithm, we choose the best classification accuracy and its corresponding feature number for different IEMs and ML algorithms as the benchmark results. Table 4 displays the comparison results between the IEMs and the proposed algorithm. As shown in Table 4, the proposed algorithm outperforms the IEMs in terms of classification accuracy in all cases. Moreover, the number of features selected by MCMI is significantly smaller than that selected by the IEMs, further emphasizing the efficiency and effectiveness of the proposed algorithm.

To further analyze the data quality of the features selected by the IEMs, Figure 3 shows the Rel and NRF of the IEMs on the BPEx dataset. From Figure 3a, it is apparent that for all IEMs except SU, the Rel increases rapidly within the feature number range of 5 to 15 and reaches 1 when the feature number exceeds or equals 15. By comparing Figures 2 and 3a, it can be inferred that the Rel of features plays a key role in classification accuracy and is directly proportional to the classification accuracy.

Furthermore, in Figure 3b, for all IEMs except SU, the NRF increases rapidly when the feature number exceeds 15. By comparing Figures 2 and 3b, it can be inferred that redundant features can have both positive and negative effects on classification accuracy. This is because these features can introduce disturbances in ML algorithms, and the impact of this disturbance, however, is often random and irregular. For certain ML algorithms, some redundant features may enhance classification accuracy by providing additional information, while others may lead to overfitting and poor performance, especially when highly correlated with other features in the dataset. Since the IEMs cannot recognize the correlation between features, it is impossible to determine the redundant features in the selected feature subset using these methods.

As a result, the proposed algorithm exhibits significant superiority over the IEMs in terms of both dimension reduction and classification accuracy, as demonstrated by the mean values presented in the last two rows of Table 4.

Table 5 presents a comparison of execution times between the IEMs and the proposed algorithm. As observed, the execution times of the IEMs are significantly less than those of the proposed algorithm. This is because the IEMs compute only the relevance between each feature and the target class, without considering the interdependence of features. In contrast, the proposed algorithm takes into account the inter-feature dependence, making it more time-consuming. However, it is important to note that the proposed algorithm does not require any preset parameters, making it more convenient and intelligent compared to the IEMs.

Table 4. Experimental results of the IEMs and MCMI.

Datasets	Classifiers	Metrics	InfoGain	GainRatio	SU	ChiSquare	Fisher	MCMI
BPEx	NB	OA	0.851	0.872	0.678	0.859	0.850	**0.887**
		FN	15	25	40	20	30	**8**
	SVM	OA	0.872	0.890	0.715	0.881	0.302	**0.904**
		FN	15	20	40	15	30	**8**
	C4.5	OA	0.809	0.824	0.609	0.813	0.797	**0.844**
		FN	10	30	20	40	35	**8**
	RF	OA	0.886	0.850	0.806	0.852	0.838	**0.910**
		FN	40	40	40	40	40	**8**
BPSnp_1	NB	OA	0.904	0.878	0.863	0.909	0.892	**0.925**
		FN	40	40	40	40	40	**9**
	SVM	OA	0.903	0.863	0.845	0.916	0.888	**0.923**
		FN	15	15	10	10	**5**	9
	C4.5	OA	0.909	0.875	0.861	0.888	0.871	**0.918**
		FN	25	40	20	40	40	**9**
	RF	OA	0.917	0.896	0.855	0.904	0.890	**0.946**
		FN	40	30	40	40	40	**9**
BPSnp_2	NB	OA	0.900	0.861	0.851	0.874	0.883	**0.922**
		FN	15	20	25	25	15	**9**
	SVM	OA	0.876	0.880	0.792	0.883	0.852	**0.933**
		FN	40	40	25	40	**5**	9
	C4.5	OA	0.892	0.905	0.852	0.898	0.864	**0.931**
		FN	30	40	40	40	40	**9**
	RF	OA	0.907	0.987	0.956	0.907	0.875	**0.942**
		FN	30	40	35	35	35	**9**
BPSnp_3	NB	OA	0.918	0.916	0.916	0.912	0.879	**0.928**
		FN	10	10	15	15	10	**8**
	SVM	OA	0.918	0.913	0.912	0.921	0.930	**0.940**
		FN	25	35	10	15	10	**8**
	C4.5	OA	0.916	0.919	0.912	0.929	0.926	**0.932**
		FN	25	30	20	10	10	**8**
	RF	OA	0.918	0.915	0.912	0.929	0.930	**0.946**
		FN	35	35	10	10	10	**8**
BPPRRS	NB	OA	0.908	0.916	0.916	0.912	0.879	**0.941**
		FN	**5**	15	25	10	5	6
	SVM	OA	0.898	0.923	0.932	0.901	0.930	**0.951**
		FN	25	35	15	15	10	**6**
	C4.5	OA	0.876	0.889	0.932	0.909	0.926	**0.949**
		FN	25	30	20	10	15	**6**
	RF	OA	0.918	0.925	0.932	0.919	0.93	**0.966**
		FN	35	35	25	15	10	**6**
Mean		OA	0.895	0.895	0.852	0.896	0.857	**0.927**
		FN	25	30.25	25.75	24.25	21.75	**8**

