Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Hybrid Feature Selection Method using Novel Pasi-Luukka and Genetic Algorithm Method for Microarray Cancer Classification

Cham Rui Hong[1], Nursabillilah Mohd Ali[1,*], Johar Akbar Mohamat Gani[2], Nurul Fatiha Johan[1], Ezreen Farina Shair[1], Nur Hazahsha Shamsudin[1], Mohd Safirin Karis[2], Hairol Nizam Mohd Shah[1], Amar Faiz Zainal Abidin[2], Muhammad Zaid Aihsan[3]

[1] Fakulti Teknologi dan Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia
[2] Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia
[2] Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia

ARTICLE INFO

Article history:
Received 31 January 2025
Received in revised form 7 February 2025
Accepted 30 June 2025
Available online 10 July 2025

Keywords:
Pasi luukka; genetic algorithm; machine leaning; optimization; feature selection

ABSTRACT

Deoxyribonucleic acid (DNA) microarray technology enables the simultaneous measurement of the expression level of numerous genes, thus enabling the identification of patterns in gene expression that may cause a disease or a particular biological process. The DNA microarray technology can identify cancer cells by analysing the gene expression difference between normal cells and cancer cells. However, due to the vast number of features in the DNA microarray, the feature selection method is required to identify the most relevant subset of microarray features for subsequent analysis. In this research paper, a novel hybrid feature selection method called Pasi Luukka-Genetic Algorithm + Support Vector Machine is introduced. This approach combines the strengths of filter and wrapper methods to effectively select features from eight (8) cancer datasets. The Pasi Luukka algorithm filters irrelevant features and reduces dimensionality. A metaheuristic-based feature selection, which is the Genetic Algorithm (GA), selects the optimum features from the filtered features. This paper evaluates the performance of the proposed method, and a comparison work is conducted against other existing hybrid feature selection methods in the literature. The evaluation considers accuracy metrics and the number of selected features using the same microarray datasets.

1. Introduction

Cancer encompasses a range of diseases characterised by abnormal cells in the body growing uncontrollably and spreading to nearby tissues [1,2]. Cancer can lead to death if the spread is not controlled. According to the Cancer Statistics from the American Cancer Society, there is an estimated new cancer case of 1,918,030 and an estimated death of 609,360 from all sites (e.g., breast, leukaemia, respiratory system, digestive system and other types of cancers) in the United States in 2022 [3].

* *Corresponding author*
E-mail address: nursabillilah@utem.edu.my

https://doi.org/10.37934/ard.136.1.121

The DNA microarray is introduced to detect cancer by analysing the gene expression level in each cell [4]. A DNA microarray is a lab-on-chip technique that consists of an array of numerous microscopic DNA spots or sequences (oligo) spotted on the solid surface [5,6]. Specific probes are used to measure the relative abundance of transcripts or gene expression levels in the target sample [7]. The analysis of gene expression levels is essential for detecting cancer by determining biological processes and identifying changes in the genes' expression patterns that are relevant to its development [8]. The DNA microarray is useful in identifying diseases such as cancer, discovering drugs, or developing more effective drugs for treatment purposes [9,10].

However, there are two major barriers to analysing DNA microarray datasets: high dimensionality and high complexity [11]. From the aspect of dimensionality, microarray technology generates large datasets with gene expression values for 6,000–60,000 genes/feature in a cell mixture [12]. The curse of high dimensionality brings low performance issues, such as poor classification accuracy and stalled genetic search [13]. Moreover, the DNA microarray dataset has high complexity, thus signifying a correlation between the genes. Standard machine learning methods are suited for less-dimensional data and do not perform well because these gene expression datasets have more features than the samples, as discussed as in previous works [11-13].

Dimensionality reduction methods have been introduced to reduce dimensionality [14], where these techniques can be categorised into two main groups: feature extraction and feature selection [15]. Feature extraction extracts information from original features to generate new features with lower dimensionality, while feature selection selects a subset of the original features without creating a new one [16]. Feature selection aims to discard irrelevant or redundant data based on certain criteria [17,18]. This helps to improve the accuracy of the model by reducing noise in the data and focusing on the most useful information [19]. Feature selection also aids in reducing the computational time of the learning process [20].

