processes

MDPI

Article

Feature Selection of Microarray Data Using Simulated Kalman Filter with Mutation

Nurhawani Ahmad Zamri, Nor Azlina Ab. Aziz *, Thangavel Bhuvaneswari, Nor Hidayati Abdul Aziz and Anith Khairunnisa Ghazali

Faculty of Engineering & Technology, Multimedia University, Melaka 75450, Malaysia; nurhawaniahmadzamri@gmail.com (N.A.Z.); t.bhuvaneswari@mmu.edu.my (T.B.); hidayati.aziz@mmu.edu.my (N.H.A.A.); anith.ghazali@mmu.edu.my (A.K.G.)
* Correspondence: azlina.aziz@mmu.edu.my

Abstract: Microarrays have been proven to be beneficial for understanding the genetics of disease. They are used to assess many different types of cancers. Machine learning algorithms, like the artificial neural network (ANN), can be trained to determine whether a microarray sample is cancerous or not. The classification is performed using the features of DNA microarray data, which are composed of thousands of gene values. However, most of the gene values have been proven to be uninformative and redundant. Meanwhile, the number of the samples is significantly smaller in comparison to the number of genes. Therefore, this paper proposed the use of a simulated Kalman filter with mutation (SKF-MUT) for the feature selection of microarray data to enhance the classification accuracy of ANN. The algorithm is based on a metaheuristics optimization algorithm, inspired by the famous Kalman filter estimator. The mutation operator is proposed to enhance the performance of the original SKF in the selection of microarray features. Eight different benchmark datasets were used, which comprised: diffuse large b-cell lymphomas (DLBCL); prostate cancer; lung cancer; leukemia cancer; "small, round blue cell tumor" (SRBCT); brain tumor; nine types of human tumors; and 11 types of human tumors. These consist of both binary and multiclass datasets. The accuracy is taken as the performance measurement by considering the confusion matrix. Based on the results, SKF-MUT effectively selected the number of features needed, leading toward a higher classification accuracy ranging from 95% to 100%.

Keywords: feature selection; simulated Kalman filter; microarray data; classification; mutation

Citation: Ahmad Zamri, N.; Ab. Aziz, N.A.; Bhuvaneswari, T.; Abdul Aziz, N.H.; Ghazali, A.K. Feature Selection of Microarray Data Using Simulated Kalman Filter with Mutation. *Processes* **2023**, *11*, 2409. https://doi.org/10.3390/pr11082409

Academic Editor: Bonglee Kim

Received: 13 June 2023
Revised: 24 July 2023
Accepted: 31 July 2023
Published: 10 August 2023

1. Introduction

Over the several last decades, gene expression has received increasing attention from scientists due to the development and advancements in DNA microarray technology. DNA microarray technology allows us to measure the expression levels of a large number of genes simultaneously, which is proven to be beneficial when understanding disease genetics, and have led to the development of new drugs and therapies. Cancer research has benefited from the unique experimental capabilities of this technology. Microarray data can be used to assess many different types of cancer.

Another alternative in cancer study is to use the spatial transcriptomics (ST) data that capture the expression levels of genes in individual cells in a tissue. An adaptive graph model method named spaCI was recently proposed in [1] to infer cell–cell interactions from ST data. However, ST data are more expensive and time consuming to collect compared to microarray data.

The advancement of artificial intelligence, particularly machine learning, has eased the process of analyzing data, including microarray data. Countless research works have been reported that utilize machine learning algorithms to analyze microarray data and classify whether a sample is cancerous or non-cancerous [2–5].

The DNA microarray data usually contain thousands of genes. But, numerous studies have shown that most genes measured in DNA microarray experiments do not contribute to the accuracy of classification [6]. The complexity of the problem arises from the massive number of features, but the low number of samples. Therefore, selecting highly discriminative genes from the gene expression data can improve the performance of cancer classification and reduce the cost of medical diagnosis [7]. Thus, there is a need for feature selection that allows for a high predictive accuracy and better diagnostic performance.

Many existing feature selection algorithms use metaheuristics. Metaheuristics are general-purpose algorithms that can be applied to solve any optimization problem. Unlike exact methods, metaheuristics can tackle a large number of problems by trading the best with satisfactory solutions within a reasonable time. The application of metaheuristics includes various fields such as very-large-scale integration (VLSI) [8], aerodynamics [9], telecommunications [10], automotive [11], and robotics [12]. Furthermore, metaheuristics are also used in machine learning and data mining in bioinformatics [13] and computational biology [14]. Their usage in many applications shows their efficiency and effectiveness in solving complex problems. Among popular metaheuristics algorithms are particle swarm optimization (PSO), artificial bee colony (ABC), gravitational search algorithm (GSA), and firefly algorithm (FA). These are nature-inspired metaheuristics. Some researchers have proposed metaheuristics that are non-nature-inspired, but are rather inspired by other processes such as estimation-based algorithms like the simulated Kalman filter (SKF) [15].

SKF is a metaheuristic optimization algorithm introduced in 2015 [15]. It has been applied in various applications, including airport gate allocation problems [16], adaptive beamforming in wireless communication [17], PCB drill path optimization problems [18], and feature selection for the peak classification of EEG signals [19]. All these works suggested that SKF is an excellent global optimizer.

However, SKF has not been applied for the feature selection of DNA microarray data. Therefore, this work proposes an SKF-based solution to search for an optimal subset of features for DNA microarray data. This work is motivated by the no free lunch (NFL) theorem [20]. The theorem states that no single optimization algorithm can optimally solve all types of optimization problems. A particular metaheuristic algorithm may present good results for an optimization problem, while other metaheuristic algorithms may provide good results for other problems. Hence, this work investigates the performance of SKF as a feature selector of cancer microarray data. Inspired by the NFL, a new variant of SKF with mutation operator (SKF-MUT) is proposed to enhance the performance of the original SKF algorithm in the feature selection of microarray data. Eight microarray data are used to evaluate the proposed solutions. The data consist of binary and multiclass data, whereby the highest number of classes is 11. The results show that SKF is able to achieve 100% accuracy for six out of eight data, the accuracy achieved for the data with nine classes is 95%, while the data with 11 classes achieved 97.7%. Meanwhile, the SKF-MUT is able to improve the accuracy of the 11 classes to 98.3% and maintain the same best accuracy for the other data.

Related works are reviewed in Section 2. Section 3 introduces the original SKF algorithm, the agents' encoding in solving feature selection problems for microarray data, and the concept of mutation as well as how it is used to enhance the SKF algorithm's performance. The experimental findings are presented and discussed in Section 4, followed by benchmarking with other works that have used the same dataset. Finally, Section 5 concludes the findings.

2. Related Works

Feature selection problem is a problem of identifying a subset of informative features from all the features available. This is important especially in the development of the classification system of microarray data. It reduces the problem of overfitting due to the dimensionality issue of microarray [21]. Thus, this field of research has been popularly explored by researchers due to its importance toward building effective and efficient systems.

