


Article

Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data

Waleed Ali ^{1,*}  and Faisal Saeed ² 

¹ Information Technology Department, Faculty of Computing and Information Technology-Rabigh, King Abdulaziz University, Jeddah 25729, Saudi Arabia

² DAAI Research Group, Department of Computing and Data Science, School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK

* Correspondence: waabdullah@kau.edu.sa; Tel.: +966-563887947

Abstract: The advancements in intelligent systems have contributed tremendously to the fields of bioinformatics, health, and medicine. Intelligent classification and prediction techniques have been used in studying microarray datasets, which store information about the ways used to express the genes, to assist greatly in diagnosing chronic diseases, such as cancer in its earlier stage, which is important and challenging. However, the high-dimensionality and noisy nature of the microarray data lead to slow performance and low cancer classification accuracy while using machine learning techniques. In this paper, a hybrid filter-genetic feature selection approach has been proposed to solve the high-dimensional microarray datasets problem which ultimately enhances the performance of cancer classification precision. First, the filter feature selection methods including information gain, information gain ratio, and Chi-squared are applied in this study to select the most significant features of cancerous microarray datasets. Then, a genetic algorithm has been employed to further optimize and enhance the selected features in order to improve the proposed method's capability for cancer classification. To test the proficiency of the proposed scheme, four cancerous microarray datasets were used in the study—this primarily included breast, lung, central nervous system, and brain cancer datasets. The experimental results show that the proposed hybrid filter-genetic feature selection approach achieved better performance of several common machine learning methods in terms of Accuracy, Recall, Precision, and F-measure.

Keywords: cancer classification; filter feature selection; genetic algorithm; gene selection; microarray dataset



Citation: Ali, W.; Saeed, F. Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data. *Processes* **2023**, *11*, 562. <https://doi.org/10.3390/pr11020562>

Academic Editors: Xiong Luo and Zhibin Lin

Received: 20 December 2022

Revised: 14 January 2023

Accepted: 10 February 2023

Published: 12 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the last two decades, research studies in health informatics have investigated several issues related to bioinformatics, cheminformatics, cancer prediction, and others. For instance, the estimation of the number of deaths caused by heart disease was about 12 million deaths yearly worldwide according to World Health Organization (WHO).

Several methods have utilized some common machine learning for the prediction of gene selection and cancer informatics such as [1,2], prediction of new bioactive molecules [3], and heart disease prediction [4–7]. Although there are many research studies conducted on cancer informatics, the cancer disease still threatens human lives and its rare increases over time since the prediction of this dangerous disease in its earlier stage is a big issue in health informatics.

In the past few years, developing many methods based on microarray datasets for analyzing gene expression provided new ways to conduct hot research in bioinformatics, cancer prediction, and similar fields [8]. These datasets contain information about human genes and methods of their expressions. Based on the analysis of this information, several

studies could be conducted by biologists efficiently, which means they will consume less time and low cost to run their experiments [9].

Recently, many machine learning methods have been applied in the analysis of microarray datasets used for cancer classification [10]. Using the expressions of the genes in microarray datasets can be utilized as a good technique for cancer diagnoses. However, the number of existing genes is growing, about more than hundreds of thousand, while the sizes of the available datasets are still small, which contain fewer subsets of samples. This leads to the curse of dimensionality, which is one of the issues in the analysis of microarray datasets used for cancer classification [10]. In addition, there is another issue related to the nature of the existing datasets which include many redundant and irrelevant features that negatively affect the computational cost [11]. The duplicated and irrelative features do not help to provide a good classification and perdition in high-dimensional data [12]. These features reduce the performance of the prediction model and make the search for valuable knowledge more difficult. Therefore, feature selection methods are needed to be applied to improve the classifier's accuracy [13].

In order to improve the performance of these popular machine learning techniques, several feature selection methods have been utilized to select the most significant features of cancerous microarray datasets [14–20]. Although the filter feature selection methods are computationally faster and can be used to reduce the high dimension of microarray datasets, their performances are not sufficiently accurate and different since the features are evaluated independently of classifiers. In contrast, the wrapper feature selection methods interact with the classifier during the features evaluation, so they achieve better results compared to the filter method. However, the wrapper methods are time-consuming when they are applied on high-dimensional microarray datasets.

In the last few years, evolutionary algorithms are successfully employed in feature selection in many fields [21–24]. Although evolutionary algorithms-based feature selection methods overcome the filter and wrapper method, they may require a longer time for some machine learning algorithms.

Since the cancerous microarray datasets are high dimensional datasets including a vast number of features, it is impractical to use evolutionary algorithms at the beginning as feature selection methods. This encourages us to propose a hybrid filter-genetic feature selection approach that inherits the advantages of both methods and can produce promising solutions with higher performance of cancer classification in high-dimensional microarray datasets.

In this paper, combinations of filter methods and genetic algorithm-based feature selection methods are applied to identify an optimal subset of features for enhancing the cancer classification performance of machine learning methods on high-dimensional microarray datasets. In this study, information gain (IG), gain ratio (IGR), and Chi-squared (CS) are applied as three common filter methods to compute a score of each feature of microarray cancer datasets. Accordingly, only the top-ranked features are selected while the other redundant and irrelevant features are eliminated to reduce the high-dimensional microarray datasets. Then, the reduced cancer datasets with only the top-ranked features selected by the filter methods are further optimized by the genetic algorithm (GA) to achieve better cancer classification results. We can summarize the main contributions of the paper as follows:

- Compared to previous works, we used IG, IGR and CS as three popular, simple and fast filter techniques to choose highly relevant features in order to reduce high-dimensional datasets: Brain, Breast, Lung, and CNS datasets. Although many microarray datasets are used in the literature, recent work [25] reported that the popular machine learning techniques achieved the lowest classification accuracy on these specific four microarray datasets: Brain, Breast, Lung, and CNS datasets. Furthermore, the performance improvements produced by several existing works on these specific four cancer datasets were limited.

- Since IG, IGR and CS evaluate features individually by finding the relationship between each feature individually with the class label, GA is then utilized to find the relationship between a set of features together with the class label to further optimize the selected features obtained from the filter methods to enhance the cancer classification performance.
- The experimental results showed outstanding enhancements accomplished using the proposed hybrid filter-genetic feature selection approach.

The remainder of the article is structured as follows. Section 2 presents the related studies on feature selection for gene selection and machine learning methods used for cancer prediction. Filter feature selection and genetic algorithm are explained in Sections 3 and 4, respectively. Section 5 presents the research methodology of the proposed hybrid filter-GA feature selection method. Section 6 presents the experiments and evaluation, and then discusses the performance results of the proposed method. Section 7 concludes the main findings of this paper.

2. Related Work

Generally, cancer disease is considered one of the main leading reasons of death. For saving patients' lives, it is important to early identify and predict the cancer type using advanced technological solutions, such as artificial intelligence and machine learning. Several medical datasets were used in these diagnoses, including the microarray gene expression data. According to work in [26], the microarray datasets suffer from two issues, the high dimensionality, and the small sample size, which make cancer classification a nontrivial task. The authors in [27] discussed the issue of high dimensionality for the gene expression dataset, which is known as the microarray dataset, and reported that selecting the most important genes is still a challenging task in this research field.

Several feature selection and machine learning methods were used on genetic datasets. For instance, the work in [28] selected the genes that act as regulators and mediate the activity of transcription factors that have been found in all promoters of the expressed gene sets. The selected gene set was fed to Dynamic Bayesian Networks (DBNs) to classify the tumor from normal samples. The authors in [29] proposed a feature selection method using the discrete wavelet transform (DWT) and a modified genetic algorithm to identify the most important and relevant features for microarray cancer classification. The findings of this study showed, in most cases, superior results compared to the existing classification techniques. Similarly, the work in [30] used five microarray cancer datasets for cancer classification and proposed feature selection methods based on wrapper and Markov blanket models. The experimental results offered high accuracy rates compared to the traditional classification methods applied on cancer microarray datasets.

Genetic algorithm (GA) is actively used as a feature selection method in different applications. For gene selection, GA was used with a t-test in [31] as an ensemble feature selection method. In this study, the t-test was used to pre-process the data, and then Nested-GA was applied to get the optimal set of genes on colon cancer and lung datasets. Ghosh et al. [32] introduced a feature selection method with two stages on microarray datasets. In the first stage, the union and intersection of the top-n features of symmetrical uncertainty, chi-square, and ReliefF were used as ensemble filter methods. The results of this stage were fed to the GA to get the optimum set of features. The proposed method was applied on five cancer datasets and the findings showed super performance compared to the existing methods. Recently, Abasabadi et al. [33] introduced a hybrid feature selection method by combining SLI- γ filter feature selection method and genetic algorithm (GA). The proposed model showed robust prediction and less execution time, especially when 1% of the best-ranked features were used for generating the GA population. Similarly, the authors in [34] highlighted the importance of proposing feature selection for microarray datasets because of the risk of over-fitting due to the small size of the data samples. Therefore, they introduced Multi-Fitness RankAggreg Genetic Algorithm (MFRAG) that combines nine feature selection methods for evaluating the feature weights and individuals and using

ensemble models to compute the individual fitness. The experiments were conducted on several microarray datasets and the findings showed that the proposed method obtained superior accuracy comparing to the existing methods. In addition, the authors in [27] developed a feature selection method on several cancerous microarray datasets based on monarch butterfly optimization that is wrapped with the Broad Learning System (BLS).