There are three main categories of feature selection methods, including filter, wrapper, and embedded methods [21,22]. The filter feature selection method selects subsets of features by evaluating the relationship between features and target variables based on their general statistical properties and a score is assigned to each of them [23]. These features are ranked via specific criteria and used to filter out the features with the highest scores. One of the advantages of the filter method is that filter methods have faster computation speed [24]. However, since the performance of selected features on the classifier is ignored, the resulting accuracy may be low due to the loss of useful information [24,25].

Examples of filter methods are the Fisher Scoring algorithm, Laplacian Score (LS), Independent Component Analysis (ICA), Minimum Redundancy Maximum Relevance (mRMR), Information Gain (IG), Mutual Information (MI), Correlation-Based Feature Selection (CFS), and Pasi Luukka, among others [26,27]. Figure 1 shows the general workflow for the filter method.

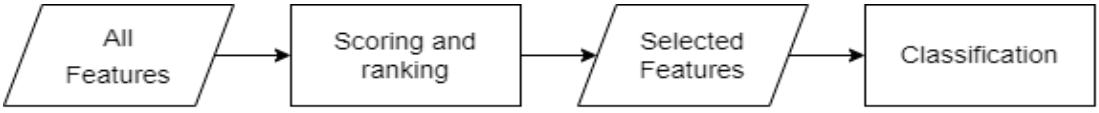


Fig. 1. Workflow of filter method

Wrapper methods are a feature selection method that utilises the performance of the selected classifier algorithm as a metric to select the best feature subset [28]. The wrapper works by applying a learning algorithm during the evaluation process of possible feature subsets by using the fitness function in each iteration and selecting the optimal subset among these results. The wrapper has higher performance than the filter, but the computation time is slower than that of the filter and

embedded methods [28]. Examples of wrapper methods are Genetic Algorithm (GA), Bat Algorithm (BA), Firefly Algorithm (FA), Particle Swarm Optimisation (PSO), Whale Optimisation Algorithm (WOA), Ant Colony Optimisation (ACO), etc [29]. Figure 2 shows the general wrapper's workflow.

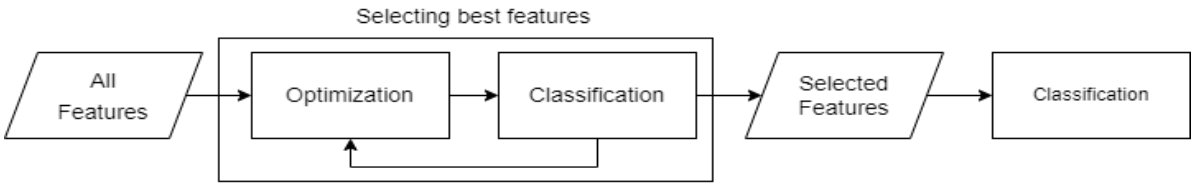


Fig. 2. Workflow of wrapper method

The final category is known as embedded methods [30]. An embedded method is known as integrated feature selection as embedded approaches attempt to discover an optimal subset of features while building a model [31]. In the training phase, the classifier fine-tunes its internal parameters and assigns appropriate weights to each feature to attain the highest possible classification accuracy [32]. As a result, the feature selection process and model construction are performed simultaneously in a single step [33-38]. Figure 3 demonstrates the workflow of the embedded method.

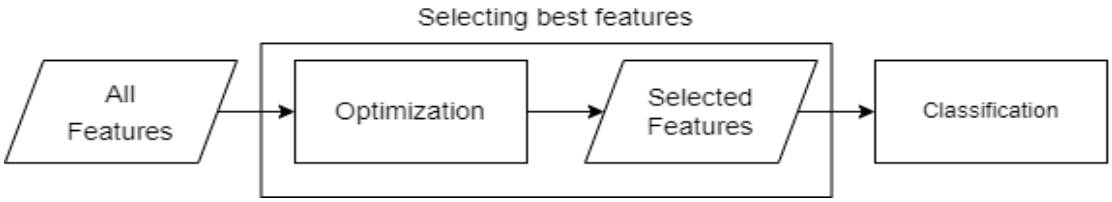


Fig. 3. Workflow of embedded method

Classification is a method of data analysis that involves predicting the class or category to which a given data point belongs, based on a predefined set of classes [39]. A classification is a form of supervised learning, in which the class labels for the training data are already known. To construct a classifier, the model undergoes two processes. Initially, the model is trained on a collection of labelled data, enabling it to learn the patterns and relationships within the data. Subsequently, the trained model is employed to make predictions on new and unseen data, assigning appropriate class labels based on its learned knowledge. The performance of the classifier is evaluated based on how well it can make these predictions. The accuracy of the selected features can be evaluated by using a classifier to determine their effectiveness in predicting or classifying certain outcomes or targets. In the wrapper model, a classifier can be employed as a fitness function to assess the performance of selected features at that instance and further improve the performance [40].