PSO-based feature selection is proposed in [22]. The accuracy of the features selected by PSO is measured using decision tree (DT), support vector machine (SVM), k nearest neighbor (k-NN), naïve Bayes (NB), and ensemble classifiers. The findings show that the addition of feature selection improved the accuracy of all classifiers applied.

The NFL theory encouraged many researchers to work on improved variants of metaheuristic algorithms for feature selection. A modified ABC with cuckoo search (CS) is proposed in [23] for the feature selection of cancer microarray data. The CS is incorporated into the onlooker phase of the ABC to improve the exploitation for better feature selection. In addition to feature selection, the work also employed feature reduction using independent component analysis (ICA) prior to the feature selection. NB and SVM are used as the classifiers. The results show that the proposed algorithm is able to improve the unbiased accuracy as well as reduce the number of features, and the application of the proposed algorithm with NB is better than SVM for most of the data tested. A comparison of the proposed algorithm performance with several other metaheuristics, namely PSO, GA, original CS, original ABC, and genetic bee colony (GBC), shows that the proposed algorithm is the best for five out of six data. ABC is also chosen by the researchers of [24]. The ABC selected the best subset of features after the feature reduction phase using ICA algorithm. The artificial neural network (ANN) is used as the classifier. From the six microarray datasets tested, the proposed ICA+ABC is able to provide the best result for five datasets. Meanwhile, the concept of altruism is introduced to whale optimization algorithm (WOA) in [25]. The altruism WOA (AltWOA) give chance to a less fit solution with potential, at the scarification of a fitter solution. The selected features by AltWOA are found to contribute to a better classification accuracy in comparison to the original WOA and 10 other metaheuristics.

GA variants have been popular among researchers in this field. This is observed by the authors of [26]. One of the work that used GA for feature selection is [27], where nested GA is proposed. The nested GA uses two layers of GA, inner and outer GA, to select features of two different types of microarray data. The outer GA selects from gene expression data, while inner GA selects the features from DNA methylation data. The nested GAs update each other based on their best selected features. The accuracy based on the features selected by the outer GA is measured using SVM, while the inner GA is evaluated using deep neural network (DNN). This work also used *t*-test in the pre-processing stage. In [24], GA is applied after feature filtering using either information gain, information gain ratio, or chi square to select the most optimal subset of features from the reduced subset. The proposed method is evaluated using several machine learning classifiers; SVM, NB, kNN, DT, and random forest (RF). The application of GA is found to be able to reduce 50% of the irrelevant features.

From the existing works reviewed here, it can be seen that feature selection is an important phase in development of a microarray classification system. The feature selection is paired with a variety of machine learning classifiers like the SVM, ANN, NB, DT, kNN, and others. In all of the works reviewed, the incorporation of feature selection is observed to be able to improve classification accuracy. A variety of metaheuristic algorithms had been used by researchers for feature selection. The algorithms are either in their original form or the improved variants. Thus, this shows that there is room to explore the application of new metaheuristics for the feature selection of microarray data. In [28], the authors proposed the Kalman filter (KF) for the pre-processing of microarray-based molecular diagnosis to eliminate noise from the data and improve the compatibility with classification algorithms for better accuracy. None of the existing research, however, reported the application of SKF for microarray feature selection. Based on the successful application of SKF in other optimization problems, this work proposes and investigates its application for microarray feature selection.

3. Methodology

Metaheuristic algorithms are commonly inspired by nature, where a collection of agents collaborate to find the optimum solution, and their interaction mimics the natural phenomenon. On the contrary, some researchers look away from nature for their source of inspiration, such as the heuristic Kalman algorithm (HKA) [29], single-agent finite impulse response optimizer (SAFIRO) [30], and SKF [15]. These algorithms explicitly consider the optimization problem as a measurement process intended to produce an estimate of the optimum [15].

3.1. SKF for Microarray Feature Selection

SKF is a metaheuristic optimization algorithm inspired by the estimation capability of the Kalman filter. However, instead of relying on the properties of Gaussian distribution as in HKA, SKF simulates the measurement process as an individual agent's update mechanism in estimating the optimum without being tied to any distribution. There are five steps in SKF. The algorithm starts with the random generation of an initial population. Then, the fitness of each agent is evaluated. Based on the fitness evaluation, the best agent at the current iteration ($X_{best}\ (t)$) and the best agent for the search (X_{true}) are updated. Next, the Kalman filtering steps of prediction, measurement, and estimation are conducted before evaluating the stopping criterion. Finally, these steps are repeated until the stopping criterion is met.

In the initialization phase, the parameters of SKF are initialized. In addition to the number of agents and the maximum number of iteration, there are several other parameters to be initialized in the original SKF, namely the error covariance estimate, the process noise value, and the measurement noise value. Since parameter tuning can be tedious and the process itself can be considered an optimization problem, the work in [31] proposed the three parameters to be replaced with any random value between 0 and 1 from a normal distribution in every dimension. This approach is able to reach on-par performance as the original SKF algorithm. The initial agents' state values, $X_i\ (t)$, are also initialized during the initialization phase. These values are randomly initialized according to the problem to be optimized. The initialization phase is conducted only once at the beginning of the experiment.

After the initialization phase, the iterative procedure of SKF starts with the fitness evaluation. First, the fitness of the state values is evaluated according to the problem to be optimized. Then, the $X\_best\ (t)$ is identified. For a maximization problem, the $X\_best\ (t)$ is the state value of the agent with the highest fitness in the corresponding iteration. While for the minimization problem, the $X\_best\ (t)$ takes the lowest fitness. The $X\_best\ (t)$ with the best fitness value starting from the first iteration is chosen as the best solution so far, X_true.

The next step is obeying the Kalman filtering method of state prediction, measurement, and estimation. The state prediction uses the equation:

$$X_i(t|t-1) = X_i(t-1) \tag{1}$$

where $X_i\ (t-1)$ and $X_i(t|t-1)$ are the previous and transition states, respectively. The subscript i represents agent number. Meanwhile, the measurement is simulated based on the following equation:

$$Z_i(t) = X_i(t|t-1) + sin(rand \times 2\pi) \times |X_i(t|t-1) - X_{true}| \tag{2}$$

The stochastic element of SKF is incorporated in the simulated measurement equation through the random term $rand$, which is a uniformly distributed random number between 0 and 1. The sine function balances the exploration and exploitation to the measurement value.