Other previous studies applied features selection and machine learning methods on the same datasets used in this work such as Brain, Breast, Lung, and CNS datasets. For instance, Hameed et al. [35] applied the combination of Pearson's Correlation Co-efficient (PCC) with Genetic Algorithm (GA) or Binary Particle Swarm Optimization (BPSO) for on these microarray datasets. They obtained a good performance when SVM was applied with PCC and GA feature selection combination (up to 98.33% of accuracy for CNS datasets). However, these methods obtained lower performance for the other datasets (up to 88.66% of accuracy for the same model on Breast dataset). In addition, the authors in [36] applied fusion-based feature selection method on Brain, Breast and CNS microarray datasets. The highest accuracy was achieved (95%) when SVM was applied on Brain dataset. However, the model achieved lower performance with the other datasets. Similarly, a hybrid feature selection method on these four microarray datasets was applied in [37]. The method combined the Gini index and support vector machine with Recursive Feature Elimination (GI-SVM-RFE). However, the highest achieved accuracy by this model was 90.67% for Breast dataset. In addition, Almugren and Alshamlan [38] conducted a survey on the existing hybrid filter feature selection and wrapper feature selection with machine learning methods that were applied on microarray datasets. It can be observed that most of the conducted studies [39–45] worked on the datasets with lower dimensionality (comparing to the Brain and Breast datasets applied in this study) such as Colon, leukemia 1, leukemia 2, Prostate and SRBCT (Small round blue cell tumors) datasets.

Although evolutionary algorithms have been utilized in the feature selection process on microarray datasets for cancer classification, using evolutionary algorithms as filter or wrapper feature selection methods in microarray datasets is still being investigated in recent studies. Furthermore, there is still a need to conduct more research works to investigate different hybridizations and combinations of filter methods with evolutionary algorithms on different microarray datasets.

3. Filter Feature Selection

Many microarray datasets suffer from the problem of high-dimensional data with noisy data, which can cause inaccurate prediction and low classification accuracy, and slow performance of machine learning techniques [26]. Feature selection is one of the most crucial pre-processing steps used to identify the most influential features in order to increase the performance of machine learning. Due to limited resources, it is impracticable or complicated to use all features of high-dimensional microarray datasets with machine learning algorithms. Thus, it is crucial to utilize a feature selection method in cancer classification problems of high-dimensional microarray datasets to remove noisy data and eliminate redundant and irrelevant features [27].

The feature selection methods are broadly classified into filter and wrapper approaches based on the process of feature evaluation. In the filter approaches, the features are evaluated based on certain criteria independently of a classifier. The wrapper approaches, by contrast, employ a classifier to evaluate the features and then select the best features. The wrapper methods are computationally intensive since they train a machine learning algorithm several times with many potential subsets of features. In contrast, the filter approaches are easier and faster compared to the wrapper approaches as they are accomplished before the training of a machine learning algorithm [22,46].

4. Genetic Algorithm

The genetic algorithm [47] is one of the most effective evolutionary algorithms inspired by the biological evolution of chromosomes. The genetic algorithm (GA) is successfully

utilized for solving several searching and optimization problems in many real-world applications. In recent years, GA has been used effectively to identify the optimal features set in many different fields [21–24,48].

GA starts by initializing a population consisting of a set of chromosomes created arbitrarily. Each chromosome in the population represents a potential solution and includes several genes. Then, GA reproduces new better chromosomes (solutions) by evaluating the current chromosomes and then recombining the fittest chromosomes.

At each GA generation, a pair of fit chromosomes are chosen depending on fitness function to be parents for mating. In GA, the tournament and roulette wheel methods are the two most popular selection methods used in the literature. The genetic crossover and mutation operators are then applied to create new offspring chromosomes used for the next generation. In the GA crossover, a crossover point in the parent chromosomes is arbitrarily chosen and then genes after that point are exchanged to produce new children. In the GA mutation, GA alters randomly the gene values in the offspring chromosome.

Over consecutive generations, the population iteratively evolves toward an optimal solution using selection, crossover, and mutation until the termination criterion is satisfied.

5. Proposed Methodology

This section provides a detailed description of the proposed hybrid filter genetic algorithm-based feature selection approach used for cancer classification in high-dimensional microarray datasets. As shown in Figure 1, the methodology includes three phases: collection of high-dimensional microarray data, training phase and classification phase.

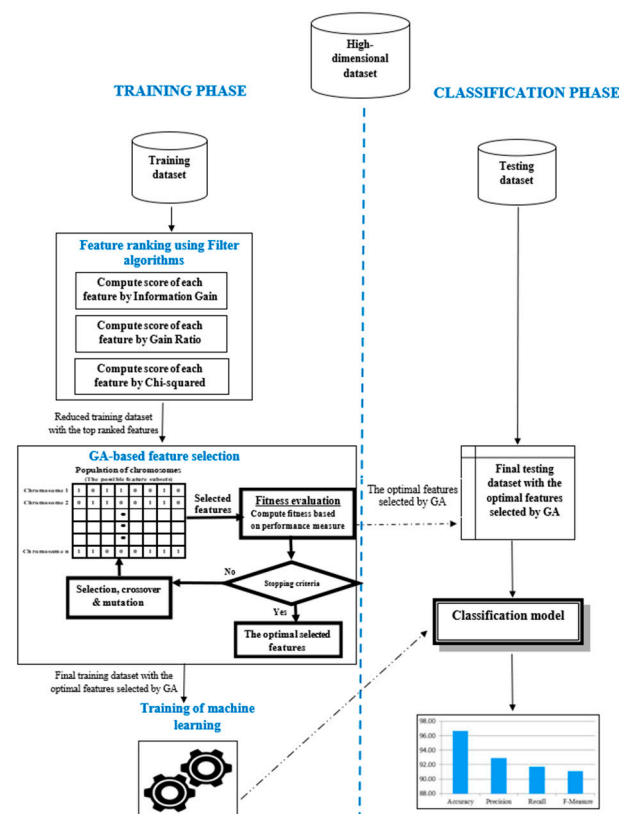


Figure 1. The methodology of the hybrid filter genetic algorithm-based feature selection approach proposed for cancer classification in high-dimensional microarray datasets.

5.1. Collection of High-Dimensional Microarray Data

In this paper, we used four high-dimensional cancerous microarray datasets to assess the performance of the proposed hybrid filter-GA feature selection method. These four datasets are Lung cancer [49], Central Nervous System (CNS) [49], Breast cancer [50], and

Brain cancer [51,52]. The description of the high-dimensional datasets used in this study is displayed in Table 1.

Table 1. Description of the high-dimensional microarray datasets used in this study.

Dataset	No. of Features	No. of Instances	No. of Classes
Breast	24,481	97	2
Lung	12,600	203	5
CNS	7129	60	2
Brain	5597	42	5

The Breast cancer dataset has 97 samples or instances including 24,481 features or genes. The Breast cancer dataset used in this study consists of 46 cancer samples that had cancer that spread in a different part or created distant metastases within 5 years, and 51 stayed free of distant metastasize for at least 5 years. The Lung cancer dataset has 203 samples with five classes and 12,600 features or genes. The samples in the Lung cancer dataset are labeled with normal lung class (17 samples) and four lung tumors classes: adenocarcinoma (139 samples), small cell lung cancer (6 samples), squamous cell carcinoma (21 samples), and pulmonary carcinoid (20 samples). The Brain tumor dataset has 42 microarray samples with 5597 features or genes and five classes. The five classes of the Brain tumor dataset are medulloblastomas, malignant gliomas, atypical teratoid/rhabdoid tumors, primitive neuroectodermal tumors, and human cerebella. The CNS cancer dataset has 60 with 7129 genes and two classes: 21 samples are survivors of cancer and 39 are failures. The datasets used in this study are then divided into parts: training dataset is used in the training phase while the testing dataset is used in classification (testing) phase.

5.2. Training Phase

The training phase consists of three main stages: feature ranking using filter algorithms, GA-based feature selection, and training of machine learning.

5.2.1. Feature Ranking Using Filter Algorithms

Since the microarray datasets used in this paper are high dimensional datasets with too many features, it is not applicable or time-consuming to use wrapper or evolutionary algorithms at the beginning as feature selection methods. So, it is an essential stage to decrease the high dimensional datasets using filter feature selection algorithms before applying evolutionary feature selection algorithms.

In this paper, information gain, gain ratio, and Chi-squared were applied as three common filter methods to compute a score of each feature. Then, only the top 5% of ranked features were chosen while the other irrelevant and redundant were eliminated to lower the high dimensional datasets.

- Information gain

The information gain (IG) is one of the popular filter techniques that was successfully applied to choose highly relevant features in order to reduce high-dimensional datasets in many applications. The IG uses the entropy measure to determine the relevance of features by calculating the information gain of features with respect to class labels. In the IG, Equation (1) is used to evaluate the features:

$$IG(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

where Value(A) represents the values set of a feature A, while S_v denotes the subset of S for which feature A has value v.