2. Materials and Methods

A hybrid method known as the Pasi Luukka-Genetic Algorithm (GA) + Support Vector Machine (SVM) is proposed to obtain the optimal feature subset from the high dimensionality microarray datasets. This method uses the Pasi Luukka filter to reduce the dimensionality, GA wrapper to select features, and lastly, the SVM classifier for the evaluation of the results to determine the accuracy.

The Pasi Luukka was chosen as one of the filters in the hybrid method of this work based on its unique characteristics that complement other selection techniques. While the rarity of its usage by other researchers in hybrid methods is a factor, the authors' decision was primarily motivated by its

ability to contribute distinct insights to the ensemble. Notably, the hybrid selection method proposed by other researchers, Pasi Luukka + AltWOA, showcased promising results [41]. In addition to Pasi Luukka, the GA was integrated into this research approach due to its established reputation for accuracy. The survey conducted by other researchers indicated that hybrid feature selections involving the Genetic Algorithm consistently achieved superior accuracy compared to alternative methods while also yielding a minimal number of optimum subsets. This is evident through the GA's achievement of 100% accuracy on five out of the six evaluated datasets using a diverse set of algorithms [42]. This combination of approaches ensures a robust and effective feature selection process that enhances the accuracy of the classification model in this work.

The Pasi Luukka method utilises fuzzy entropy to evaluate the differences between the fuzzy set and the well-described crisp set for feature selection [43]. Pasi Luukka proposed the combinations of fuzzy entropy measures and similarity classifiers in selecting the most relevant features [44]. The workflow of the Pasi Luukka method is concluded as follows:

i. Firstly, the data is split into training and testing sets, and an ideal vector, v, is calculated for the training data. This ideal vector can be obtained by taking the generalised means of the samples belonging to specific classes. The shape of the ideal vector is (number of classes in the label, number of features).
ii. Secondly, both the ideal vector and training sets are normalised between 0 and 1.
iii. Thirdly, the similarity, S, between the feature and the ideal vector is calculated.

$$S\left(x_{j,d}, v_{i,d}\right) = \sqrt[p]{\left(1 - \left|{x_{j,d}}^{p} - {v_{i,d}}^{p}\right|\right)} \tag{1}$$

Where:
x = normalised training sets;
v = normalised ideal vector;
j = current iteration in m, and m = the number of rows in the data;
i = current iteration in l, and l = the number of classes;
d = current iteration in t, and t = number of features; and
p = real number in the Minkowski distance formula, default is 1.

iv. Next, the fuzzy entropy measures are performed.

$$H_1(A) = -\sum_{j=1}^{n}\left(\mu_A\left(x_j\right) log \mu_A\left(x_j\right) + \left(1 - \mu_A\left(x_j\right) \log\left(1 - \mu_A\left(x_j\right)\right)\right)\right) \tag{2}$$

Where, $\mu_A\left(x_j\right)$ is a degree of membership of feature x in fuzzy set A. Fuzzy set A refers to a collection of elements or features, each of which is associated with a degree of membership that represents the extent to which the element belongs to the set.

v. In this work, the degree of membership is replaced by the similarity value, $S\left(x_{j,d}, v_{i,d}\right)$.
vi. Then, a rank is assigned to each feature based on its entropy value. The lower the entropy measure, the lower the randomness in the feature, and the more useful the feature.
vii. Finally, the subsets for the first 100 ranks are selected from the training set.
viii. The workflow of Pasi Luukka is illustrated in Figure 4.