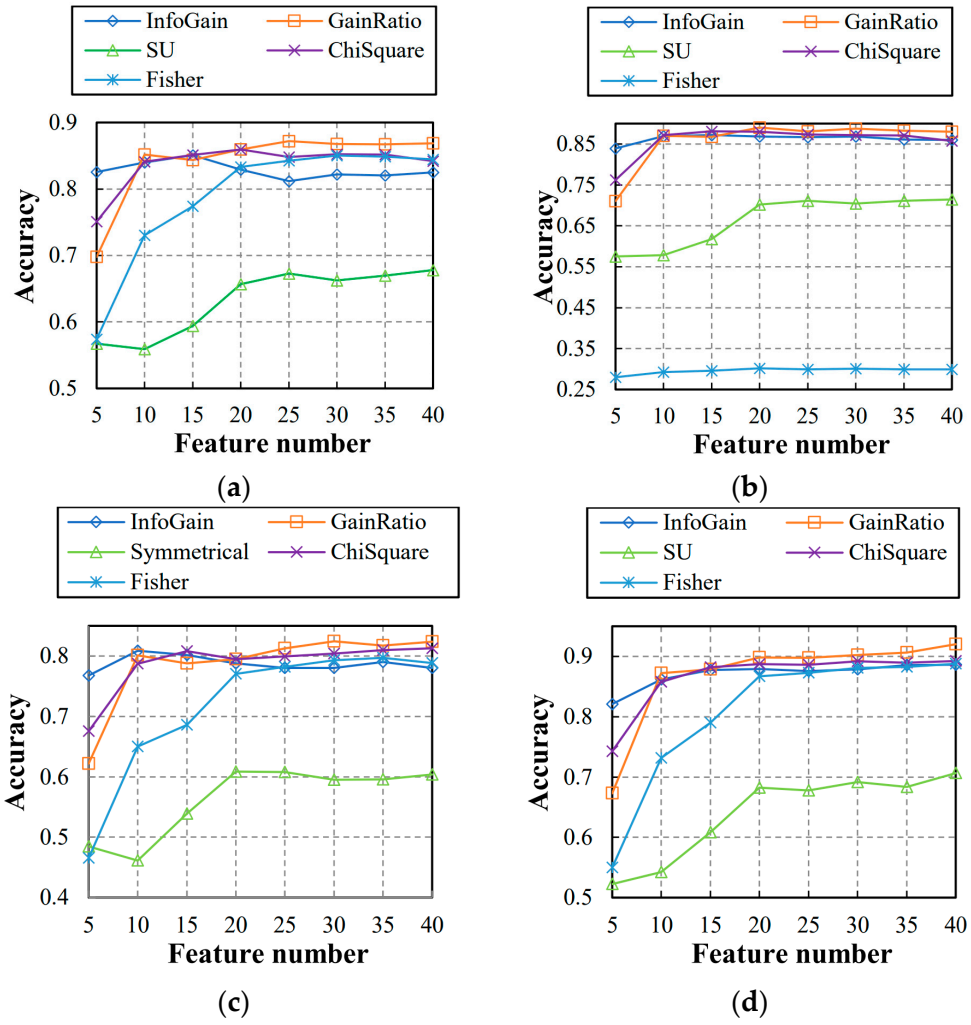

Figure 2. Classification accuracy of the IEMs on BPEx: (**a**) 10-fold cross-validation accuracy by using NB. (**b**) 10-fold cross-validation accuracy by using SVM. (**c**) 10-fold cross-validation accuracy by using C4.5. (**d**) 10-fold cross-validation accuracy by using RF.

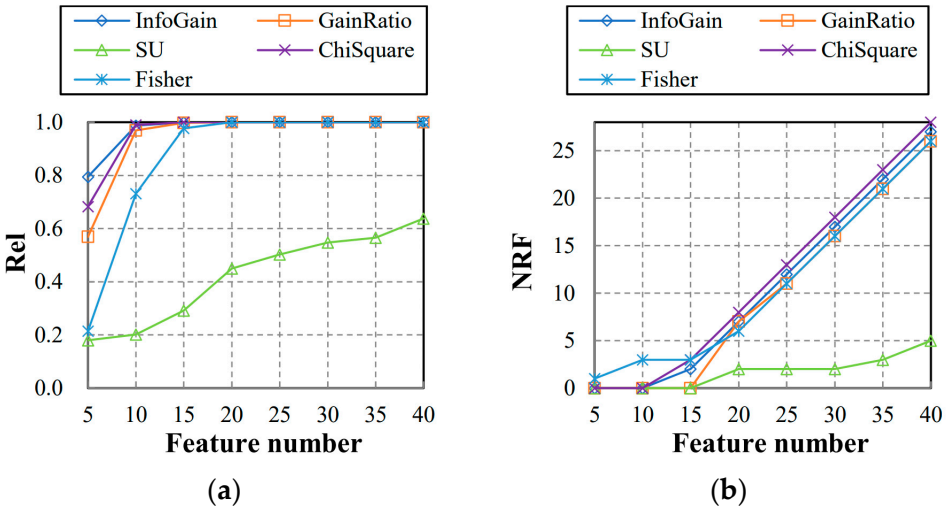

Figure 3. Feature quality of the IEMs on BPEx: (**a**) Rel. (**b**) NRF.

Table 5. Execution time of the IEMs and MCMI.