We needed to calculate $\Pr(c_j)$, which represents the probability of class c_j in S , to compute Entropy(S) as shown in Equation (2).

$$\text{Entropy}(S) = - \sum_{j=1}^{|C|} \Pr(c_j) \log_2 \Pr(c_j) \quad (2)$$

The IG is commonly used to identify the significance degree of a feature. However, IG may suffer from an overfitting problem due to it being biased towards features with many different values.

- Information gain ratio

The information gain ratio (IGR) was introduced to improve the performance of information gain by taking the number and size of branches into account when choosing an attribute to reduce its bias toward high-branch attributes.

IGR uses Equation (3) to evaluate the features:

$$\text{IGR}(S, A) = \frac{\text{IG}(S, A)}{\text{Split information}(S, A)} \quad (3)$$

Split information (S, A) is computed using Equation (4):

$$\text{Split information}(S, A) = - \sum_{i=1}^k \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (4)$$

where S and S_i represent the original dataset and the i th sub-dataset after being split while $|S|$ and $|S_i|$ are the numbers of samples belonging to S and S_i , respectively.

- Chi-squared

Chi-squared [53] is one of the simple and fast filter techniques which is used to determine the significant difference between features by examining the independence of data between two features. In general, Chi-squared (CS) computes the dependence between features and class. The null hypothesis for Chi-squared is tested by the χ^2 as shown in Equation (5) with the assumption that the feature and class label are independent. In the Chi-squared test, the summation of squared differences between observed and expected values is computed as shown in Equation (5). The importance of each feature was evaluated by calculating χ^2 with respect to the class. The feature with higher χ^2 was a more important feature for the classification decision.

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (5)$$

where O_{ij} and E_{ij} represent the observed frequency and expected frequency, respectively, while c is the class number and r is the number of bins used for the discretization of numerical features.

5.2.2. GA-Based Feature Selection

Although the filter methods can reduce the high dimensional training datasets, the performance of such methods was not sufficiently accurate since the features were evaluated based on certain criteria independently of a machine learning algorithm. Furthermore, most of the filter approaches evaluate features individually by finding the relationship between features and the class labels while they assume all features are independent. Therefore, GA was utilized to further optimize the selected features obtained from the filter methods to enhance the cancer classification performance.

GA is a global optimization searching algorithm that is effectively applied as a feature selection technique to identify the most significant features in many applications.

The feature selection based on GA is generally conducted by the following four key stages:

1. Chromosome encoding: GA population includes a set of chromosomes and denotes search space which represents all possible feature subsets. Each chromosome in the population represents a feature subset and it is encoded with a binary string containing m genes, where m is the number of available features. If the feature is selected, the gene will be encoded by one, otherwise, it will be represented by zero.
2. Population initialization: initially, GA generates arbitrarily an initial population of chromosomes that correspond to subsets of the potential attributes.
3. Fitness evaluation: GA evaluates the fitness of the individual chromosome by computing the fitness function of each individual chromosome. In the GA-based feature selection, the training dataset containing the features selected for a chromosome is utilized to train the machine learning technique and then GA calculates the classification accuracy, which is used as the fitness of that chromosome. In this step, GA tries to find the ideal subset of features that maximizes the machine learning performance.
4. Reproduction: like biological evolution, the fittest chromosomes are selected and recombined to reproduce and evolve better new chromosomes or solutions. In GA reproduction, three genetic operators are used in GA to perform the reproduction procedure:
 - Selection: the chromosomes that have better fitness values are chosen as parents to generate new children.
 - Crossover: in this process, GA exchanges the genes of two parent chromosomes after a crossover point chosen randomly in order to produce a new child chromosome.
 - Mutation: the GA mutation is performed by changing occasionally value of a gene for the child chromosome from 1 to 0 or from 0 to 1.

GA iteratively evolves the chromosomes to generate a different generation of better new solutions by repeating the fitness evaluation and reproduction process until GA meets one of the termination criteria such as obtaining satisfactorily optimal fitness or reaching maximum generations. Algorithm 1 shows the pseudocode of the hybrid filter genetic algorithm-based feature selection approach.

Algorithm 1: The pseudocode of the hybrid filter genetic algorithm-based feature selection approach

Input: F: Original feature set
 N: Size of population (Number of chromosomes)
Output: SF: The optimal selected features

- 1 **Begin**
- 2 Compute score of each feature in F using Information gain, Gain ratio, or Chi-squared
- 3 RF = Select only the top 5% of ranked features
- 4 D = Dimension of RF
- 5 Initialize population P by generating N chromosomes C including D genes(features) with random values [0, 1] for each gene g
 // Convert chromosomes to binary chromosomes (If the feature is selected, g = 1; otherwise, g = 0)
- 6 If $g \geq 0.5$ then g = 1; otherwise, g = 0
- 7 **While** termination criteria not meet **do**
- 8 Compute fitness value (classification accuracy) for each chromosome
- 9 Select two parents based on better fitness values
- 10 Perform Crossover
- 11 Perform Mutation
- 12 **End While**
- 13 Obtain the best chromosome C_{best}
- 14 Extract the optimal selected features SF from C_{best} (the genes with 1)
- 15 Return SF
- 16 **End Algorithm**

5.2.3. Training of Machine Learning Techniques

In this step, the reduced training dataset with the optimal features selected by GA was utilized to train some common machine learning algorithms for classifying cancer in high-dimensional microarray datasets. In this study, the support vector machine (SVM), naïve Bayes classifier (NB), k-Nearest neighbor (kNN), decision tree (DT), and random forest (RF) were chosen since they are commonly used in the literature to classify cancer in the high-dimensional microarray datasets. Then, we kept the trained classification models to be employed in the classification phase with the new testing datasets.

5.3. Classification Phase

In this phase, the classification models trained in the training phase were evaluated with a new dataset called the testing dataset. The initial testing dataset was reduced by selecting only the same top features ranked by filter algorithms in the training phase. Furthermore, the optimal features selected by GA in the training phase were then employed to select the substantial features of the testing dataset. Accordingly, the trained classification models were employed to classify cancer in the final testing dataset with the optimal feature subset and then their performances were evaluated using popular classification measures such as Classification Accuracy, Recall, Precision, and F-measure.

6. Experiments and Evaluation

6.1. Experimental Settings

We conducted many experiments to identify the best GA parameters. In this study, the best parameters used in the proposed hybrid filter-GA feature selection method were selected by a trial-and-error basis in order to produce the best results. Table 2 shows the settings of GA parameters used with the proposed hybrid filter-GA feature selection method on all experimental datasets.

Table 2. Settings of GA parameters used in the proposed hybrid filter-GA feature selection method on all experimental datasets.

GA Parameter	Value
Crossover rate	0.6
Mutation rate	0.02
Number of chromosomes	20
Number of generations	50
Selection scheme	Tournament (0.25)

6.2. Performance Metrics

In this study, 10-fold cross-validation was used to assess the hybrid filter-GA feature selection method proposed to enhance the performance of popular machine learning. The selected feature number and performance measures of the testing dataset were computed for each run in 10-fold cross-validation. Then, the overall selected attributes number and performance measures were the average for all runs.

In addition to the number of selected features, the Classification Accuracy, Recall, Precision, and F-Measure were used to measure the performance of the proposed hybrid filter-GA feature selection method. The Classification Accuracy, Recall, Precision, and F-Measure are briefly explained as follows:

Classification Accuracy is the percentage of instances correctly classified as shown in Equation (6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100(\%) \quad (6)$$

where TP indicates the number of positive instances correctly classified as positive instances, TN represents the number of negative instances correctly classified as negative instances, FP represents the number of the negative instances incorrectly classified as posi-

tive instances, and FN represents the number of positive instances incorrectly classified as negative instances.

The Recall is the percentage of positive instances correctly classified as belonging to the positive class as shown in Equation (7).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100(\%) \quad (7)$$

The Precision is the number of correctly classified positive instances divided by the total number of instances classified as positive as shown in Equation (8).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100(\%) \quad (8)$$

The F-Measure is the harmonic mean that combines both precision and recall as shown in Equation (9).

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100(\%) \quad (9)$$

6.3. Experimental Results and Discussion

6.3.1. Performance Comparison of Proposed Hybrid Filter-GA Feature Selection

In this section, the performances of machine learning techniques after applying the proposed hybrid filter-GA feature selection method were compared to the standalone machine learning techniques and the machine learning techniques by considering only filter feature selection methods.

Figures 2–5 and Tables 3–6 show the comparison of the classification results for machine learning techniques using all features, the features selected only by filters methods, and the features selected by the proposed hybrid filter-GA methods on four high-dimensional datasets: Brain, Breast Cancer, Lung, and CNS datasets. In the filter methods, the best classification results were achieved by training the machine learning techniques with the top 5% of ranked features on four datasets.