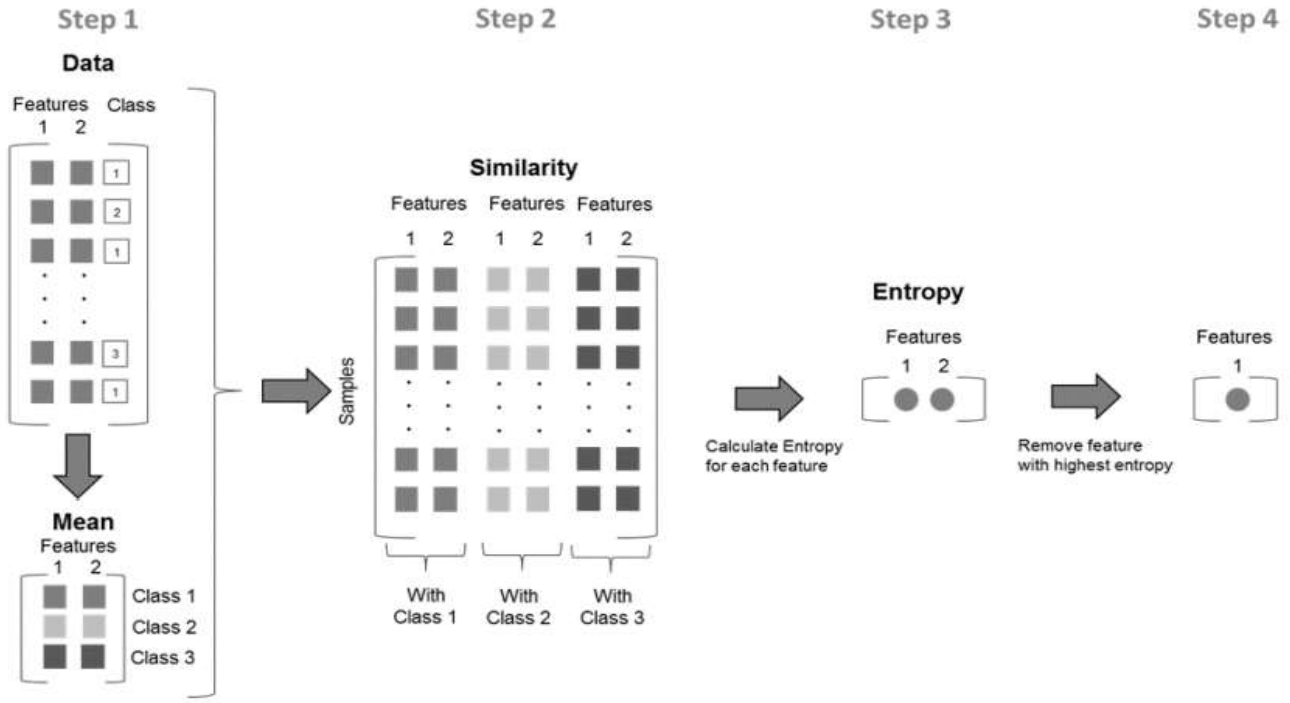


Fig. 4. Pasi Luukka framework

Algorithm 1: Pasi Luukka

Input: X_train (m, t), y_train (m,), number of filters (n)
Output: rank

Begin.
Initialise the number of rows (m), number of classes (l), number of features (t)
Initialise ideal vector (l, t)
For k in range l **do**:
  Calculate mean feature values for each class and assign to ideal vector
Initialise data_v (m, t) as features and data_c (m,) as label
Normalise data_v and ideal vector between 0 and 1.
Initialise sim (t, m, l)
Concatenate data_v and data_c to form data (m, t+1)
for j = 1 to m **do**
  for i = 1 to t **do**
    for k = 1 to l **do**
      sim[i][j][k] = (1 - idealvec[k][i] ^p - data[j][i] ^p) ^(1-p) --- Eq. (1)
    End for
  End for
End for
Reshape sim (t, m*l)
Calculate entropy, H using Eq. (2) taking sim variable as μ_A
Sort H using magnitude in descending order.
Return rank

Metaheuristic algorithms have been employed to address intricate problems that are difficult to solve using traditional approaches by searching through the search space, guided by a fitness function

[45]. Metaheuristic algorithms can be classified into two categories: population-based and single-based, depending on the strategies used for conducting searches. GA is one of the population-based metaheuristic algorithms based on the principles of natural evolution and natural selection. The GA comprises three primary operations: selection, crossover and mutation [46].

During the selection operation, the fittest chromosomes are chosen based on their fitness values, and these selected chromosomes are permitted to proceed to the next generation. This process ensures that the genetic traits associated with higher fitness have a higher chance of being inherited and propagated in subsequent generations. In the crossover operation, two selected individuals undergo a random exchange of genetic information at a crossover point, producing new offspring with a blend of traits from both parents. A mutation is used to maintain diversity within the population by introducing new genetic material that may not have been present in the original population. This can help prevent the population from becoming too homogenous and can also help to explore a wider range of potential solution [47].