Dataset	Execution Time (s)					
	InfoGain	**GainRatio**	**SU**	**ChiSquare**	**Fisher**	**MCMI**
BPEx	26.01	24.82	26.98	19.53	**15.06**	44.74
BPSnp_1	33.75	33.77	34.99	26.70	**23.18**	62.21
BPSnp_2	33.20	34.05	32.96	27.92	**21.10**	68.16
BPSnp_3	34.09	34.46	34.61	24.52	**22.15**	62.49
BPPRRS	23.35	21.87	23.58	18.73	**13.87**	38.47

3.3.2. Comparison with SEMs

The advantage of SEMs over IEMs is their ability to consider the interdependence of features. In this study, we compare CIFE, CMIM, JMIM, mRMR, and maxMIFS as representative SEMs with the proposed algorithm. Similar to the comparison with IEMs, we present the results of selecting between 5 and 40 features using these SEMs. Figure 4 shows the average classification accuracy of the SEMs using four ML algorithms on the BPEx dataset. As seen in Figure 4, for the four ML algorithms, the classification accuracy of all the SEMs shows a significant upward trend between 5 and 10 features, then becomes stable with minor fluctuations. Furthermore, all the SEMs exhibit comparable levels of classification accuracy, and it should be noted that the CIFE and CMIM methods slightly underperform compared to the other three methods.

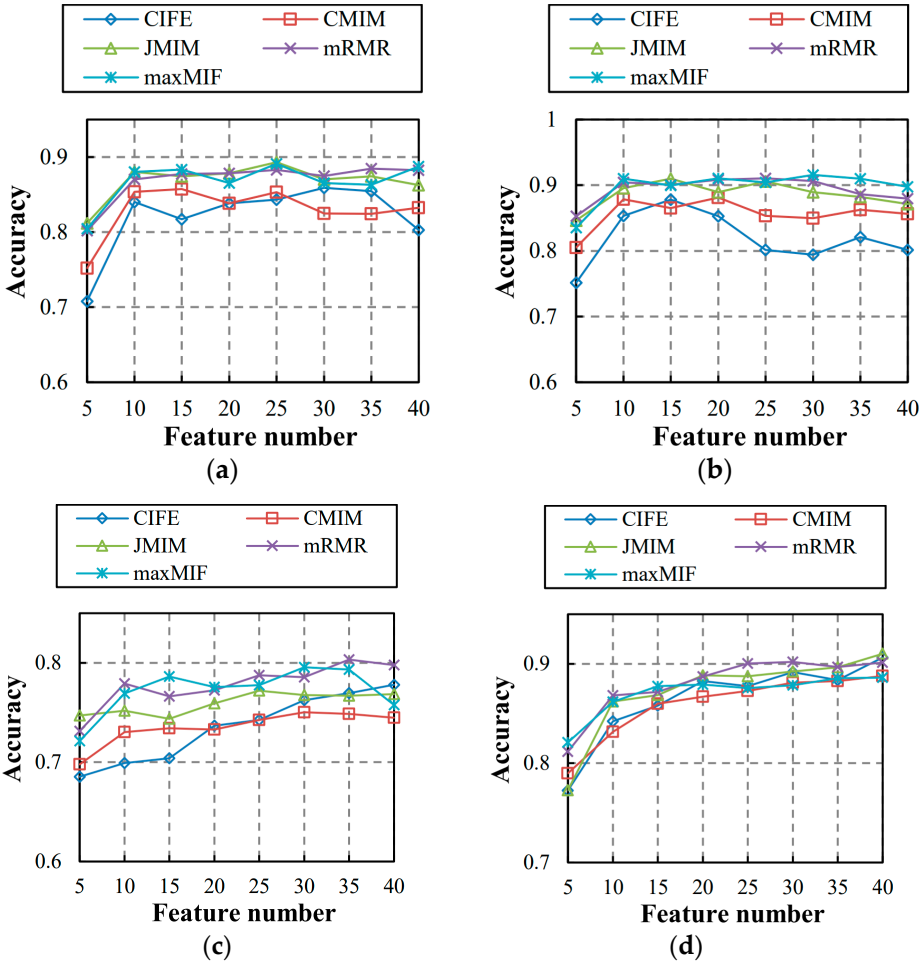

Figure 4. Classification accuracy of the SEMs on B16Ex: (**a**) 10-fold cross-validation accuracy by using NB. (**b**) 10-fold cross-validation accuracy by using SVM. (**c**) 10-fold cross-validation accuracy by using C4.5. (**d**) 10-fold cross-validation accuracy by using RF.

As with the comparison to IEMs, we chose the best classification accuracy and corresponding feature number of the SEMs using different ML algorithms, which then served as the benchmark results to be compared with those of the proposed algorithm. Table 6 presents the comparison results of the SEMs and the proposed algorithm. In the majority of cases, the proposed algorithm achieves the highest or near-highest classification accuracy. Additionally, the number of features obtained by the proposed algorithm is significantly smaller than that obtained by the SEMs in all instances. By further comparing the mean of classification accuracy and feature numbers shown in the last two rows of Table 6, it is evident that the proposed algorithm outperforms the SEMs in terms of classification accuracy and feature quality.

Table 6. Experimental results of the SEMs and MCMI.