In the estimation state, the state estimation is evaluated using the equation:

$$X_i(t) = X_i(t|t-1) + K(t) \times (Z_i(t) - X_i(t|t-1)) \tag{3}$$

The estimation equation performs further exploitation weighted by the Kalman gain *K(t)*, which is calculated using Equation (4):

$$K_i(t) = \frac{P_i(t|t-1)}{P_i(t|t-1) + randn_i} \tag{4}$$

The predicted error covariant estimate, $P(t \mid t-1)$, and the previous error covariant estimates, $P(t-1)$, are updated based on Equations (5) and (6):

$$P_i(t|t-1) = P_i(t-1) + randn_i \tag{5}$$

$$P_i(t) = (1 - K_i(t)) \times P_i(t|t-1) \tag{6}$$

Finally, the stopping criteria are evaluated. As long as the stopping criteria (maximum number of iterations) are not met, the iterative procedure of SKF is repeated. The flowchart of the SKF algorithm is shown in Figure 1.

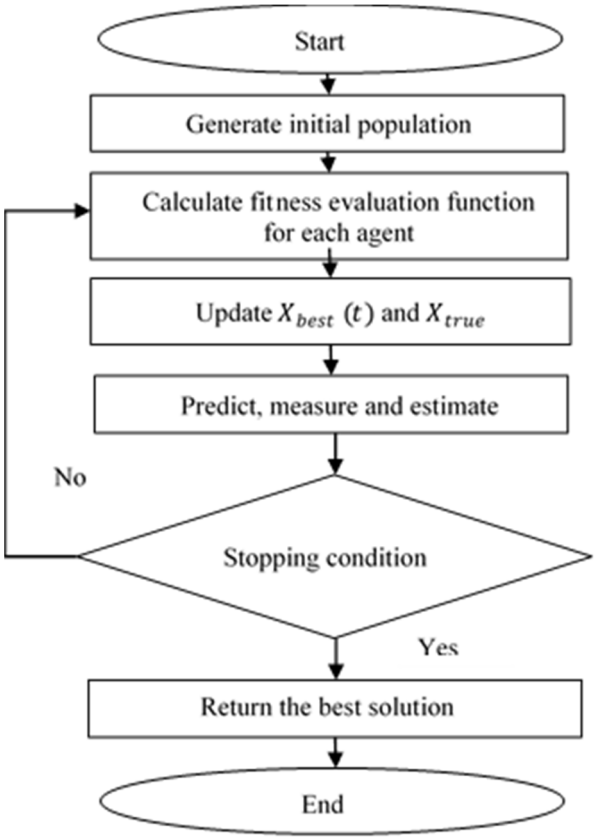


Figure 1. Flowchart of SKF algorithm.

3.1.1. Agent Encoding

Figure 2 shows the encoding of the agent. A matrix dimension is created by which the columns represent the number of features, and the row is the number of agents. The column number represents feature number. The state value at each dimension represents the probability of the corresponding feature to be selected. Two columns are added to the matrix dimension. The first column added is the percentage of features to be selected, S_p. The second column added is the training ratio. The agents randomly choose the training ratio between 0.5 and 0.8.

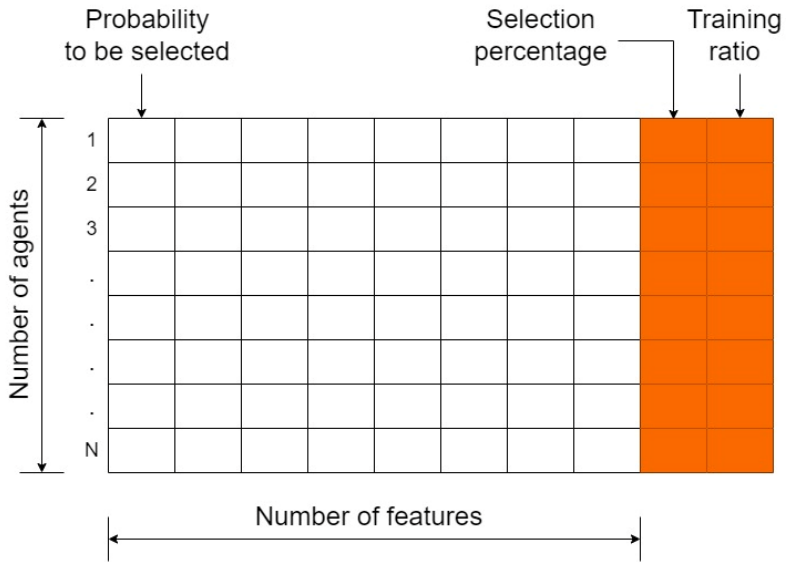


Figure 2. SKF's agents encoding.

From the value of the selection percentage, S_p, the number of selected features, N_{fs}, is calculated using Equation (7):

$$N_{fs} = N_f \times S_p \tag{7}$$

where N_f is the total number of features. N_{fs} features with the highest probability are selected. According to [32], only a few genes are relevant to achieve a 100% classification accuracy. Nevertheless, in earlier studies, some works reported more than 80% accuracy with a larger number of more than 100 genes [33,34].

3.1.2. Fitness Evaluation

In this study, an ANN classifier is used to evaluate the features selected by SKF. The rate of misclassification, also known as accuracy, is used as the fitness, and it is calculated as follows:

$$Rate\ of\ Misclassification = \frac{False\ Positive + False\ Negative}{Total\ Sample} \tag{8}$$

In the fitness evaluation process, the data are trimmed with the selected features only. The classifier is trained by the training data consisting of selected features only. Then, the trained classifier is tested. This study uses supervised learning, where the network's inputs and expected outputs are known. The error, the difference between actual and expected results, is then calculated. The backpropagation algorithm reduces this error until the ANN learns the training data, but Levenberg–Marquardt (LM) algorithm trains ANN 10 to 100 times faster than the standard backpropagation algorithm [35]. The LM algorithm is considered one of the most successful algorithms in increasing the convergence speed of the ANN with multilayer perceptron (MLP) architectures [36]. It inherits speed from the Newton method, but it also has the convergence capability of the steepest descent method. Therefore, it suits training a neural network in which the performance index is calculated in mean squared error (MSE) [37]. This study utilized the Neural Network toolbox available in Matlab. The proposed SKF for microarray feature selection is shown in Figure 3.