The equation of the GA can be expressed by the Schema theorem [47]. The Schema theorem is used to understand and predict the evolution of a population of chromosomes in a GA. The Schema theorem also improves GA by obtaining a scheme with a higher fitness.

$$m_H(i+1) = F_H(i) m_H(i) \left[1 - p_c \frac{l_H}{l-1}\right] [(1-p_m)^{O_H}] \quad (3)$$

The workflow of the GA method is shown below:

i. Firstly, the parameters required by the GA are declared, such as the number of chromosomes, the maximum number of iterations, the crossover rate, and the mutation rate.
ii. Secondly, the filtered data are taken from the Pasi Luukka filter as input prior to the initialisation of the position of chromosomes. The fitness of each chromosome is calculated and the best fitness is obtained. Note that the fitness function used is SVM.
iii. Thirdly, a loop that iterates up to the maximum number of iterations is created. Within this loop, the number of crossovers is determined based on the crossover rate. If a randomly generated number is smaller than the crossover rate, the count of crossovers is incremented.
iv. Next, inside another loop of the number of crossovers, a crossover was performed between the parents, P_1 and P_2. The parents are generated using probability from inverse fitness. Two new chromosomes, X_1 and X_2, are produced under the crossover process by exchanging the elements from the two parents.
v. Then, mutation is applied randomly on chromosomes X_1 and X_2 based on the mutation rate.
vi. The following step involves concentrating X_1 and X_2 into an offspring chromosome, X_{new}.
vii. Next, the fitness of each feature in X_{new} is evaluated, and the best fitness and best feature subsets are updated.
viii. Finally, the above steps are repeated until the while loop terminates, and the optimal features are now selected.

Algorithm 2: Genetic Algorithm

Input: Number of chromosomes (N), the maximum number of iterations (T), crossover rate (CR), mutation rate (MR), lower bound, upper bound, threshold, x_train (m, t), y_train (m,) Index of best feature subset

Output:

```
        Begin.
1:      Initialise X (N, t) as chromosomes, t =0.
2:      for x in X:
3:        Set the position of x to a random value between the lower bound and upper bound.
4:      End for
5:      for x in X:
6:      if (x > threshold)
7:        Set the position of x to 1.
8       Else Set the position of x to 0.
9:      End for
10:     Initialise fit (N, 1) to store fitness value
11:     for i in range N:
12:           Evaluate fitness of X [i,:]
13:           Update the best fitness
14:           Update the best feature
15:     End for
16:     t+=1
11:     While (t < T):
12:       for i in range(N):
13:         if (random number < CR):
14:           Increment the Number of cross over (Nc) by 1
15:         End for
16:     Initialise parents, P1 (t,) and P2 (t,) using an inverse fitness function
17:     Initialise x1 (Nc, t) and x2 (Nc, t)
18:     Perform crossover on P1 and P2 and store results into x1 and x2
19:     Perform mutation in x1 and x2
20:     Merge x1 and x2 to produce offspring X_new (2*Nc, t)
21:     for i in range 2*Nc:
22:       Evaluate fitness of X_new [i, :]
23:       Update the best fitness
24:       Update the best feature
25:     End for
26:     Update X as concatenaion of X and X_new
27:     Update fitness as the concatenation of fit and best fitness
28:     End while
29:     Return index of best feature subset
30:
```

The operations of the Pasi Luukka filter and GA are discussed. The overall operation of the Pasi Luukka-GA + SVM hybrid method is shown below:

i. First, the desired microarray dataset is loaded, and the data are split into x and y.
ii. Secondly, the train test splitting is performed by taking x and y as inputs and setting the test size at 0.2 and the random state at 42.
iii. Thirdly, the Pasi Luukka filter is applied by taking the number of filters of 100 to reduce the large number of features to 100 only.
iv. Next, a GA wrapper is performed using the filtered training sets as input.

v. Then, the performance of the selected features is evaluated using the SVM classifier.
vi. Finally, graphs of the "Convergence of Fitness over Iteration" and "Convergence of Number of Feature over Iteration" are drawn.