Dataset	Classifiers	Matrices	CIFE	CMIM	JMIM	mRMR	maxMIF	MCMI
BPEx	NB	OA	0.869	0.867	**0.893**	0.885	0.891	0.887
		FN	30	15	25	35	25	**8**
	SVM	OA	0.888	0.891	**0.906**	0.911	0.916	0.904
		FN	15	20	25	25	30	**8**
	C4.5	OA	0.798	0.780	0.802	0.803	0.796	**0.844**
		FN	40	30	25	35	30	**8**
	RF	OA	0.906	0.898	**0.910**	0.902	0.906	0.910
		FN	40	40	40	30	40	**8**
BPSnp_1	NB	OA	0.889	0.898	0.900	0.898	0.901	**0.925**
		FN	15	25	35	40	35	**9**
	SVM	OA	0.909	0.922	0.925	**0.928**	0.922	0.923
		FN	20	25	15	10	25	**9**
	C4.5	OA	0.831	0.839	0.837	0.825	0.854	**0.918**
		FN	25	15	25	25	30	**9**
	RF	OA	0.922	0.926	0.932	0.920	0.930	**0.946**
		FN	15	15	25	35	40	**9**
BPSnp_2	NB	OA	0.887	0.888	0.902	0.882	**0.928**	0.922
		FN	15	35	20	30	40	**9**
	SVM	OA	0.897	0.899	0.914	0.875	0.912	**0.933**
		FN	25	25	35	40	35	**9**
	C4.5	OA	0.880	0.898	0.915	0.918	0.909	**0.931**
		FN	15	10	15	15	25	**9**
	RF	OA	0.898	0.906	0.902	0.937	0.929	**0.942**
		FN	15	10	15	35	35	**9**
BPSnp_3	NB	OA	0.897	0.909	0.929	**0.930**	0.919	0.928
		FN	30	15	35	20	25	**8**
	SVM	OA	0.919	0.911	0.917	**0.949**	0.929	0.940
		FN	20	25	15	25	20	**8**
	C4.5	OA	0.921	0.920	0.929	**0.936**	0.936	0.932
		FN	15	25	25	20	35	**8**
	RF	OA	0.921	0.929	0.937	0.934	0.942	**0.946**
		FN	25	20	20	25	35	**8**
BPPRRS	NB	OA	0.927	0.928	0.930	**0.950**	**0.950**	0.941
		FN	30	30	10	**5**	20	6
	SVM	OA	0.932	0.940	0.949	0.949	**0.959**	0.951
		FN	20	15	**5**	**5**	20	6
	C4.5	OA	0.939	0.932	0.946	**0.966**	0.948	0.949
		FN	15	10	10	15	25	**6**
	RF	OA	0.941	0.945	0.954	0.964	0.957	**0.966**
		FN	25	20	25	25	25	**6**
Mean		OA	0.898	0.901	0.911	0.913	0.917	**0.927**
		FN	22.5	21.25	22.25	24.75	29.75	**8**

Figure 5 shows the feature quality of the SEMs on the BPEx dataset. From Figure 5, it is evident that the classification accuracy of the SEMs significantly improves as the value of Rel increases. However, it's important to note that after reaching the peak value of Rel, the classification accuracy of the SEMs does not show notable changes and may slightly fluctuate with the addition of redundant features. This observation highlights that the classification performance and feature quality of the SEMs are highly dependent on the setting of the feature subset size.

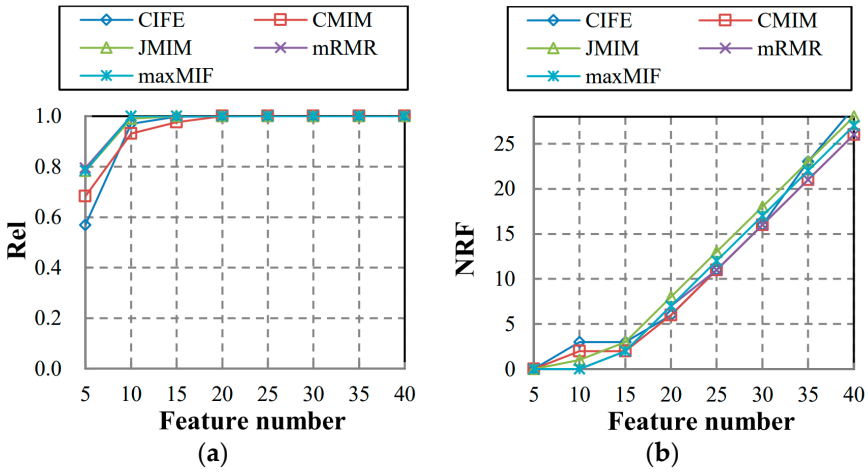

Figure 5. Feature quality of the SEMs on BPEx: (**a**) Rel. (**b**) NRF.

Table 7 shows the comparison of execution times between the SEMs and the proposed algorithm. Since the execution time of the SEMs varies with their preset feature number, we used the average number of features shown in Table 6 as the preset feature number for the experiment. From Table 7, it can be seen that the execution time of the proposed algorithm is less than that of the SEMs. In practical applications, the execution time of the SEMs far exceeds that of the proposed method, as they require iterative searches to find the optimal number of features.

Table 7. Execution times of the SEMs and MCMI.

Dataset	Execution Time (s)					
	CIFE	**CMIM**	**JMIM**	**mRMR**	**maxMIF**	**MCMI**
BPEx	74.81	74.14	78.21	68.01	67.73	**44.74**
BPSnp_1	93.21	93.87	97.45	84.74	85.47	**62.21**
BPSnp_2	88.04	87.51	92.04	80.04	80.69	**68.16**
BPSnp_3	94.74	95.78	99.04	86.13	85.34	**62.49**
BPPRRS	53.35	51.87	53.58	48.73	43.87	**38.47**

3.4. Verification of Selected Genes

To verify the effectiveness of the proposed FS algorithm, we conducted a literature review for each selected gene to confirm whether they are validated genes affecting pig traits. Table 8 summarizes the verification results.