For the Brain dataset, Figure 2 and Table 3 show that the classification accuracies of SVM (69.05%), NB (69.05%), kNN (78.57%), DT (50%), and RF (78.57%) were enhanced by applying IG to 73.81%, 88.1%, 80.95%, 61.9%, and 90.48%, while enhanced by applying IGR to 78.57%, 85.71%, 83.33%, 64.29%, and 92.86%, and improved by applying CS to 83.33%, 83.33%, 80.95%, 69.05%, and 88.1, respectively. Furthermore, the proposed hybrid IG-GA feature selection method increased further the classification accuracies of SVM, NB, kNN, DT, and RF to 85.71%, 92.86%, 92.86%, 85.71%, and 100%, while they were enhanced by the proposed hybrid IGR-GA feature selection method to 97.62%, 95.24%, 97.62%, 88.1%, and 100%, respectively. In addition, Figure 2 and Table 3 show that the proposed hybrid CS-GA feature selection method increased further the classification accuracies of SVM, NB, kNN, DT, and RF to 97.62%, 95.24%, 97.62%, 85.71%, and 100%, respectively. In terms of Recall and Precision, the results shown in Table 3 demonstrate that SVM, NB, kNN, DT, and RF that applied the proposed hybrid filter-GA feature selection methods achieved better performance compared to the performance of the stand-alone classifiers or their performances with considering only filter algorithms. Consequently, SVM, NB, kNN, DT, and RF with considering the proposed hybrid filter-GA feature selection methods produced the best F-measure among other approaches since F-measure combines both precision and recall. It can be noticed also from Figure 2 and Table 3 that RF after applying the proposed hybrid IG-GA, IGR-GA and CS-GA methods accomplished the best performance among the classifiers with other feature selection methods.

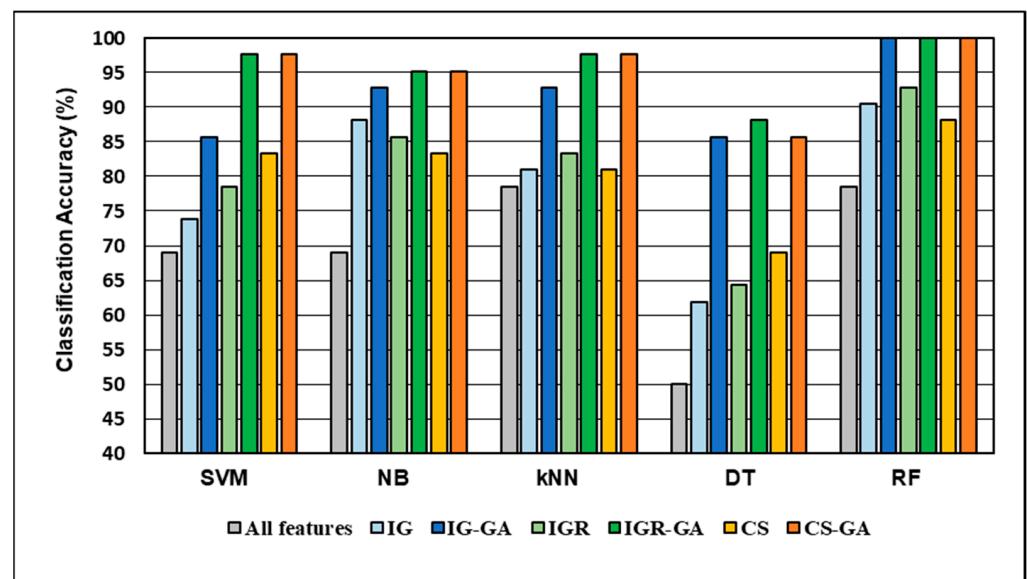


Figure 2. The accuracy comparison of classifiers using all features, the features selected only by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Brain dataset.

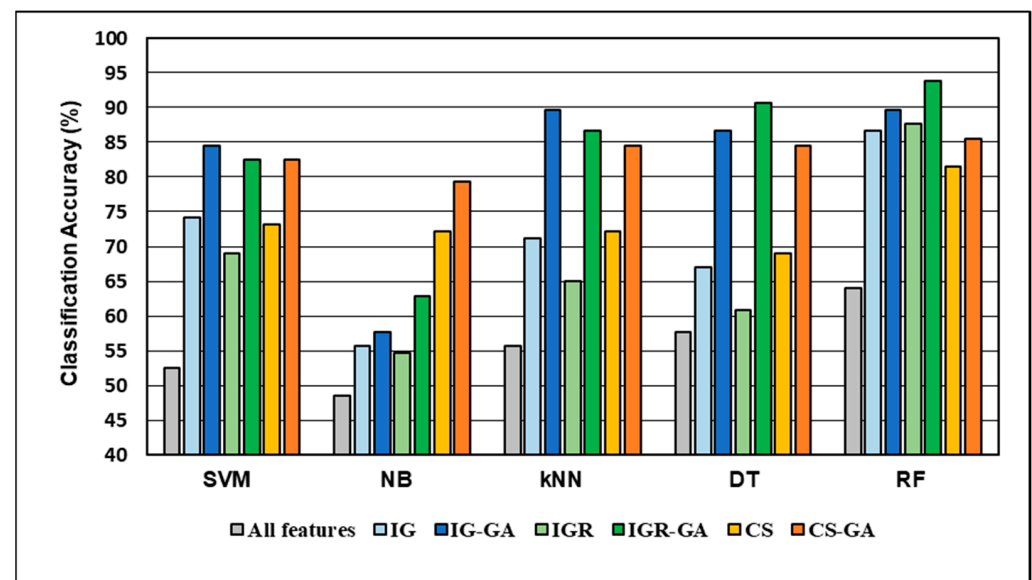


Figure 3. The accuracy comparison of classifiers using all features, the features selected only by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Breast Cancer dataset.

For Breast Cancer dataset, Figure 3 and Table 4 show that IG contributed to improving the classification accuracies of SVM (52.58%), NB (48.45%), kNN (55.67%), DT (57.73%) and RF (63.92%) to 74.23%, 55.67%, 71.13%, 67.01%, and 86.6%, while they were improved by applying IGR to 69.07%, 54.64%, 64.95%, 60.82%, and 87.63%, respectively. Furthermore, SVM, NB, kNN, DT, and RF were enhanced by applying CS to 73.2%, 72.16%, 72.16%, 69.07%, and 81.44%, respectively. Figure 3 and Table 4 also show that the SVM, NB, kNN, DT, and RF were enhanced further by the proposed hybrid IG-GA, IGR-GA, and CS-GA methods compared to using only filter algorithms. The classification accuracies of SVM, NB, kNN, DT, and RF were enhanced further by the proposed hybrid IG-GA method to 84.54%, 57.73%, 89.69%, 86.6%, and 89.69%, while improved by the proposed hybrid IGR-GA method to 82.47%, 62.89%, 86.6%, 90.72%, and 93.81%, and enhanced by the proposed hybrid CS-GA method to 82.47%, 79.38%, 84.54%, 84.54%, and 85.57%, respectively. In

addition to the classification accuracy, Table 4 shows the performance in terms of Recall and Precision, and F-measure of SVM, NB, kNN, DT, and RF before and after applying the proposed hybrid filter-GA feature selection methods. As can be observed from results in Table 4, Recall and Precision, and F-measure of SVM, NB, kNN, and DT were remarkably enhanced by applying the proposed hybrid filter-GA, compared to performances of the stand-alone classifiers or their performances with considering only filter algorithms. From Figure 3 and Table 4, we can observe also that RF and DT after employing the proposed hybrid IGR-GA method achieved the best performance among the classifiers that applied other feature selection methods.

For Lung Cancer dataset, Figure 4 and Table 5 demonstrate that the classification accuracies of SVM (78.82%), NB (90.15%), and RF (83.74%) were enhanced by applying IG to 92.12%, 95.07%, and 93.6%, while they were enhanced by applying IGR to 83.25%, 93.6%, and 91.13%, respectively. In addition, they are enhanced by applying CS to 84.24%, 92.12%, and 92.61%, respectively. Figure 4 and Table 5 also show that the performances of kNN and DT after applying IG, IGR, and CS were almost the same or slightly higher than the performances of the stand-alone kNN and DT. Compared to using only filter algorithms, the proposed hybrid IG-GA, IGR-GA, and CS-GA methods achieved substantially better classification results. The proposed hybrid IG-GA method increased further the classification accuracies of SVM, NB, kNN, DT, and RF to 94.09%, 98.52%, 97.04%, 96.55%, and 96.06%, while they were enhanced by applying the proposed hybrid IGR-GA method to 94.58%, 97.54%, 96.06%, 96.06%, and 95.57%, respectively. Furthermore, they were enhanced by applying the proposed hybrid CS-GA method to 95.07%, 97.04%, 95.57%, 96.55%, and 96.06%, respectively. In terms of Recall and Precision, and F-measure, Table 5 shows that SVM, NB, kNN, DT, and RF with applying the proposed hybrid filter-GA methods performed significantly better Recall and Precision, and F-measure compared to the stand-alone SVM, NB, kNN, DT and RF, and their performances with considering only filter algorithms. As can be seen also in Figure 4 and Table 5, NB classifier after applying the proposed hybrid IG-GA, IGR-GA and CS-GA, and kNN classifier after employing the proposed hybrid IG-GA method achieved the best performance among the classifiers that applied other feature selection methods.