The pseudocode form of the proposed Pasi Luukka-GA + SVM method is shown below:

Algorithm 3: Pasi Luukka-GA + SVM

Input: Dataset (M, N+1)
Output: Classification accuracy, number of the selected feature, convergence graphs

Begin.
Initialise feature_x as features (M, N) from the dataset.
Initialise label_y as label (M,) from the dataset.
Split feature_x and label_x into training and testing sets (x_train, x_test, y_train, y_test) taking test size = 0.2, random state = 42
Set the number of filters (n) as 100
Apply Pasi Luukka filter on x_train to obtain x_train_luukka (0.8M, n) - **Algorithm 1**
Declare parameters for GA:
Number of chromosomes (N) = 200, maximum number of iterations (T) = 500, crossover rate (CR)= 0.8, mutation rate (MR) = 0.1, classifier = SVM
Apply GA on x_train_luukka to obtain x_train_genetic - **Algorithm 2**
Evaluate training accuracy of x_train_GA using SVM
Obtain the index of selected genes.
Plot the "Convergence of Classification Error over Iteration" and "Convergence of Number of Feature over Iteration"

After obtaining the index of selected genes from the feature selection process, the symbol and official name of the gene can be further obtained from the list ID, as listed in the datasets. The procedures are shown below:

i. Firstly, all indices obtained were increased by 1, as the index in coding starts from 0, while the dataset's index starts from 1.
ii. Secondly, search is done along the dataset to find the list ID for the match index.
iii. Lastly, search for the gene's details is done via; 1) the National Library of Medicine [48,49]; or 2) the Gene ID Conversion Tool by DAVID Bioinformatics Resources, NIAID/NIH [49].
iv. For the SRBCT dataset [50], the gene names can be accessed through several steps:
v. Firstly, the postgenomic library and SRBCT dataset are loaded through R online compiler in the 'Examples' section on the dataset website: [51]
vi. Next, the indices of obtained genes are inserted to the matrix 'gene.names' to find the exact gene names for each index.
vii. The gene names are printed out and searched for the gene symbol and gene ID using gene names, as discussed in the above.

In this paper, the SVM classifier is selected. SVMs are supervised machine learning algorithms specifically designed for classification tasks. They aim to identify the hyperplane in a high-dimensional space that maximally separates different classes of data [52]. A hyperplane is a flat surface that divides a higher-dimensional space into two separate regions.

The reason for choosing SVM is due to fewer classification errors in the training data and better generalisation ability [53,54]. The hardware used in this study as in the following:

i. Device brand: Acer
ii. Processor: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz
iii. Installed RAM: 12.0 GB
iv. System type 64-bit operating system, x64-based processor

The software used in this research is:

i. Code editor: Visual Studio Code
ii. Programming language: Python

3. Results and Discussion

3.1 Dataset

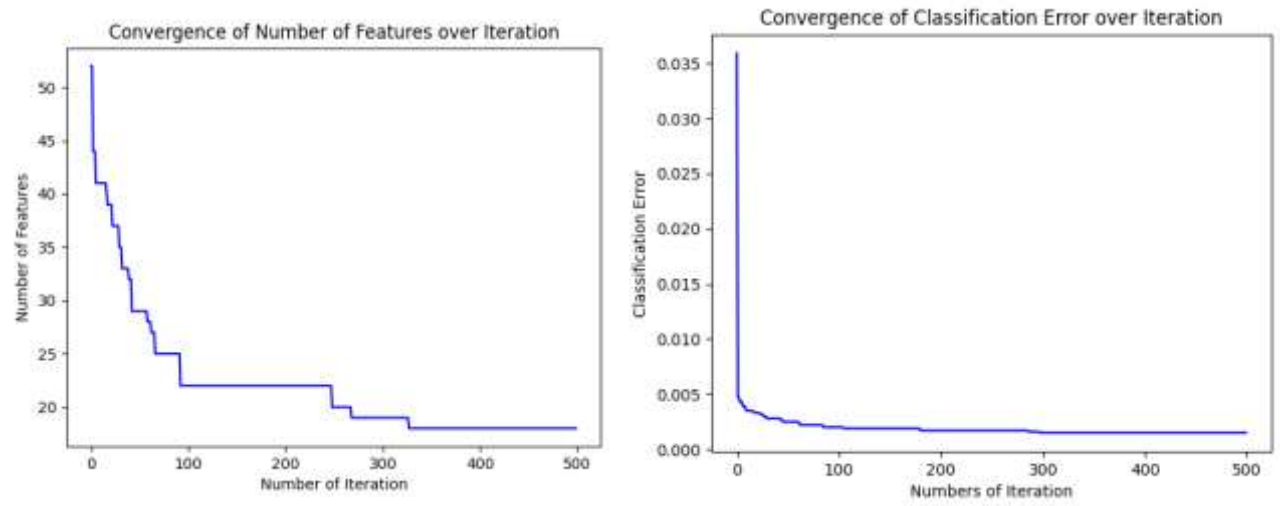
The datasets used in this paper are lung cancer II, leukaemia, SRBCT, CNS, AML-ALL-3C, AML-ALL-4C and breast datasets. Lung, leukaemia, breast and CNS are binary-class, while lymphoma, ALL-AML-3C, ALL-AML-4C and SRBCT are multiple classes. All of them are high-dimensionality microarray datasets containing samples from patients and these samples are classified into cancer classes using gene monitoring. All the datasets are found at https://csse.szu.edu.cn/staff/zhuzx/Datasets.html, while Leukaemia is available at https://zenodo.org/record/2709491#.Y_G1y3ZBzIV. **Error! Reference source not found.** lists the details such as classes, genes, samples and descriptions of the datasets used in this research.