For the BPEx dataset, our algorithm selected eight genes, six of which are documented in the literature as being related to pork quality, resulting in an effectiveness rate of 75%. For the BPSnp_1 dataset, nine genes were selected, with eight supported by research on body size and weight in pigs, yielding an effectiveness rate of 89%. For the BPSnp_2 dataset, nine genes were selected, with seven confirmed by the literature as influencing muscle development and fat distribution, resulting in an effectiveness rate of 77.78%. For the BPSnp_3 dataset, eight genes were chosen, with five affecting growth rate, giving an effectiveness rate of 62.5%. For the BPPRRS dataset, six genes were selected, with four

documented for PRRS resistance, resulting in an effectiveness rate of 66.67%. Overall, these results demonstrate the robustness and accuracy of our feature selection method across various traits in pigs, confirming its utility in genetic research and breeding programs.

Table 8. Gene verification for pig traits.

Dataset	Selected Genes	Known Genes	Validated Count	Effectiveness Rate
BPEx	IGF2, TNNT3, PPARδ, LEP, SIRT1, APOE, TRIM55, FTO	IGF2 [32], TNNT3 [33], PPARδ [34], LEP [35], SIRT1 [36], APOE [37]	6	75.00%
BPSnp_1	GHR, IGF1, LEP, MC4R, INSL3, PPARδ, SLC10A2, TNFAIP3, RYR1	GHR [38], IGF1 [39], LEP [40], MC4R [41], INSL3 [42], PPARδ [43], SLC10A2 [44], TNFAIP3 [45]	8	89.00%
BPSnp_2	IGF2, LEP, FABP3, MC4R, CAST, PLIN1, UCP3, ASIP, ADRB3	IGF2 [46], LEP [47], FABP3 [48], MC4R [49], CAST [50], PLIN1 [51], UCP3 [52]	7	77.78%
BPSnp_3	IGF1, GHR, GH1, LEP, IGFBP3, MTOR, ASIP, NOS3	IGF1 [53], GHR [54], GH1 [55], LEP [40], IGFBP3 [56]	5	62.50%
BPPRRS	CD163, GBP5, MX1, OAS1, MTOR, IGF2BP1	CD163 [57], GBP5 [58], MX1 [59], OAS1 [60]	4	66.67%

4. Discussion

FS for high-dimensional microarray datasets has always been a challenging task because of the complicated interdependence between features. In high-dimensional microarray datasets, the most critical issues in FS are the effectiveness of the evaluation criterion and the computational complexity of the search strategy. The former ensures that the selected feature subset has high quality, such as high relevance, low redundancy, small size, and good classification performance, while the latter ensures that the feature search can be completed in a short time.

Our experimental results on three high-dimensional microarray datasets demonstrate that our algorithm successfully selects the minimum MRNR feature subset, as expected, without using any wrapper methods. Compared to existing IEMs and SEMs, our algorithm achieves better feature quality (as shown in Tables 3, 4 and 6). In contrast, the IEMs perform poorly in FS for high-dimensional datasets due to their neglect of interdependency between features, resulting in the failure to identify and eliminate redundant features in selected subsets (as shown in Figure 3). Additionally, the IEMs struggle to determine the number of features to select, leading to the need for personal experience or wrapper methods, which increases instability and computational complexity. While SEMs can consider feature interdependence, existing SEMs lack effective evaluation criteria to accurately determine maximum relevance with the target class and identify redundant features in the selected subset. Although the SEMs demonstrate more excellence than the IEMs in classification performance and redundancy reduction (as shown in Figures 2–5, and Tables 4 and 6), they still require preset parameters to determine the number of features to select. Determining the optimal feature number is critical for the SEMs, as too few features lead to insufficient relevance, while too many leads to numerous redundant features. Unfortunately, accurately determining the optimal number for high-dimensional datasets is difficult without resorting to more complex wrapper methods.

In addition, we analyzed the relationship between feature quality and the classification accuracy of feature subsets using Rel and NFS. We find that the Rel of a feature subset is positively correlated with its classification accuracy, while a small number of redundant features within the feature subset have an uncertain (both positive and negative) impact on its classification accuracy. However, including too many redundant features in a feature subset has a significant negative impact on some machine learning classifiers, such as SVM. The conclusion indirectly highlights the advantages of the minimum MRNR feature subset in terms of feature quality and classification accuracy.

In terms of execution time, the IEMs have obvious advantages because they ignore the relevance between features (as shown in Table 5). In contrast, the SEMs are more time-consuming. Our algorithm, however, achieves relatively short execution times compared to SEMs, though it is more time-consuming than IEMs (as shown in Tables 5 and 7). The main time consumption of our algorithm is the calculation of the joint probability of multiple variables.

In summary, it can be concluded that the IEMs are not suitable for high-dimensional datasets because they ignore feature interdependence, leading to suboptimal feature subsets. Although SEMs consider interdependence, existing SEMs lack effective evaluation criteria, resulting in the selection of redundant features that negatively impact dimensionality reduction and classification performance. In contrast, the proposed algorithm has demonstrated strong performance in terms of feature quality, classification accuracy, and execution time, establishing its overall superiority over existing algorithms.

Finally, through the literature review verification for each selected gene in various datasets, the results demonstrate that our algorithm reliably identifies key genetic factors. This validation underscores the utility of our FS algorithm in genetic research and breeding programs.