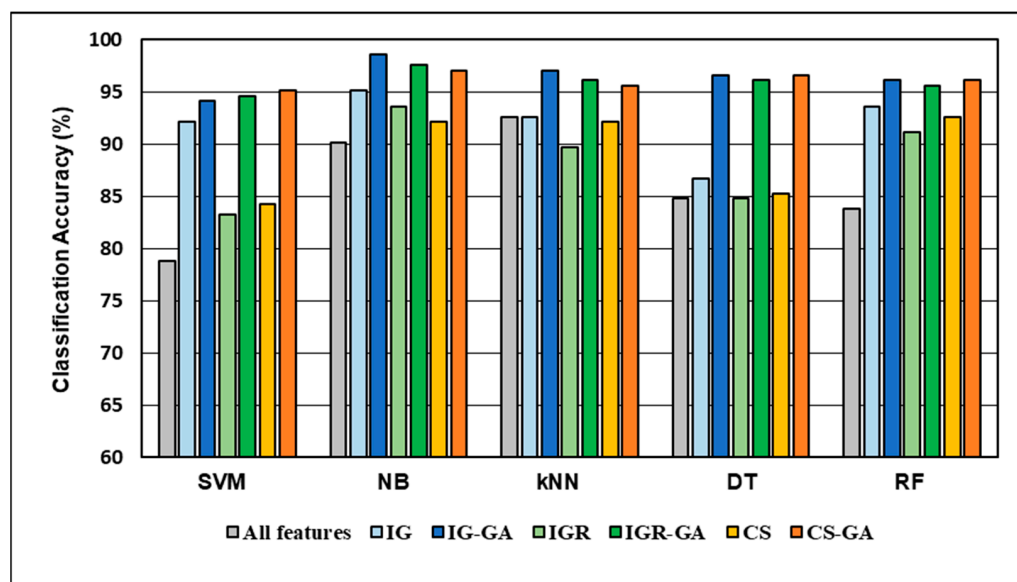


Figure 4. The accuracy comparison of classifiers using all features, the features selected only by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Lung Cancer dataset.

For the CNS dataset, Figure 5 and Table 6 demonstrate that IG contributed to improving the classification accuracies of NB (61.67%), kNN (61.67%), DT (58.33%), and RF (53.33%) to 75%, 75%, 70%, and 80% while they were improved by applying IGR to 78.33%, 68.33%, 68.33%, and 80%, respectively. Furthermore, they were enhanced by applying CS to 70%, 75%, 61.67%, and 83.33%, respectively. It can be observed that the performance of SVM was not enhanced by applying IG, IGR, or CS. Table 6 and Figure 5 also show that further improvements were conducted using the proposed hybrid IG-GA, IGR-GA, and CS-GA methods. The classification accuracies of SVM, NB, kNN, DT, and RF were enhanced further by the proposed hybrid IG-GA method to 86.67%, 90%, 93.33%, 93.33%, and 91.67%, while enhanced by the proposed hybrid IGR-GA method to 65%, 88.33%, 83.33%, 93.33%, and 90%, and enhanced by the proposed hybrid CS-GA method to 83.33%, 83.33%, 88.33%, 88.33%, and 88.33%, respectively. In addition to enhancing the classification accuracy, results in Table 6 demonstrate that Recall and Precision, and F-measure of SVM, NB, kNN, and DT were outstandingly enhanced by applying the proposed hybrid filter-GA, compared to performances of the stand-alone classifiers or their performances with considering only filter algorithms. As can be noticed also in Figure 5 and Table 6, the kNN and DT classifier after applying the proposed IG-GA, and DT classifier after applying IGR-GA methods accomplished the best performance among the classifiers that applied other feature selection methods.

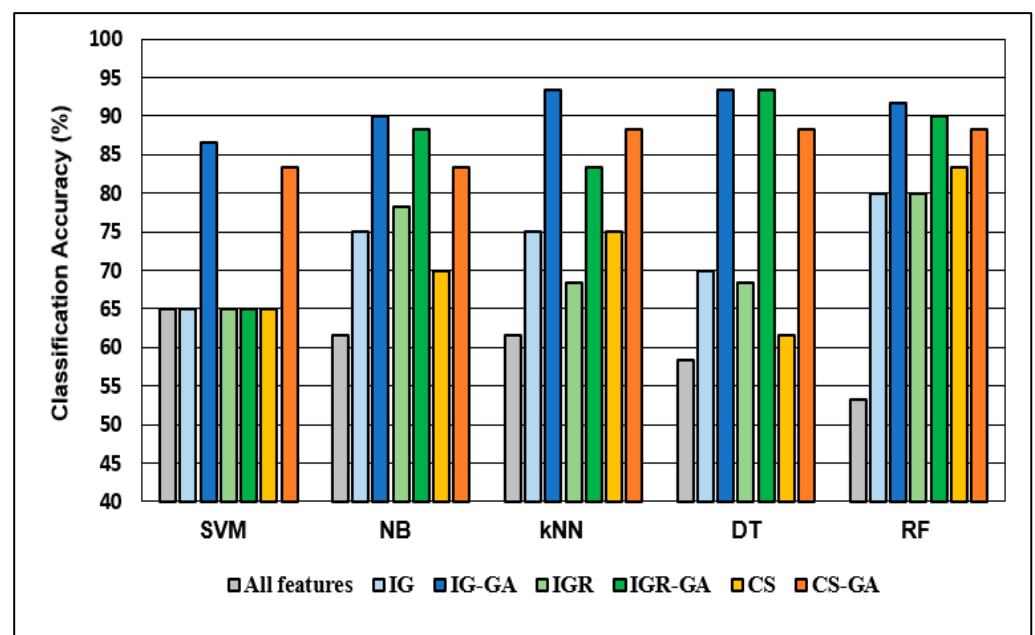


Figure 5. The accuracy comparison of classifiers using all features, the features selected only by filters methods, and the features selected by the proposed hybrid filter-GA methods on the CNS dataset.

Table 3. The performance comparison of classifiers using all features, the features selected by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Brain dataset.

	All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
SVM	Accuracy	69.05	73.81	85.71	78.57	97.62	83.33
	Recall	58	63	75	78	97.5	83
	Precision	42.28	64.05	68.57	66.52	98.18	88.41
	F-measure	48.91	63.52	71.64	71.8	97.84	85.62

Table 3. Cont.

		All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
NB	Accuracy	69.05	88.1	92.86	85.71	95.24	83.33	95.24
	Recall	60.5	80.5	88	78	93	76.5	93
	Precision	58.69	90.91	94.55	88.72	95.96	88.77	95.96
	F-measure	59.58	85.39	91.16	83.02	94.46	82.18	94.46
kNN	Accuracy	78.57	80.95	92.86	83.33	97.62	80.95	97.62
	Recall	75.5	80.5	92.5	83	97.5	80.5	97.5
	Precision	86.36	84.33	94.85	88.33	98.18	87.05	98.18
	F-measure	80.57	82.37	93.66	85.58	97.84	83.65	97.84
DT	Accuracy	50	61.9	85.71	64.29	88.1	69.05	85.71
	Recall	48.5	65	87	61.5	89.5	69	83.5
	Precision	53.06	61.73	87.22	63.83	88.5	69.89	88.01
	F-measure	50.68	63.32	87.11	62.64	89	69.44	85.7
RF	Accuracy	78.57	90.48	100	92.86	100	88.1	100
	Recall	76.5	88	100	93	100	88	100
	Precision	80.21	92.73	100	94.36	100	91.33	100
	F-measure	78.31	90.3	100	93.68	100	89.63	100

Table 4. The performance comparison of classifiers using all features, the features selected by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Breast dataset.

		All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
SVM	Accuracy	52.58	74.23	84.54	69.07	82.47	73.2	82.47
	Recall	50	73.89	84.34	68.67	82.27	73.34	82.48
	Precision	26.29	74.41	84.69	69.21	82.6	73.3	82.43
	F-measure	34.46	74.15	84.51	68.94	82.43	73.32	82.45
NB	Accuracy	48.45	55.67	57.73	54.64	62.89	72.16	79.38
	Recall	46.93	53.47	55.54	52.71	60.87	72.04	79.33
	Precision	45.32	62.94	70.63	56.23	79.31	72.09	79.33
	F-measure	46.11	57.82	62.18	54.41	68.88	72.06	79.33
kNN	Accuracy	55.67	71.13	89.69	64.95	86.6	72.16	84.54
	Recall	54.43	70.52	89.77	63.58	85.98	71.5	84.44
	Precision	55.78	71.93	89.67	69.3	88.89	73.25	84.53
	F-measure	55.1	71.22	89.72	66.32	87.41	72.36	84.48
DT	Accuracy	57.73	67.01	86.6	60.82	90.72	69.07	84.54
	Recall	57.25	66.71	86.3	60.51	90.43	68.88	84.34
	Precision	57.52	66.97	87.09	60.67	91.31	69	84.69
	F-measure	57.38	66.84	86.69	60.59	90.87	68.94	84.51
RF	Accuracy	63.92	86.6	89.69	87.63	93.81	81.44	85.57
	Recall	63.55	86.51	89.66	87.49	93.8	81.29	85.64
	Precision	63.85	86.6	89.66	87.71	93.8	81.48	85.54
	F-measure	63.7	86.55	89.66	87.6	93.8	81.38	85.59

Table 5. The performance comparison of classifiers using all features, the features selected by filters methods, and the features selected by the proposed hybrid filter-GA methods on the Lung dataset.