Table 1
Description of the datasets used in this work

Dataset	Classes	Genes	Samples	Descriptions
SRBCT [50]	4	2308	83	A small-round-blue-cell tumour (SRBCT) from childhood
Breast [55]	2	24481	97	From 97 breast cancer patients, 46 developed distant metastases, while the other 51 remained disease-free for at least five years after diagnosis.
CNS [55]	2	7129	60	21 survivors and 39 treatment failures among 60 Central Nervous System (CNS) cancer patients.
ALL–AML-3C [55]	3	7129	72	AML, ALL B-cell, and ALL T-cell
ALL–AML-4C [55]	4	7129	72	AML-bone marrow, AML-peripheral blood, ALL B-cell, and T-cell.
Lung cancer II [55]	2	12533	181	Malignant Pleural Mesothelioma (MPM) and Adenocarcima (ADCA) of the lung
Leukaemia [56]	2	7129	72	Acute lymphoblastic leukaemia (AML) and acute myeloid leukaemia (ALL).
Lymphoma [57]	3	4026	62	Follicular Lymphoma (FL), Diffuse Large B-Cell Lymphoma (DLBCL), and Chronic Lymphocytic Leukaemia (CLL).

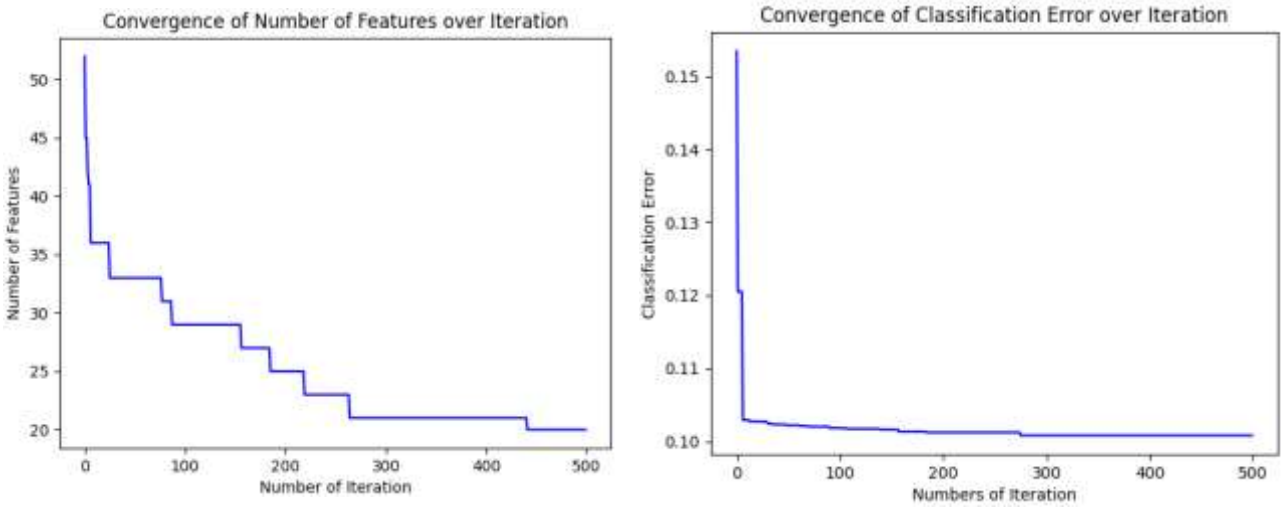
3.2 Results

The convergence graphs of the Pasi Luukka-GA + SVM method are presented in Figure 5 to Figure 12. The convergence graphs visualise how the number of features and classification errors decrease with increased iterations.