5. Conclusions

To address the challenges of FS in high-dimensional microarray datasets, in this paper, we have introduced the concept of the MRNR feature subset and proposed a novel FS algorithm based on MCMI and a greedy algorithm. Compared to the existing filter FS algorithms, the proposed algorithm demonstrates better performance in terms of feature quality and classification accuracy. In addition, because there is no need to preset the number of selected features, this algorithm avoids dependence on a wrapper method.

In terms of running time, the proposed algorithm outperforms the traditional SFS algorithm because it eliminates features that are conditionally irrelevant to the target class in each iteration. Although it is less efficient than the IEMs, it achieves better time complexity compared to the SEMs.

In conclusion, our work provides a powerful and efficient approach for feature selection (FS) in high-dimensional microarray datasets, demonstrating its intelligence and effectiveness on the specific datasets discussed in this paper. Additionally, the results of the literature review verification highlight the utility of our FS algorithm in genetic research and breeding programs. However, further validation on a broader range of datasets is necessary to generalize its effectiveness. Future work will involve extending the application of our algorithm to more diverse datasets to confirm its robustness and general applicability.

Author Contributions: Conceptualization, J.Z. and H.S.; Methodology, J.Z. and H.S.; Writing—original draft, J.Z., H.Y. and J.J.; Writing—review and editing, H.Y., J.J. and S.L.; Supervision, S.L. and H.S.; Funding acquisition, S.L. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of Shandong Province (ZR2022MD025), the Action Plan Project for Rural Revitalization, Scientific and Technological Innovation of Shandong Province (2022TZXD0012-2, 2022TZXD007-5), the Key R&D Program of Shandong Province (Soft Science Project) (2022RKY06004), the Agricultural Improved Seed Project of Shandong Province (2020LZGC010, 2022LZGC021), the Advanced Talents Foundation of Qingdao Agricultural University (6631120066), and the Crosswise Research Tasks of Qingdao Agricultural University (6602422206, 6602423101).

Data Availability Statement: The program and some test data are available at https://github.com/HongtaoShi/MCMI (accessed on 23 June 2024). The full datasets are available upon request due to copyright restrictions by contacting sht@qau.edu.cn.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

FS	Feature Selection
ML	Machine Learning
MI	Mutual Information
IEMs	Individual Evaluation Methods
SEM	Subset Evaluation Methods
MIM	Mutual Information Maximization
JMI	Joint Mutual Information
CIFE	Conditional Infomax Feature Extraction
CMIM	Conditional Mutual Information Maximization
JMIM	Joint Mutual Information Maximization
MIFS	Mutual Information Feature Selection
mRMR	Minimum Redundancy Maximum Relevance
maxMIFS	Maximum Mutual Information Feature Selection
MRNR	Maximum Relevant No-Redundant
SU	Symmetrical Uncertainty
MCMI	Maximum Conditional Mutual Information
SFS	Sequential Forward Selection
SNPs	Single Nucleotide Polymorphisms
PRRS	Porcine Reproductive and Respiratory Syndrome
OA	Overall Accuracy
FN	Feature Number
Rel	Relevance
NRF	Number of Redundant Features
NB	Naive Bayes
SVM	Support Vector Machine
RF	Random Forest

References

1. Kyselová, J.; Tichý, L.; Jochová, K. The role of molecular genetics in animal breeding: A minireview. *Czech J. Anim. Sci.* **2021**, *66*, 107–111. [CrossRef]
2. Alhenawi, E.A.; Al-Sayyed, R.; Hudaib, A.; Mirjalili, S. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Comput. Biol. Med.* **2022**, *140*, 105051. [CrossRef] [PubMed]
3. Bellman, R. *Adaptive Control Processes: A Guided Tour*; Princeton University Press: Princeton, NJ, USA, 1961.
4. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
5. Guyon, I.; Weston, J.; Barhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
6. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **2005**, *3*, 185–205. [CrossRef]
7. Chuang, L.-Y.; Chang, H.-W.; Tu, C.-J.; Yang, C.-H. Improved binary PSO for feature selection using gene expression data. *Comput. Biol. Chem.* **2008**, *32*, 29–38. [CrossRef] [PubMed]
8. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; de Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1106–1119. [CrossRef] [PubMed]
9. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
10. Agrawal, P.; Abutarboush, H.F.; Ganesh, T.; Mohamed, A.W. Metaheuristic algorithms on feature selection: A survey of one decade of research (2009–2019). *IEEE Access* **2021**, *9*, 26766–26791. [CrossRef]
11. Nguyen, B.H.; Xue, B.; Zhang, M. A survey on swarm intelligence approaches to feature selection in data mining. *Swarm Evol. Comput.* **2020**, *54*, 100663. [CrossRef]
12. Su, C.; Hsu, J. An extended chi2 algorithm for discretization of real value attributes. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 437–441.
13. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2006.
14. Blum, L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271. [CrossRef]
15. Dash, M.; Liu, H. Consistency-based search in feature selection. *Artif. Intell.* **2003**, *151*, 155–176. [CrossRef]
16. Li, T.; Zhan, Z.H.; Xu, J.C.; Yang, Q.; Ma, Y.Y. A binary individual search strategy-based bi-objective evolutionary algorithm for high-dimensional feature selection. *Inf. Sci.* **2022**, *610*, 651–673. [CrossRef]