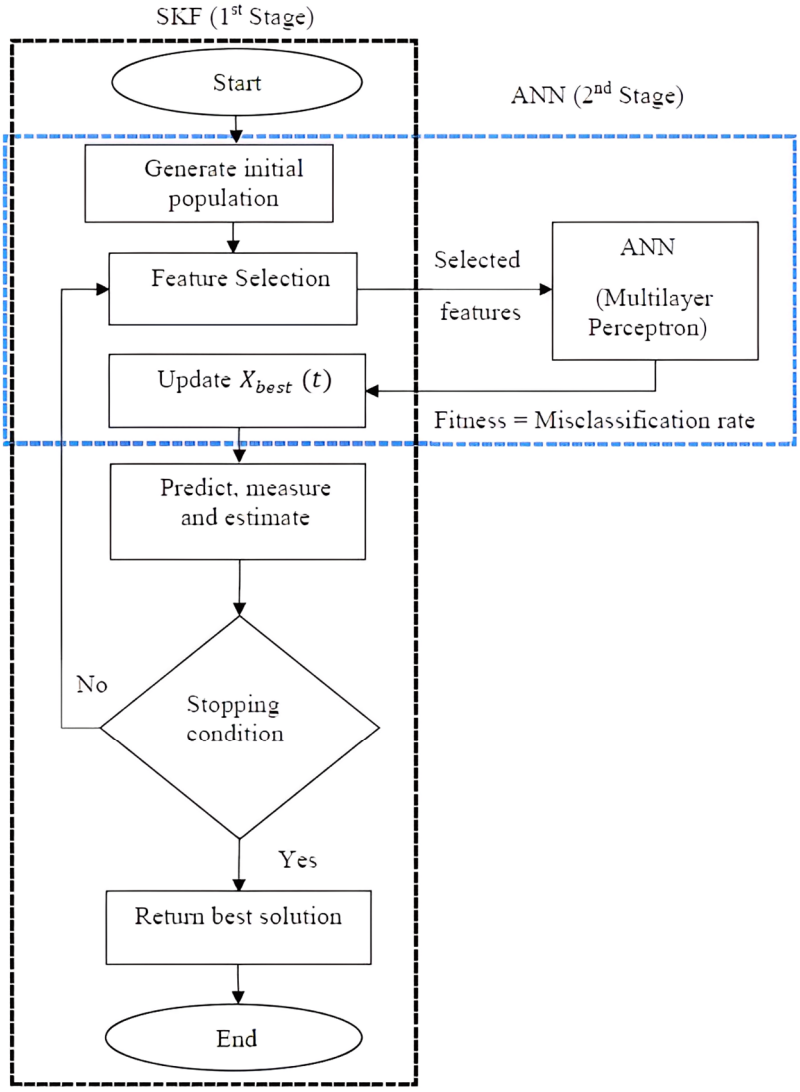


Figure 3. SKF for microarray feature selection.

3.1.3. SKF with Mutation (SKF-MUT)

In this study, a mutation operator is introduced in the search strategy of SKF to increase the exploration and the chance of the agent in finding the best solution for the feature selection of microarray data. The technique of mutation adopted is the scramble mutation. Scramble mutation happens by randomly changing the sequence of positions of the solutions in the matrix dimension. This mutation involves *X_best* and is conducted after the estimation phase. The number of dimensions involved in the mutation is dependent on the number of features selected by *X_best* according to the equation:

$$Number\ of\ Mutation = N_{fs} \tag{9}$$

The starting dimension involved in mutation is randomly selected. This concept is shown in Figure 4. After mutation, the fitness of the new solution created after mutation, *X_mut*, is evaluated. The best agent, *X_best*, will be replaced by *X_mut* if it is a better solution. Nevertheless, suppose the mutated solution, *X_mut*, does not hold a better misclassification rate; in that case, the solution will still be adopted by a randomly selected agent to increase the exploration in the SKF algorithm (Figure 4). The flowchart of SKF-MUT algorithm is shown in Figure 5.

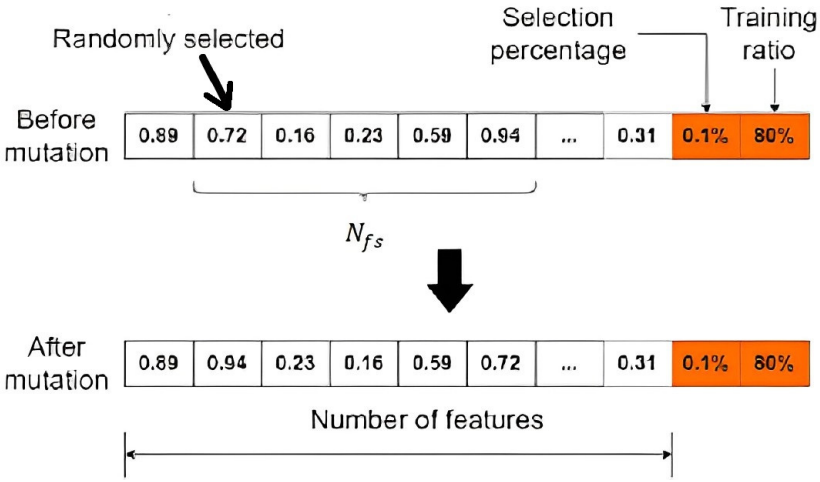


Figure 4. Concept of mutation.

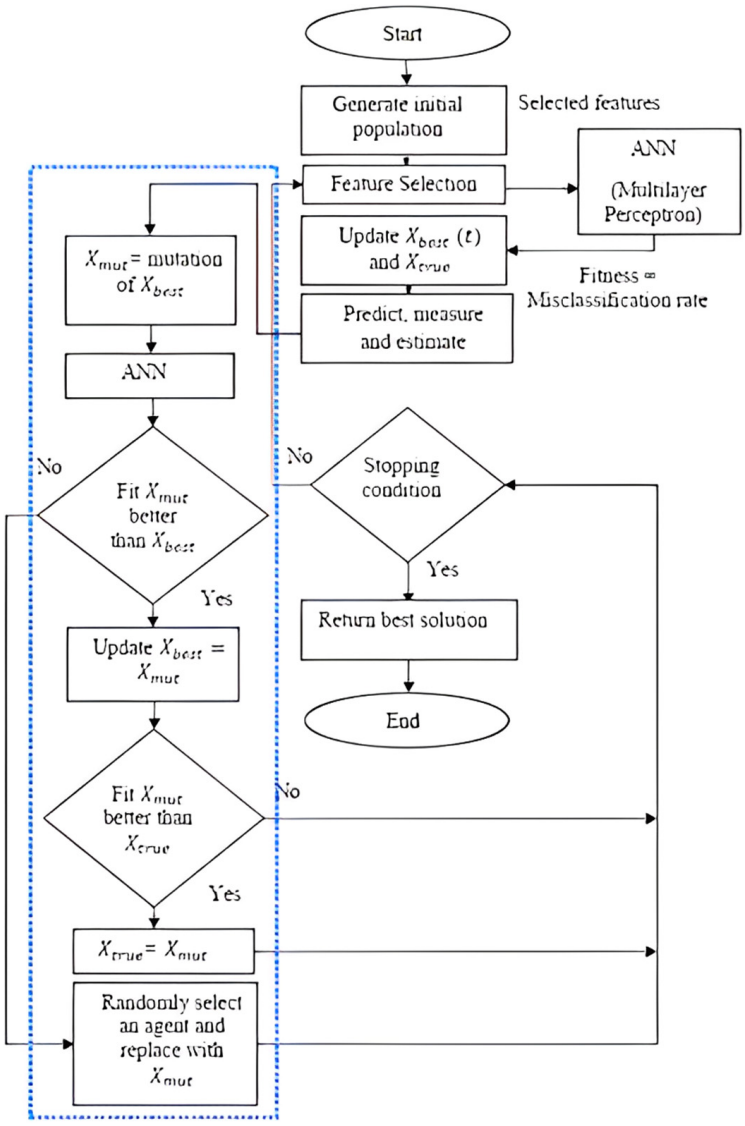


Figure 5. Flowchart of SKF-MUT algorithm.