		All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
SVM	Accuracy	78.82	92.12	94.09	83.25	94.58	84.24	95.07
	Recall	41.13	71.6	75.68	52.8	86.47	55.15	90.03
	Precision	75.27	75.72	76.41	54.47	98.53	54.9	98.66
	F-measure	53.19	73.6	76.04	53.62	92.11	55.02	94.15
NB	Accuracy	90.15	95.07	98.52	93.6	97.54	92.12	97.04
	Recall	79.07	88.5	97.73	93.64	97.44	93.21	97.3
	Precision	88.21	94.36	98.54	88.9	93.92	84.92	93.05
	F-measure	83.39	91.34	98.13	91.21	95.65	88.87	95.13
kNN	Accuracy	92.61	92.61	97.04	89.66	96.06	92.12	95.57
	Recall	80.73	87.91	94.87	73.98	92.74	80.36	91.79
	Precision	95.22	89.1	95.4	93.1	97.78	95.09	97.65
	F-measure	87.38	88.5	95.13	82.45	95.19	87.11	94.63
DT	Accuracy	84.73	86.7	96.55	84.73	96.06	85.22	96.55
	Recall	69.07	73.69	94.68	69.07	92.2	72.4	93.15
	Precision	84.15	80.63	94.35	84.15	95.21	85.92	96.21
	F-measure	75.87	77	94.51	75.87	93.68	78.58	94.66
RF	Accuracy	83.74	93.6	96.06	91.13	95.57	92.61	96.06
	Recall	59.1	85.6	90.58	78.46	92.01	83.47	92.97
	Precision	93.54	97.15	97.86	94.45	97.73	96.82	97.86
	F-measure	72.43	91.01	94.08	85.72	94.78	89.65	95.35

Table 6. The performance comparison of classifiers using all features, the features selected by filters methods, and the features selected by the proposed hybrid filter-GA methods d on the CNS dataset.

		All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
SVM	Accuracy	65	65	86.67	65	65	65	83.33
	Recall	50	50	82.05	50	50	50	77.29
	Precision	32.5	32.5	88.89	32.5	32.5	32.5	86.58
	F-measure	39.39	39.39	85.33	39.39	39.39	39.39	81.67
NB	Accuracy	61.67	75	90	78.33	88.33	70	83.33
	Recall	59.52	74.18	87.91	75.64	87.73	69.23	81.68
	Precision	59.03	72.92	89.86	76.25	86.96	68	81.68
	F-measure	59.27	73.54	88.87	75.94	87.34	68.61	81.68
kNN	Accuracy	61.67	75	93.33	68.33	83.33	75	88.33
	Recall	54.03	71.98	91.58	61.36	78.39	69.78	84.43
	Precision	55.12	72.5	93.71	64.44	84.44	73.01	90.06
	F-measure	54.57	72.24	92.63	62.86	81.3	71.36	87.15

Table 6. Cont.

		All Features	IG	IG-GA	IGR	IGR-GA	CS	CS-GA
DT	Accuracy	58.33	70	93.33	68.33	93.33	61.67	88.33
	Recall	54.76	67.03	91.58	63.55	91.58	54.03	84.43
	Precision	54.67	67.03	93.71	64.68	93.71	55.12	90.06
	F-measure	54.71	67.03	92.63	64.11	92.63	54.57	87.15
RF	Accuracy	53.33	80	91.67	80	90	83.33	88.33
	Recall	45.42	75.82	89.19	72.53	85.71	78.39	85.53
	Precision	44.44	78.93	92.46	72.53	93.33	84.44	88.49
	F-measure	44.92	77.34	90.8	72.53	89.36	81.3	86.98

6.3.2. Comparison of Features Reduced by Applying Proposed Hybrid Filter-GA Methods

As mentioned in Section 6.1, the four datasets used in this study were high-dimensional datasets with too many features. So, it was a critical step to remove the irrelevant and redundant features, and then select only the optimal feature subsets that could be utilized in improving the performance of the machine learning techniques. In our experiments, the top 5% of features ranked by IG, IGR, and CS were selected to remove the redundant and irrelevant features in order to reduce the high dimensional datasets. Then, GA was utilized to further find the more significant and relevant features and eliminate less relevant features in order to maximize the performance of the machine learning techniques.

Table 7 shows the number of features selected by the proposed hybrid filter-GA feature selection method compared to considering only filter IG, IGR and CS methods on four datasets. As can be seen, the IG, IGR, and CS contributed to reducing the number of features of Brain, Breast Cancer, Lung, and CNS datasets from 5597, 24,481, 12,600, and 7129 features to 280, 1224, 630, and 356, respectively. Furthermore, for Brain data, only 135 significant features on average were selected by both the proposed hybrid IG-GA and hybrid IGR-GA methods while 139 important features on average were selected by the proposed hybrid CS-GA method. For the Breast Cancer dataset, only 617, 616, and 623 significant features on average were selected by the proposed hybrid IG-GA, IGR-GA, and CS-GA methods, respectively. For the Lung dataset, only 326, 318, and 320 influential features on average were selected by the proposed hybrid IG-GA, IGR-GA, and CS-GA methods, respectively. For the CNS dataset, only 183, 179, and 172 relevant features on average were selected by the proposed hybrid IG-GA, IGR-GA, and CS-GA methods, respectively.

Table 7. Comparison of the selected features numbers before and after applying the proposed hybrid filter-GA feature selection method.

		All Features	No. of Features after Applying IG	No. of Features after Applying IG-GA	No. of Features after Applying IGR	No. of Features after Applying IGR-GA	No. of Features after Applying CS	No. of Features after Applying CS-GA
Brain dataset	SVM			136		138		134
	NB			148		133		147
	kNN	5597	280	132	280	129	280	136
	DT			121		131		122
	RF			138		144		155
	Average			135		135		139

Table 7. Cont.

		All Features	No. of Features after Applying IG	No. of Features after Applying IG-GA	No. of Features after Applying IGR	No. of Features after Applying IGR-GA	No. of Features after Applying CS	No. of Features after Applying CS-GA
Breast dataset	SVM	24,481	1224	611	1224	624	1224	612
	NB			642		604		617
	kNN			621		613		641
	DT			602		623		606
	RF			610		618		639
	Average			617		616		623
Lung dataset	SVM	12,600	630	329	630	310	630	330
	NB			327		316		326
	kNN			323		329		327
	DT			313		323		296
	RF			336		310		321
	Average			326		318		320
CNS dataset	SVM	7129	356	194	356	178	356	157
	NB			168		181		183
	kNN			182		189		196
	DT			182		189		159
	RF			190		156		167
	Average			183		179		172

6.3.3. Comparison with Existing Works

In this section, the proposed hybrid filter-genetic feature selection methods are compared to several previous studies that applied filter-based and hybrid-based feature selection to reduce the dimensionality of the used four microarray datasets: Brain, Breast, Lung, and CNS datasets. Hameed et al. [35] suggested applying PCC-GA and PCC-BPSO in the microarray datasets that combined Pearson's Correlation Coefficient (PCC) with Genetic Algorithm (GA) or Binary Particle Swarm Optimization (BPSO). A fusion-based feature selection method was used by [36] on the microarray datasets to improve the effectiveness of cancer classification. Recently, the authors in [37] applied a hybrid feature selection method on these four microarray datasets. The method combined the Gini index and support vector machine with Recursive Feature Elimination (GI-SVM-RFE).

The performances of the proposed hybrid filter-genetic feature selection methods were compared against these related studies that used the same datasets. The accuracy comparison reported in Table 8 shows the superior performance of the proposed methods on most of the datasets used in this study. As can be seen in Table 8, the experimental results demonstrated that the proposed IG-GA, IGR-GA, and CS-GA methods outperformed the competitor methods for the five machine learning algorithms on most of the datasets used in this study.

Table 8. The accuracy comparison of the proposed hybrid filter-genetic feature selection methods with the previous studies.