17. Lewis, D.D. Feature selection and feature extraction for text categorization. In Proceedings of the Workshop on Speech and Natural Language, Association for Computational Linguistics, New York, NY, USA, 23–26 February 1992; pp. 212–217.
18. Yang, H.H.; Moody, J. Data visualization and feature selection: New algorithms for nonGaussian data. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1999; pp. 687–693.
19. Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006*; Leonardis, A., Bischof, H., Pinz, A., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; pp. 68–82.
20. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
21. Bennasar, M.; Hicks, Y.; Setchi, R. Feature selection using joint mutual information maximization. *Expert Syst. Appl.* **2015**, *42*, 8520–8532. [CrossRef]
22. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef]
23. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
24. Pascoal, C.; Oliveira, M.R.; Pacheco, A.; Valadas, R. Theoretical evaluation of feature selection methods based on mutual information. *Neurocomputing* **2017**, *226*, 168–181. [CrossRef]
25. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Online Library: Hoboken, NJ, USA, 1991; Volume 6.
26. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes*; Cambridge University Press: Cambridge, MA, USA, 1988.
27. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2001.
28. Mitchell, T. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
29. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
30. Quinlan, J. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: Burlington, MA, USA, 1993.
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Burgos, C.; Galve, A.; Moreno, C.; Altarriba, J.; Reina, R.; García, C.; López-Buesa, P. The effects of two alleles of IGF2 on fat content in pig carcasses and pork. *Meat Sci.* **2012**, *90*, 309–313. [CrossRef]
33. Ji, J.; Zhou, L.; Huang, Y.; Zheng, M.; Liu, X.; Zhang, Y.; Huang, C.; Peng, S.; Zeng, Q.; Zhong, L.; et al. A whole-genome sequence based association study on pork eating quality traits and cooking loss in a specially designed heterogeneous F6 pig population. *Meat Sci.* **2018**, *146*, 160–167. [CrossRef] [PubMed]
34. Luo, H.F.; Wei, H.K.; Huang, F.R.; Zhou, Z.; Jiang, S.W.; Peng, J. The effect of linseed on intramuscular fat content and adipogenesis related genes in skeletal muscle of pigs. *Lipids* **2009**, *44*, 999. [CrossRef] [PubMed]
35. Kennes, Y.M.; Murphy, B.D.; Pothier, F.; Palin, M.F. Characterization of swine leptin (LEP) polymorphisms and their association with production traits. *Anim. Genet.* **2001**, *32*, 215–218. [CrossRef] [PubMed]
36. Gao, Y.; Li, Z.; Zhang, Q.; Hao, T.; Liu, H.; Liu, Q.; Liu, L.; Zhang, Z.; Yu, Y.; Li, N. Comparison of meat quality, muscle fiber characteristics and the Sirt1/AMPK/PGC-1α pathway in different breeds of pigs. *Anim. Prod. Sci.* **2024**, *in press*.
37. Passols, M.; Llobet-Cabau, F.; Sebastià, C.; Castelló, A.; Valdés-Hernández, J.; Criado-Mesas, L.; Sánchez, A.; Folch, J.M. Identification of genomic regions, genetic variants and gene networks regulating candidate genes for lipid metabolism in pig muscle. *Animal* **2023**, *17*, 101033. [CrossRef]
38. Brameld, J.M. Molecular mechanisms involved in the nutritional and hormonal regulation of growth in pigs. *Proc. Nutr. Soc.* **1997**, *56*, 607–619. [CrossRef]
39. Niu, P.; Kim, S.W.; Choi, B.H.; Kim, T.H.; Kim, J.J.; Kim, K.S. Porcine insulin-like growth factor 1 (IGF1) gene polymorphisms are associated with body size variation. *Genes Genom.* **2013**, *35*, 523–528. [CrossRef]
40. Balatsky, V.; Oliinychenko, Y.; Sarantseva, N.; Getya, A.; Saienko, A.; Vovk, V.; Doran, O. Association of single nucleotide polymorphisms in leptin (LEP) and leptin receptor (LEPR) genes with backfat thickness and daily weight gain in Ukrainian Large White pigs. *Livest. Sci.* **2018**, *217*, 157–161. [CrossRef]
41. Ovilo, C.; Fernández, A.; Rodríguez, M.C.; Nieto, M.; Silió, L. Association of MC4R gene variants with growth, fatness, carcass composition and meat and fat quality traits in heavy pigs. *Meat Sci.* **2006**, *73*, 42–47. [CrossRef]
42. Krupova, Z.; Krupa, E.; Žáková, E.; Zavadilová, L.; Kvašná, E. Candidate genes for congenital malformations in pigs. *Acta Fytotechn. Zootech.* **2021**, *24*, 309–314. [CrossRef]
43. Wang, Z.; Li, Y.; Wu, L.; Guo, Y.; Yang, G.; Li, X.; Shi, X.E. Rosiglitazone-induced PPARγ activation promotes intramuscular adipocyte adipogenesis of pig. *Anim. Biotechnol.* **2023**, *34*, 3708–3717. [CrossRef]
44. Liu, M.; Lan, Q.; Yang, L.; Deng, Q.; Wei, T.; Zhao, H.; Peng, P.; Lin, X.; Chen, Y.; Ma, H.; et al. Genome-wide association analysis identifies genomic regions and candidate genes for growth and fatness traits in Diannan small-ear (DSE) pigs. *Animals* **2023**, *13*, 1571. [CrossRef] [PubMed]
45. Zhang, H.; Zhuang, Z.; Yang, M.; Ding, R.; Quan, J.; Zhou, S.; Gu, T.; Xu, Z.; Zheng, E.; Cai, G.; et al. Genome-wide detection of genetic loci and candidate genes for body conformation traits in Duroc × Landrace × Yorkshire crossbred pigs. *Front. Genet.* **2021**, *12*, 664343. [CrossRef] [PubMed]