3.2. Dataset

Eight microarray public datasets associated with cancer diseases, ranging from diffuse large b-cell lymphomas (DLBCL) [38], prostate cancer [39], lung cancer [40], leukemia cancer [41], small, round blue cell tumor (SRBCT) [42], brain tumor [43], nine tumors [44], and 11 tumors [6] are used in this study.

Table 1 shows a summary description of these eight (8) datasets. All data were taken from www.gems-system.org [45] in June 2017. Readers are referred to the original author's paper for the details of data collection techniques for each dataset. All the datasets are in matrices in Matlab file (.mat). Input and target data are given in the matrix dimension, where input is the features and samples, and the target is the correct output for every sample data. These datasets are commonly used by researchers in this field as seen in [32–34,46–48].

Table 1. Summary of microarray datasets.

Dataset and Ref.	Binary/ Multiclass	No. of Classes	Description	No. of Features	No. of Samples
DLBCL [38]	Binary	2	Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas	5469	77
Prostate [39]	Binary	2	Prostate tumor and normal tissues	10,509	102
Lung [40]	Multiclass	5	4 lung cancer types (adeno, squamous, COID, SMCL) and normal tissues	12,600	203
Leukemia [41]	Multiclass	3	AML, ALL, and mixed-lineage leukemia (MLL)	11,225	72
SRBCT [42]	Multiclass	4	Small, round blue cell tumor (SRBCT) of childhood; EWS, RMS, BL and NB	2308	83
Brain Tumors [43]	Multiclass	5	5 human brain tumor types; medulloblastoma, malignant glioma, AT/RT, normal cerebellum and PNET.	5920	90
9 Tumors [44]	Multiclass	9	9 various human tumor types; NSLC, colon cancer, breast cancer, ovarian cancer, leukemia, renal cancer, melanoma cancer, prostate cancer, and CNS	5726	60
11 Tumors [6]	Multiclass	11	11 various human tumor types; ovarian cancer, bladder or ureter cancer, breast cancer, colorectal cancer, gastroesophageal cancer, liver cancer, prostate cancer, pancreas cancer, lung adeno, and lung squamous	12,533	174

4. Results

SKF and SKF-MUT are executed for feature selection using 30 agents and the number of iteration is set to 100. Each experiment is run 30 times. This parameter setting is based on work performed on the feature selection of EEG signals using previous variants of SKF [19]. In addition, the maximum percentage for SKF and SKF-MUT to select features was set according to two conditions, 0.1% and 1%, relative to the data size. Finally, the classification accuracy for classifying test data based on the selected features was recorded. The results presented are based on the best solution found from 30 runs of each dataset.

Five works using the same datasets are used as benchmarks, out of these six works, four works are variants of the PSO algorithm—improved binary PSO (IBPSO) [33], modified binary PSO (MBPSO) [46], and information gain and improved simplified swarm optimization (IG-ISSO) [32]—and the other two algorithms are information gain genetic algorithm (IG-GA) [34] and multiobjective binary biogeography optimization (MOBBO) [48].

4.1. Accuracy of the Proposed Algorithms

Accuracy/rate of misclassification is the primary performance measurement used for the comparison. The best accuracy found by every algorithm is tabulated in Table 2. All

proposed algorithms with 0.1% and 1% selected features achieved a maximum accuracy of 100% in the two binary datasets: prostate and DLBCL. The complexity of the dataset classification arose with the number of classes. With a lower number of features (0.1%), the SKF feature selector performed better than SKF-MUT. While SKF was able to achieve 100% accuracy for two out of the six multiclass data, SKF-MUT only achieved 100% accuracy for leukemia data. Interestingly, it was observed that the best accuracy achieved by SKF (0.1%) decreased with the increase in number of classes. Both SKF and SKF-MUT were not able to achieve an accuracy of more than 90% with 0.1% number of features.

Table 2. Best accuracy (%) of proposed algorithms according to dataset.

Dataset	Class of Dataset	SKF (0.1%)	SKF-MUT (0.1%)	SKF (1%)	SKF-MUT (1%)
DLBCL	2	100	100	100	100
Prostate	2	100	100	100	100
Lung	5	99.5	99	100	100
Leukemia	3	100	100	100	100
SRBCT	4	100	96.4	100	100
Brain Tumors	5	98.9	98.9	100	100
9 Tumors	9	86.7	80	95	95
11 Tumors	11	86.2	87.4	97.7	98.3

As the number of selected features increased to 1%, both SKF and SKF-MUT were able to achieve more than a 100% accuracy rate; specifically, 100% accuracy was achieved by SKF and SKF-MUT for DLBCL, prostate, lung, leukemia, SRBCT, and brain tumor data. The nine and eleven tumor data have a high number of classes, thus the classification task was more complex. Both SKF and SKF-MUT (1%) achieved 95% accuracy for nine tumors, while SKF-MUT performed better than SKF for eleven tumors with up to 98.3% accuracy.

The performance of the proposed SKF and SKF-MUT is further analyzed using the confusion matrix shown in Figure 6. Only the SKF and SKF-MUT with 1% are presented here due to the better performance compared to the variants with 0.1% number of features. Both SKF and SKF-MUT were able to correctly classify data from all classes for the DLBCL, prostate, lung, leukemia, SRBCT, and brain tumor datasets; the confusion matrices are shown in Figure 6a–f. For the nine tumors dataset, the SKF-MUT was able to correctly classify data from seven target classes. This is better than SKF. The algorithm wrongly classified three data as class 1, when they were actually from class 6 and 9. For the 11 tumor dataset, SKF correctly labeled data from seven classes with respect to the target class, while SKF-MUT correctly labeled data from eight classes. Meanwhile, the SKF wrongly labeled data from four target classes into three classes, and SKF-MUT wrongly labeled data from three target classes into four output classes. Thus, the accuracy of both SKT and SKF-MUT for the 11 tumors dataset is equivalent.

4.2. Performance Benchmarking

The performance of the proposed algorithms against benchmark works is tabulated in Table 3. It can be seen that the proposed SKF and SKF-MUT are able to achieve a higher rate of 100% accuracy in comparison to the other algorithms. For the nine tumors data, none of the algorithms are able to achieve 100% accuracy. Nonetheless, both SKF and SKF-MUT achieved the highest accuracy at 95%. The SKF-MUT achieved the best accuracy at 98.3% for eleven tumors.