		The Proposed Methods			GI-SVM-RFE [37]	Fusion [36]	PCC-GA [35]	PCC-BPSO [35]
		IG-GA	IGR-GA	CS-GA				
Brain dataset	SVM	85.71	97.62	97.62	N/A	95	97.62	97.62
	NB	92.86	95.24	95.24	88	N/A	90.48	92.86
	kNN	92.86	97.62	97.62	87.50	N/A	95.24	97.62
	DT	85.71	88.1	85.71	71.50	N/A	N/A	N/A
	RF	100	100	100	90	88.67	95.24	85.71
Breast dataset	SVM	84.54	82.47	82.47	N/A	75.11	88.66	90.72
	NB	57.73	62.89	79.38	90.67	N/A	85.57	88.66
	kNN	89.69	86.6	84.54	87.67	N/A	86.60	87.63
	DT	86.6	90.72	84.54	72.22	N/A	N/A	N/A
	RF	89.69	93.81	85.57	88.67	84.65	84.54	85.57
Lung dataset	SVM	94.09	94.58	95.07	N/A	N/A	97.54	97.04
	NB	98.52	97.54	97.04	91.17	N/A	97.04	98.03
	kNN	97.04	96.06	95.57	92.62	N/A	97.54	96.06
	DT	96.55	96.06	96.55	88.71	N/A	N/A	N/A
	RF	96.06	95.57	96.06	93.64	N/A	96.06	96.06
CNS dataset	SVM	86.67	65	83.33	N/A	75.00	98.33	91.94
	NB	90	88.33	83.33	85	N/A	90	91.94
	kNN	93.33	83.33	88.33	81.67	N/A	96.67	93.55
	DT	93.33	93.33	88.33	75	N/A	N/A	N/A
	RF	91.67	90	88.33	83.33	76.48	85.00	91.94

6.3.4. Discussion

Figures 2–5 and Tables 3–6 show almost all the standalone machine learning techniques trained with all features did not obtain good classification results on four datasets as these datasets suffer from the curse of dimensionality. This was expected since they trained with redundant and irrelevant features. Furthermore, Figures 2–5 and Tables 3–6 also show that IG, IGR, and CS filter methods contributed to improving the performance of SVM, NB, kNN, DT, and RF on most of the datasets used in this study. However, the performances of SVM, NB, kNN, DT, and RF with considering only IG, IGR, and CS were not good enough since the filter methods usually evaluate features independently of a classifier and ignore the relationship between features. On the other hand, the training of the classifier and the relationship between features are taken into consideration during the process of feature selection in the proposed hybrid filter-GA methods. Therefore, SVM, NB, kNN, DT, and RF with considering the proposed hybrid filter-GA feature selection methods achieved considerably better performance compared to the performance of the stand-alone classifiers or their performances with considering only filter algorithms.

The results in Table 7 show that the IG, IGR, and CS accomplished better classification results although we used only the top 5% of features. The IG, IGR, and CS contributed to reducing the number of features of Brain, Breast Cancer, Lung, and CNS datasets to 280, 1224, 630, and 356 features, respectively.

The results in Table 7 also demonstrate that the GA omitted about 50% of the irrelevant features from the datasets that were reduced by filter methods in the first step, and only fewer important attributes were used for training the machine learning techniques. Thus, after applying the proposed hybrid filter-GA feature selection methods, the classifiers accomplished better classification results with only a smaller number of features because the proposed hybrid IG-GA, IGR-GA, and CS-GA methods were capable of excluding redundant, irrelevant, and less relevant features.

From Tables 3–6, we can also observe that the machine learning that applied the proposed hybrid IG-GA, IGR-GA, and CS-GA methods achieved higher classification results in cancer datasets with smaller number of the selected features such as brain, CNS and lung dataset. On the other hand, the higher redundant and irrelevant features included in the breast data caused less improvements in classification results of the machine learning techniques used in this study.

The results in Table 8 show that the proposed IG-GA, IGR-GA, and CS-GA methods achieved better performance than previous studies that applied filter-based and hybrid-based feature selection for most of the microarray datasets. Table 8 also shows the proposed IG-GA, IGR-GA, and CS-GA methods performed well but they did not produce the best results in the CNS dataset among the compared other methods. However, the proposed IG-GA, IGR-GA, and CS-GA methods contributed to achieving better classification results of classifiers in CNS dataset as shown in Table 6 compared to stand-alone classifiers and considering only filter algorithms. Furthermore, the proposed IG-GA, IGR-GA, and CS-GA methods were effectively utilized in reducing the number of features and eliminating irrelevant and redundant features in the CNS dataset as shown in Table 7.

7. Conclusions and Future Work

To overcome the difficulties arising from the high-dimensional microarray datasets, this paper suggested a hybridization of filter feature selection methods and GA-based feature selection method. In the first phase of the proposed hybrid filter-genetic feature selection approach, the top 5% of features ranked by information gain, gain ratio, and Chi-squared are selected while the other redundant and irrelevant are eliminated to reduce the high dimensional microarray datasets. The cancer classification performances of the machine learning techniques with considering only filter feature selection methods are not sufficiently accurate since the features are evaluated based on certain criteria independently of a machine learning algorithm. Therefore, in the second phase of the proposed hybrid filter-genetic feature selection approach, the reduced datasets with only the top-ranked features selected by the filter methods are further optimized by the GA to achieve better cancer classification results. The experimental results demonstrated that the GA in the proposed methods omitted about 50% of irrelevant features, from datasets reduced by the filter methods in the first step, and only the remaining important features were used to maximize the cancer classification performance of the classifiers. In addition, the proposed hybrid filter-GA feature selection methods achieved significantly better performances compared to the performance of the stand-alone classifiers or their performances with considering only filter algorithms. Furthermore, the proposed hybrid filter-GA approach outperformed other existing feature selection methods on most of the high-dimensional microarray datasets used in this study. The future work of this research will focus on utilizing other filter feature selection methods with other evolutionary algorithms for further enhancement of cancer classification for high-dimensional microarray datasets.

Author Contributions: Conceptualization, W.A. and F.S.; methodology, W.A. and F.S.; software, W.A.; validation, W.A. and F.S.; formal analysis, W.A. and F.S.; investigation, W.A. and F.S.; resources, W.A.; data curation, F.S.; writing—original draft preparation, W.A.; writing—review and editing, W.A. and F.S.; project administration, W.A.; funding acquisition, W.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (G: 558-830-1441).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Available upon request.

Acknowledgments: This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (G: 558-830-1441). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

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

References

1. Hameed, S.S.; Petinrin, O.O.; Hashi, A.O.; Saeed, F. Filter-Wrapper Combination and Embedded Feature Selection for Gene Expression Data. *Int. J. Adv. Soft Comput. Appl.* **2018**, *10*, 90–105.
2. Hameed, S.S.; Hassan, R.; Muhammad, F.F. Selection and Classification of Gene Expression in Autism Disorder: Use of a Combination of Statistical Filters and a GBPSO-SVM Algorithm. *PLoS ONE* **2017**, *2*, e0187371. [[CrossRef](#)] [[PubMed](#)]
3. Afolabi, L.T.; Saeed, F.; Hashim, H.; Petinrin, O.O. Ensemble Learning Method for the Prediction of New Bioactive Molecules. *PLoS ONE* **2018**, *13*, e0189538. [[CrossRef](#)]
4. Anbarasi, M.; Anupriya, E.; Iyengar, N.C.S.N. Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. *Int. J. Eng. Sci. Technol.* **2010**, *2*, 5370–5376.
5. Srinivas, K.; Rani, B.; Govrdhan, A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *Int. J. Comput. Sci. Eng.* **2010**, *2*, 250–255.
6. Soni, S.; Vyas, O.P. Using Associative Classifiers for Predictive Analysis in Health Care Data Mining. *Int. J. Comput. Appl.* **2010**, *4*, 33–37. [[CrossRef](#)]
7. Rajkumar, A.; Reena, G.S. Diagnosis Of Heart Disease Using Datamining Algorithm. *Glob. J. Comput. Sci. Technol.* **2010**, *5*, 1678–1680.
8. Bolón-Canedo, V.; Sánchez-Marroño, N.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. A Review of Microarray Datasets and Applied Feature Selection Methods. *Inf. Sci.* **2014**, *282*, 111–135. [[CrossRef](#)]
9. Cosma, G.; Brown, D.; Archer, M.; Khan, M.; Pockley, A.G. A Survey on Computational Intelligence Approaches for Predictive Modeling in Prostate Cancer. *Expert Syst. Appl.* **2017**, *70*, 1–19. [[CrossRef](#)]
10. Singh, R.K.; Sivabalakrishnan, M. Feature Selection of Gene Expression Data for Cancer Classification: A Review. *Procedia Comput. Sci.* **2015**, *50*, 52–57. [[CrossRef](#)]
11. Wang, L. Feature Selection in Bioinformatics. In *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X*; SPIE: Bellingham, WA, USA, 2012.
12. Song, Q.; Ni, J.; Wang, G. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 1–14. [[CrossRef](#)]
13. Saeys, Y.; Inza, I.; Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [[CrossRef](#)] [[PubMed](#)]
14. Wang, A.; Liu, H.; Yang, J.; Chen, G. Ensemble Feature Selection for Stable Biomarker Identification and Cancer Classification from Microarray Expression Data. *Comput. Biol. Med.* **2022**, *142*, 105208. [[CrossRef](#)]
15. Liu, S.; Xu, C.; Zhang, Y.; Liu, J.; Yu, B.; Liu, X.; Dehmer, M. Feature Selection of Gene Expression Data for Cancer Classification Using Double RBF-Kernels. *BMC Bioinform.* **2018**, *19*, 1–14. [[CrossRef](#)]
16. Taveira De Souza, J.; Carlos De Francisco, A.; Macedo, D.C. De Dimensionality Reduction in Gene Expression Data Sets. *IEEE Access* **2019**, *7*, 61136–61144. [[CrossRef](#)]
17. Bolón-Canedo, V.; Sánchez-Marroño, N.; Alonso-Betanzos, A. Distributed Feature Selection: An Application to Microarray Data Classification. *Appl. Soft Comput. J.* **2015**, *30*, 136–150. [[CrossRef](#)]
18. Bhui, N. Ensemble of Deep Learning Approach for the Feature Selection from High-Dimensional Microarray Data. In *Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences, Kurukshetra, India, 7–9 May 2021*; Springer: Berlin/Heidelberg, Germany, 2022.
19. Alhenawi, E.; 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](#)]
20. Abdulla, M.; Khasawneh, M.T. G-Forest: An Ensemble Method for Cost-Sensitive Feature Selection in Gene Expression Microarrays. *Artif. Intell. Med.* **2020**, *108*, 101941. [[CrossRef](#)]
21. Tao, P.; Sun, Z.; Sun, Z. An Improved Intrusion Detection Algorithm Based on GA and SVM. *IEEE Access* **2018**, *6*, 13624–13631. [[CrossRef](#)]
22. Ghareb, A.S.; Bakar, A.A.; Hamdan, A.R. Hybrid Feature Selection Based on Enhanced Genetic Algorithm for Text Categorization. *Expert Syst. Appl.* **2016**, *49*, 31–47. [[CrossRef](#)]
23. Ali, W.; Malebary, S. Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection. *IEEE Access* **2020**, *8*, 116766–116780. [[CrossRef](#)]
24. Ali, W.; Ahmed, A.A. Hybrid Intelligent Phishing Website Prediction Using Deep Neural Networks with Genetic Algorithm-Based Feature Selection and Weighting. *IET Inf. Secur.* **2019**, *13*, 659–669. [[CrossRef](#)]
25. Almutiri, T.; Saeed, F. Review on Feature Selection Methods for Gene Expression Data Classification. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2020.