46. Aslan, O.; Hamill, R.M.; Davey, G.; McBryan, J.; Mullen, A.M.; Gispert, M.; Sweeney, T. Variation in the IGF2 gene promoter region is associated with intramuscular fat content in porcine skeletal muscle. *Mol. Biol. Rep.* **2012**, *39*, 4101–4110. [CrossRef] [PubMed]
47. Tempfli, K.; Simon, Z.; Kovács, B.; Posgay, M.; Papp, Á.B. PRLR, MC4R and LEP polymorphisms, and ADIPOQ, A-FABP and LEP expression in crossbred Mangalica pigs. *J. Anim. Plant Sci.* **2015**, *25*, 1746–1752.
48. Xue, W.; Wang, W.; Jin, B.; Zhang, X.; Xu, X. Association of the ADRB3, FABP3, LIPE, and LPL gene polymorphisms with pig intramuscular fat content and fatty acid composition. *Czech J. Anim. Sci.* **2015**, *60*, 60–66. [CrossRef]
49. Galve, A.; Burgos, C.; Silió, L.; Varona, L.; Rodríguez, C.; Ovilo, C.; López-Buesa, P. The effects of leptin receptor (LEPR) and melanocortin-4 receptor (MC4R) polymorphisms on fat content, fat distribution and fat composition in a Duroc × Landrace/Large White cross. *Livest. Sci.* **2012**, *145*, 145–152. [CrossRef]
50. Kušec, I.D.; Kušec, G.; Vuković, R.; Has-Schön, E.; Kralik, G. Differences in carcass traits, meat quality and chemical composition between the pigs of different CAST genotype. *Anim. Prod. Sci.* **2015**, *56*, 1745–1751. [CrossRef]
51. Li, B.; Weng, Q.; Dong, C.; Zhang, Z.; Li, R.; Liu, J.; Jiang, A.; Li, Q.; Jia, C.; Wu, W.; et al. A key gene, PLIN1, can affect porcine intramuscular fat content based on transcriptome analysis. *Genes* **2018**, *9*, 194. [CrossRef]
52. Damon, M.; Vincent, A.; Lombardi, A.; Herpin, P. First evidence of uncoupling protein-2 (UCP-2) and-3 (UCP-3) gene expression in piglet skeletal muscle and adipose tissue. *Gene* **2000**, *246*, 133–141. [CrossRef] [PubMed]
53. Casas-Carrillo, E.; Kirkpatrick, B.W.; Prill-Adams, A.; Price, S.G.; Clutter, A.C. Relationship of growth hormone and insulin-like growth factor-1 genotypes with growth and carcass traits in swine. *Anim. Genet.* **1997**, *28*, 88–93. [CrossRef]
54. Te Pas, M.F.W.; Visscher, A.H.; de Greef, K.H. Molecular genetic and physiologic background of the growth hormone–IGF-I axis in relation to breeding for growth rate and leanness in pigs. *Domest. Anim. Endocrinol.* **2004**, *27*, 287–301. [CrossRef] [PubMed]
55. Urban, T.; Kuciel, J.; Mikolasova, R. Polymorphism of genes encoding for ryanodine receptor, growth hormone, leptin and MYC protooncogene protein and meat production in Duroc pigs. *Czech J. Anim. Sci.* **2002**, *47*, 411–417.
56. Liu, D.W.; Zhang, H.; Wu, Z.F.; Li, J.Q.; Yang, G.F.; Zhang, X.Q. Identification of SNPs and Their Effects on Swine Growth and Carcass Traits for Porcine IGFBP-3 Gene. *Agric. Sci. China* **2008**, *7*, 630–635. [CrossRef]
57. Torricelli, M.; Fratto, A.; Ciullo, M.; Sebastiani, C.; Arcangeli, C.; Felici, A.; Giovannini, S.; Sarti, F.M.; Sensi, M.; Biagetti, M. Porcine Reproductive and Respiratory Syndrome (PRRS) and CD163 Resistance Polymorphic Markers: What Is the Scenario in Naturally Infected Pig Livestock in Central Italy? *Animals* **2023**, *13*, 2477. [CrossRef] [PubMed]
58. Khatun, A.; Nazki, S.; Jeong, C.G.; Gu, S.; Mattoo, S.U.S.; Lee, S.I.; Yang, M.S.; Lim, B.; Kim, K.S.; Kim, B.; et al. Effect of polymorphisms in porcine guanylate-binding proteins on host resistance to PRRSV infection in experimentally challenged pigs. *Vet. Res.* **2020**, *51*, 1–14. [CrossRef]
59. Niu, P.; Shabir, N.; Khatun, A.; Seo, B.J.; Gu, S.; Lee, S.M.; Lim, S.K.; Kim, K.S.; Kim, W.I. Effect of polymorphisms in the GBP1, Mx1 and CD163 genes on host responses to PRRSV infection in pigs. *Vet. Microbiol.* **2016**, *182*, 187–195. [CrossRef]
60. Zhao, J.; Feng, N.; Li, Z.; Wang, P.; Qi, Z.; Liang, W.; Zhou, X.; Xu, X.; Liu, B. 2′, 5′-Oligoadenylate synthetase 1 (OAS1) inhibits PRRSV replication in Marc-145 cells. *Antivir. Res.* **2016**, *132*, 268–273. [CrossRef]