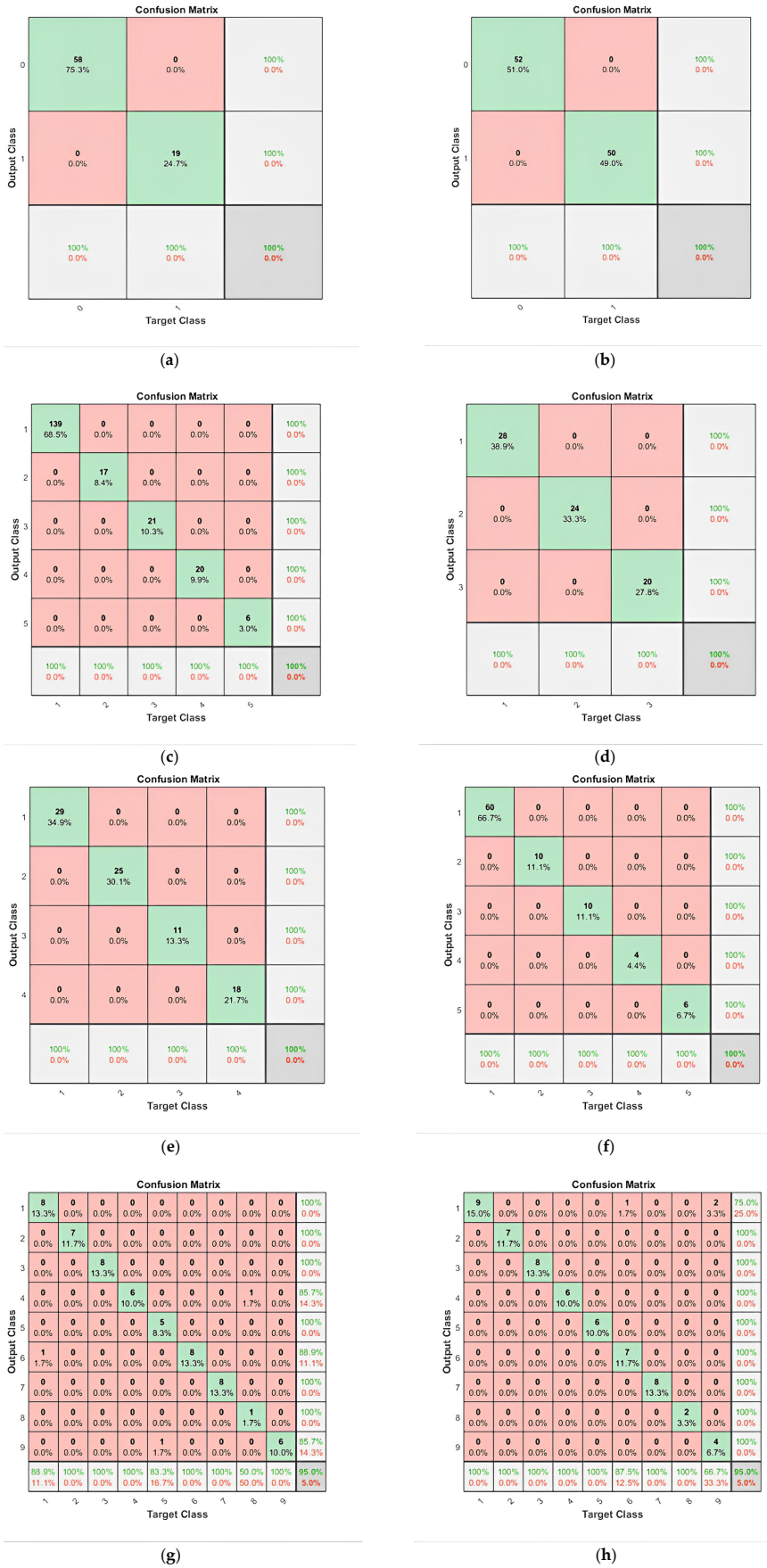
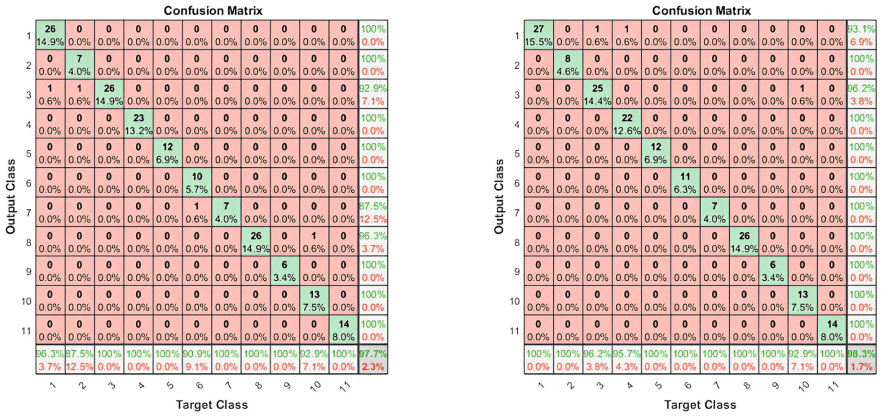


Figure 6. *Cont.*



(**i**) (**j**)

Figure 6. Confusion matrix of: (**a**) SKF (1%) and SKF-MUT (1%) for DLBCL dataset; (**b**) SKF (1%) and SKF-MUT (1%) for prostate dataset; (**c**) SKF (1%) and SKF-MUT (1%) for lung dataset; (**d**) SKF (1%) and SKF-MUT (1%) for leukemia dataset; (**e**) SKF (1%) and SKF-MUT (1%) for SRBCT dataset; (**f**) SKF (1%) and SKF-MUT (1%) for brain tumor dataset; (**g**) SKF (1%) for 9 tumors dataset; (**h**) SKF-MUT (1%) for 9 tumors dataset; (**i**) SKF (1%) for 11 tumors dataset; (**j**) SKF-MUT (1%) for 11 tumors dataset.

Table 3. Comparison of best solution (accuracy, %).

Dataset	Class of Dataset	SKF (1%)	SKF-MUT (1%)	IBPSO	MBPSO	IG-ISSO	IG-GA	MOBBBO
DLBCL	2	100	100	100	100	100	100	100
Prostate	2	100	100	92.16	97.94	99.0196	96.08	98.33
Lung	5	100	100	96.55	95.86	100	95.57	98.47
Leukemia	3	100	100	100	100	100	98.61	100
SRBCT	4	100	100	100	100	100	100	100
Brain Tumor	5	100	100	94.40	92.56	98	93.33	96.67
9 Tumors	9	95	95	78.33	75.50	85	85	80.50
11 Tumors	11	97.7	98.3	93.10	92.41	92.53	92.53	92.41

For an unbiased study, statistical analysis is carried out to further compare the performance of all algorithms [49,50]. The Friedman rank test used in this work and significance level used is $\alpha = 0.05$. The test is firstly performed among the proposed algorithms using different sets of selection percentage. Based on Table 4, SKF-MUT (1%) ranked the first in terms of accuracy. The statistical value is 6.45; this is smaller than the critical value of 7.81, indicating that there is no significant difference between the algorithms.

Table 4. Friedman rank of proposed algorithms.