26. Shah, S.H.; Iqbal, M.J.; Ahmad, I.; Khan, S.; Rodrigues, J.J.P.C. Optimized Gene Selection and Classification of Cancer from Microarray Gene Expression Data Using Deep Learning. *Neural Comput. Appl.* **2020**, *1*–12. [[CrossRef](#)]
27. Parhi, P.; Bisoi, R.; Dash, P.K. Influential Gene Selection From High-Dimensional Genomic Data Using a Bio-Inspired Algorithm Wrapped Broad Learning System. *IEEE Access* **2022**, *10*, 49219–49232. [[CrossRef](#)]
28. Kourou, K.; Rigas, G.; Papaloukas, C.; Mitsis, M.; Fotiadis, D.I. Cancer Classification from Time Series Microarray Data through Regulatory Dynamic Bayesian Networks. *Comput. Biol. Med.* **2020**, *116*, 103577. [[CrossRef](#)]
29. Saeid, M.M.; Nossair, Z.B.; Saleh, M.A. A Microarray Cancer Classification Technique Based on Discrete Wavelet Transform for Data Reduction and Genetic Algorithm for Feature Selection. In Proceedings of the Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020, Tirunelveli, India, 15–17 June 2020.
30. Passi, K.; Nour, A.; Jain, C.K. Markov Blanket: Efficient Strategy for Feature Subset Selection Method for High Dimensional Microarray Cancer Datasets. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, Kansas City, MO, USA, 13–16 November 2017.
31. Sayed, S.; Nassef, M.; Badr, A.; Farag, I. A Nested Genetic Algorithm for Feature Selection in High-Dimensional Cancer Microarray Datasets. *Expert Syst. Appl.* **2019**, *121*, 233–243. [[CrossRef](#)]
32. Ghosh, M.; Adhikary, S.; Ghosh, K.K.; Sardar, A.; Begum, S.; Sarkar, R. Genetic Algorithm Based Cancerous Gene Identification from Microarray Data Using Ensemble of Filter Methods. *Med. Biol. Eng. Comput.* **2019**, *57*, 159–176. [[CrossRef](#)]
33. Abasabadi, S.; Nematzadeh, H.; Motameni, H.; Akbari, E. Hybrid Feature Selection Based on SLI and Genetic Algorithm for Microarray Datasets. *J. Supercomput.* **2022**, *78*, 19725–19753. [[CrossRef](#)]
34. Xie, W.; Fang, Y.; Yu, K.; Min, X.; Li, W. MFRAG: Multi-Fitness RankAggreg Genetic Algorithm for Biomarker Selection from Microarray Data. *Chemom. Intell. Lab. Syst.* **2022**, *226*, 104573. [[CrossRef](#)]
35. Hameed, S.S.; Muhammad, F.F.; Hassan, R.; Saeed, F. Gene Selection and Classification in Microarray Datasets Using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers. *J. Comput. Sci.* **2018**, *14*, 868–880. [[CrossRef](#)]
36. Almutiri, T.; Saeed, F.; Alassaf, M.; Hezzam, E.A. A Fusion-Based Feature Selection Framework for Microarray Data Classification. In *Lecture Notes on Data Engineering and Communications Technologies*; Springer: Berlin/Heidelberg, Germany, 2021.
37. Almutiri, T.; Saeed, F. A Hybrid Feature Selection Method Combining Gini Index and Support Vector Machine with Recursive Feature Elimination for Gene Expression Classification. *Int. J. Data Min. Model. Manag.* **2022**, *14*, 41–62. [[CrossRef](#)]
38. Almugren, N.; Alshamlan, H. A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification. *IEEE Access* **2019**, *7*, 78533–78548. [[CrossRef](#)]
39. Aziz, R.; Verma, C.K.; Srivastava, N. A Novel Approach for Dimension Reduction of Microarray. *Comput. Biol. Chem.* **2017**, *71*, 161–169. [[CrossRef](#)] [[PubMed](#)]
40. Jain, I.; Jain, V.K.; Jain, R. Correlation Feature Selection Based Improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification. *Appl. Soft Comput.* **2018**, *62*, 203–215. [[CrossRef](#)]
41. Alshamlan, H.; Badr, G.; Alohal, Y. MRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling. *Biomed Res. Int.* **2015**, *2015*, 604910. [[CrossRef](#)]
42. Vafae Sharbaf, F.; Mosafer, S.; Moattar, M.H. A Hybrid Gene Selection Approach for Microarray Data Classification Using Cellular Learning Automata and Ant Colony Optimization. *Genomics* **2016**, *107*, 231–238. [[CrossRef](#)]
43. Dashtban, M.; Balafar, M. Gene Selection for Microarray Cancer Classification Using a New Evolutionary Method Employing Artificial Intelligence Concepts. *Genomics* **2017**, *109*, 91–107. [[CrossRef](#)]
44. Lu, H.; Chen, J.; Yan, K.; Jin, Q.; Xue, Y.; Gao, Z. A Hybrid Feature Selection Algorithm for Gene Expression Data Classification. *Neurocomputing* **2017**, *256*, 56–62. [[CrossRef](#)]
45. Vijay, S.A.A.; GaneshKumar, P. Fuzzy Expert System Based on a Novel Hybrid Stem Cell (HSC) Algorithm for Classification of Micro Array Data. *J. Med. Syst.* **2018**, *42*, 61. [[CrossRef](#)]
46. Hancer, E.; Xue, B.; Zhang, M. Differential Evolution for Filter Feature Selection Based on Information Theory and Feature Ranking. *Knowl.-Based Syst.* **2018**, *140*, 103–119. [[CrossRef](#)]
47. Holland, J.H. *Adaption in Natural and Artificial Systems*; The University of Michigan: Ann Arbor, MI, USA, 1975.
48. Kawamura, A.; Chakraborty, B. A Hybrid Approach for Optimal Feature Subset Selection with Evolutionary Algorithms. In Proceedings of the 2017 IEEE 8th International Conference on Awareness Science and Technology, iCAST 2017, Taichung, Taiwan, 8–10 November 2017.
49. Li, J.; Liu, H. Kent Ridge Biomedical Data Set Repository. Available online: <http://sdmc-lit.org.sg/GEDatasets> (accessed on 15 December 2020).
50. Van't Veer, L.J.; Dai, H.; Van de Vijver, M.J.; He, Y.D.; Hart, A.A.M.; Mao, M.; Peterse, H.L.; Van Der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* **2002**, *415*, 530–536. [[CrossRef](#)] [[PubMed](#)]
51. Pomeroy, S.L.; Tamayo, P.; Gaasenbeek, M.; Sturla, L.M.; Angelo, M.; McLaughlin, M.E.; Kim, J.Y.H.; Goumnerova, L.C.; Black, P.M.; Lau, C.; et al. Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature* **2002**, *15*, 436–442. [[CrossRef](#)] [[PubMed](#)]

52. Whitehead Institute Center for Genomic Research Cancer Genomics. Available online: <http://www-genome.wi.mit.edu/cancer> (accessed on 15 November 2022).
53. Liu, H.; Setiono, R. Chi2: Feature Selection and Discretization of Numeric Attributes. In Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, 5–8 November 1995.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.