Algorithm	Friedman Rank
SKF (0.1%)	2.875
SKF-MUT (0.1%)	3.25
SKF (1%)	2
SKF-MUT (1%)	1.875

The SKF-MUT (1%) that ranked first in the previous Friedman test is benchmarked with the five algorithms from previous works. The Friedman rank is shown in column 2 of Table 5. SKF-MUT (1%) is ranked the best among all the algorithms compared. The statistical value is higher than the critical value (11.75 > 11.07), suggesting that there is a significant difference between the algorithms. Thus, a Holm post hoc procedure is carried for further analysis. From columns 3 and 4 of Table 5, it can be seen that SKF-MUT (1%) is significantly better than MBPSO and IG-GA (p-value is smaller than Holm value).

Table 5. Friedman rank and Holm post hoc values.

Algorithm	Friedman Rank	*p*	Holm
SKF-MUT (1%)	1.9375		
IBPSO	3.875	0.038333	0.016667
MBPSO	4.5625	0.005012	0.01
IG-ISSO	2.6875	0.42267	0.05
IG-GA	4.375	0.009166	0.0125
MOBBO	3.5625	0.082352	0.025

5. Conclusions

There is a vast amount of research on the feature selection and classification of microarray data involving different metaheuristics algorithms. This biological dataset is typically highly dimensional, with only a small number of samples and many features, making the task of microarray analysis incredibly challenging. In this paper, two approaches have been introduced for the feature selection of microarray data: SKF and SKF-MUT. This study only focused on the algorithms for feature selection and did not study different classification methods. The ANN was used as a classifier in this work. The proposed algorithms, SKF and SKF-MUT, randomly selected the number of features needed for a higher accuracy of classification. Based on the experimental results, SKF-MUT, whose number of selected features was 1% of the total number of features, was able to achieve a 100% accuracy for six out of eight datasets and was able to achieve 95% and 98.3% accuracy for the other two datasets, respectively. SKF and SKF-MUT accuracy improved when the number of features increased. This is because when the algorithms are allowed to select a more significant number of features, the chances of obtaining informative features is increased. A decrease in the number of genes may decrease the prediction accuracy. The proposed algorithm is found to be significantly better than several existing works and on par with others.

Author Contributions: Conceptualization, N.A.Z., N.A.A.A., T.B. and N.H.A.A.; formal analysis, N.A.Z.; funding acquisition, N.H.A.A.; investigation, N.A.Z.; methodology, N.A.Z., N.A.A.A. and N.H.A.A.; resources, N.H.A.A.; data curation, N.A.Z.; writing—original draft preparation, N.A.Z., N.A.A.A. and T.B.; writing—review and editing, N.A.A.A., T.B., N.H.A.A. and A.K.G.; visualization, N.A.Z.; supervision, N.A.A.A. and T.B.; validation, N.A.Z. and N.A.A.A.; project administration, N.H.A.A. and A.K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research is fully supported by Ministry of Higher Education (MOHE) (FRGS/1/2015/TK04/MMU/03/2) and Multimedia University (MMUI/210015). In addition, the authors fully acknowledged Multimedia University for the approved fund, which makes this vital research viable and effective.

Data Availability Statement: The data that support the findings of this study are open access and the sources of the data are cited. The readers are referred to the original sources.

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

References

1. Tang, Z.; Zhang, T.; Yang, B.; Su, J.; Song, Q. spaCI: Deciphering spatial cellular communications through adaptive graph model. *Brief. Bioinform.* **2023**, *24*, bbac563. [CrossRef] [PubMed]
2. Musheer, R.A.; Verma, C.K.; Srivastava, N. Novel machine learning approach for classification of high-dimensional microarray data. *Soft Comput.* **2019**, *23*, 13409–13421. [CrossRef]
3. Dwivedi, A.K. Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Comput. Appl.* **2018**, *29*, 1545–1554. [CrossRef]
4. Maurya, S.; Tiwari, S.; Mothukuri, M.C.; Tangeda, C.M.; Nandigam, R.N.S.; Addagiri, D.C. A review on recent developments in cancer detection using Machine Learning and Deep Learning models. *Biomed. Signal Process. Control* **2023**, *80*, 104398. [CrossRef]
5. Bhatt, H.; Shah, V.; Shah, K.; Shah, R.; Shah, M. State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: A comprehensive review. *Intell. Med.* **2022**. [CrossRef]

6. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537. [CrossRef]
7. Han, X.; Li, D.; Liu, P.; Wang, L. Feature selection by recursive binary gravitational search algorithm optimization for cancer classification. *Soft Comput.* **2020**, *24*, 4407–4425. [CrossRef]
8. Shanavas, I.H.; Gnanamurthy, R.K. Application metaheuristic technique for solving VLSI global routing problem. In Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, India, 27–28 October 2009; pp. 915–917. [CrossRef]
9. Herbert-Acero, J.F.; Martínez-Lauranchet, J.; Probst, O.; Méndez-Díaz, S.; Castillo-Villar, K.K.; Valenzuela-Rendón, M.; Réthoré, P.-E. A Hybrid Metaheuristic-Based Approach for the Aerodynamic Optimization of Small Hybrid Wind Turbine Rotors. *Math. Probl. Eng.* **2014**, *2014*, 746319. [CrossRef]
10. Fernandez, S.A.; Juan, A.A.; De Armas Adrian, J.; Silva, D.G.E.; Terren, D.R. Metaheuristics in Telecommunication Systems: Network Design, Routing, and Allocation Problems. *IEEE Syst. J.* **2018**, *12*, 3948–3957. [CrossRef]
11. Fuellerer, G.; Doerner, K.F.; Hartl, R.F.; Iori, M. Metaheuristics for vehicle routing problems with three-dimensional loading constraints. *Eur. J. Oper. Res.* **2010**, *201*, 751–759. [CrossRef]
12. Huang, H.C.; Tsai, C.C. Global path planning for autonomous robot navigation using hybrid metaheuristic GA-PSO algorithm. In Proceedings of the SICE Annual Conference 2011, Tokyo, Japan, 13–18 September 2011; pp. 1338–1343.
13. Pelta, D.A.; González, J.R.; Vega, M.M. A simple and fast heuristic for protein structure comparison. *BMC Bioinform.* **2008**, *9*, 161. [CrossRef]
14. Sun, J.; Garibaldi, J.M.; Hodgman, C. Parameter estimation using metaheuristics in systems biology: A comprehensive review. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2012**, *9*, 185–202. [CrossRef]
15. Ibrahim, Z.; Aziz, N.A.; Aziz, N.A.A.; Razali, S.; Shapiai, M.I.; Nawawi, S.W.; Mohamad, M.S. A Kalman Filter Approach for Solving Unimodal Optimization Problems. *ICIC Express Lett.* **2015**, *9*, 3415–3422.
16. Yusof, Z.M.; Satiman, S.N.; Azmi, K.M.; Muhammad, B.; Razali, S.; Ibrahim, Z.; Aspar, Z.; Ismail, S. I-ECO-084: Solving Airport Gate Allocation Problem using Simulated Kalman Filter Faculty of Electrical and Electronics Engineering Faculty of Electrical Engineering. In Proceedings of the International Conference on Knowledge Transfer, Putrajaya, Malaysia, 1–3 December 2015.
17. Lazarus, K.; Noordin, N.H.; Ibrahim, Z.; Abas, K.H. Adaptive Beamforming Algorithm based on Simulated Kalman Filter. In Proceedings of the Asia Multi Conference on Modelling and Simulation, Sabah, Malaysia, 5–6 December 2016; pp. 19–23.
18. Aziz, N.H.A.; Aziz, N.A.A.; Ibrahim, Z.; Razali, S.; Abas, K.H.; Mohamad, M.S. A Kalman Filter approach to PCB drill path optimization problem. In Proceedings of the 2016 IEEE Conference on Systems, Process and Control (ICSPC), Melaka, Malaysia, 16–18 December 2016; pp. 33–36. [CrossRef]
19. Adam, A.; Ibrahim, Z.; Mokhtar, N.; Shapiai, M.I.; Mubin, M.; Saad, I. Feature selection using angle modulated simulated Kalman filter for peak classification of EEG signals. *SpringerPlus* **2016**, *5*, 1580. [CrossRef]
20. Wolpert, D.H.; Macready, W.G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
21. Osama, S.; Shaban, H.; Ali, A.A. Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Syst. Appl.* **2023**, *213*, 118946. [CrossRef]
22. Alrefai, N.; Ibrahim, O. Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Comput. Appl.* **2022**, *34*, 13513–13528. [CrossRef]
23. Aziz, R.M. Application of nature inspired soft computing techniques for gene selection: A novel frame work for classification of cancer. *Soft Comput.* **2022**, *26*, 12179–12196. [CrossRef]
24. 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. [CrossRef]
25. Kundu, R.; Chattopadhyay, S.; Cuevas, E.; Sarkar, R. AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets. *Comput. Biol. Med.* **2022**, *144*, 105349. [CrossRef]
26. Vahmiyan, M.; Kheirabadi, M. Feature selection methods in microarray gene expression data: A systematic mapping study. *Neural Comput. Appl.* **2022**, *34*, 19675–19702. [CrossRef]
27. 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]
28. Kelemen, J.Z.; Kertész-Farkas, A.; Kocsor, A.; Puskás, L.G. Kalman filtering for disease-state estimation from microarray data. *Bioinformatics* **2006**, *22*, 3047–3053. [CrossRef] [PubMed]
29. Toscano, R. Structured Controllers for Uncertain Systems. In *A Stochastic Optimization Approach*; Springer London Limited: London, UK, 2013; pp. 1–298.
30. Rahman, T.A.; Ibrahim, Z.; Aziz, N.A.A.; Zhao, S.; Aziz, N.H.A. Single-Agent Finite Impulse Response Optimizer for Numerical Optimization Problems. *IEEE Access* **2018**, *6*, 9358–9374. [CrossRef]
31. Aziz, N.H.A.; Ibrahim, Z.; Aziz, N.A.A.; Razali, S. Parameter-less Simulated Kalman Filter. *Int. J. Softw. Eng. Comput. Syst.* **2017**, *3*, 129–137. [CrossRef]
32. Lai, C.M.; Yeh, W.C.; Chang, C.Y. Gene selection using information gain and improved simplified swarm optimization. *Neurocomputing* **2016**, *218*, 331–338. [CrossRef]

33. 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]
34. Yang, C.-H.; Chuang, L.-Y.; Yang, C.-H. IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data. *J. Med. Biol. Eng.* **2010**, *30*, 23–28.
35. Yadav, D.; Naresh, R.; Sharma, V. Stream flow forecasting using Levenberg-Marquardt algorithm approach. *Environ. Eng.* **2011**, *3*, 30–40.
36. Hagan, M.T.; Menhaj, M.B. Training Feedforward Networks with the Marquardt Algorithm. *IEEE Trans. Neural Netw.* **1994**, *5*, 989–993. [CrossRef]
37. Haykin, S.; Nie, J.; Currie, B. Neural network-based receiver for wireless communications. *Electron. Lett.* **1999**, *35*, 203–205. [CrossRef]
38. Shipp, M.A.; Ross, K.N.; Tamayo, P.; Weng, A.P.; Kutok, J.L.; Aguiar, R.C.; Gaasenbeek, M.; Angelo, M.; Reich, M.; Pinkus, G.S.; et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **2002**, *8*, 68–74. [CrossRef]
39. Singh, D.; Febbo, P.G.; Ross, K.; Jackson, D.G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A.A.; D'Amico, A.V.; Richie, J.P.; et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **2002**, *1*, 203–209. [CrossRef] [PubMed]
40. Bhattacharjee, A.; Richards, W.G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.; Gillette, M.; et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13790–13795. [CrossRef]
41. Armstrong, S.A.; Staunton, J.E.; Silverman, L.B.; Pieters, R.; Boer, M.L.D.; Minden, M.D.; Sallan, S.E.; Lander, E.S.; Golub, T.R.; Korsmeyer, S.J. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **2002**, *30*, 41–47. [CrossRef]
42. Tirumala, S.S.; Narayanan, A. Classification and diagnostic prediction of prostate cancer using gene expression and artificial neural networks. *Neural Comput. Appl.* **2019**, *31*, 7539–7548. [CrossRef]
43. 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**, *415*, 436–442. [CrossRef]
44. Staunton, J.E.; Slonim, D.K.; Coller, H.A.; Tamayo, P.; Angelo, M.J.; Park, J.; Scherf, U.; Lee, J.K.; Reinhold, W.O.; Weinstein, J.N.; et al. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10787–10792. [CrossRef] [PubMed]
45. Statnikov, A.; Tsamardinos, I.; Dosbayev, Y.; Aliferis, C.F. GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int. J. Med. Inform.* **2005**, *74*, 491–503. [CrossRef] [PubMed]
46. Mohamad, M.S.; Omatu, S.; Deris, S.; Yoshioka, M.; Ibrahim, Z. A Modified Binary Particle Swarm Optimization for Selecting the Small Subset of Informative Genes From Gene Expression Data. *Int. J. Innov. Comput. Inf. Control* **2012**, *8*, 4285–4297.
47. Lai, C.M. Multi-objective simplified swarm optimization with weighting scheme for gene selection. *Appl. Soft Comput. J.* **2018**, *65*, 58–68. [CrossRef]
48. Li, X.; Yin, M. Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Trans. Nanobiosci.* **2013**, *12*, 343–353. [CrossRef] [PubMed]
49. García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.* **2010**, *180*, 2044–2064. [CrossRef]
50. Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.* **2011**, *1*, 3–18. [CrossRef]