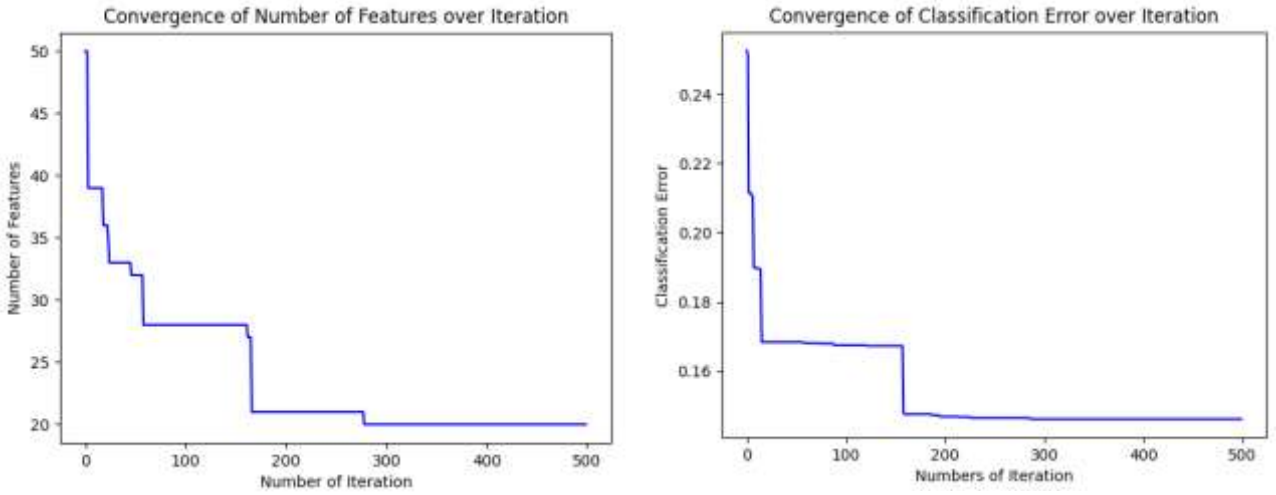
(a) (b)

Fig. 5. Convergence graphs on SRBCT [50] (a) The number of selected features over iterations (b) Classification error over iterations



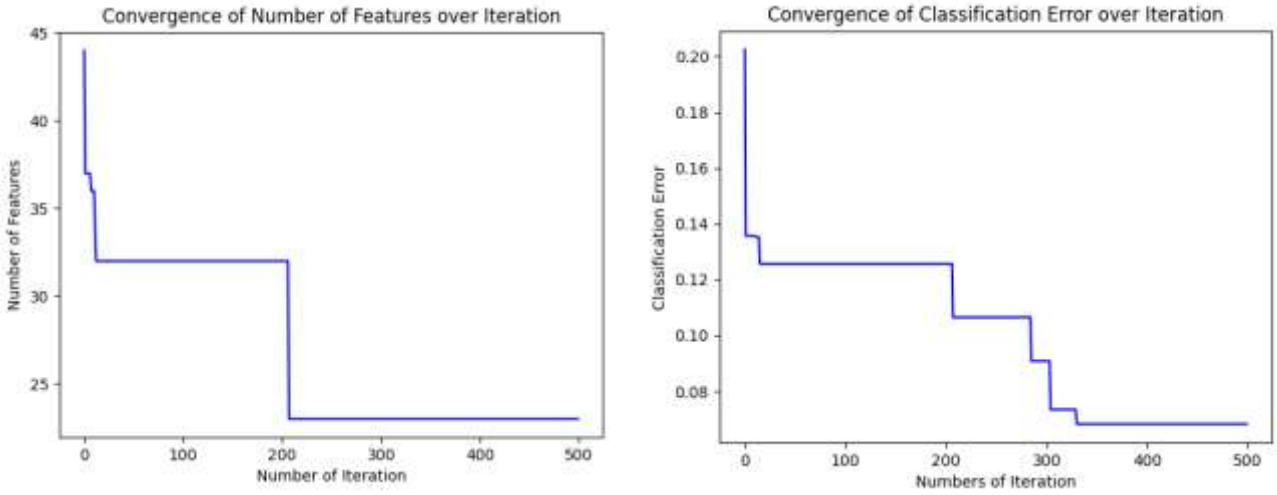
(a) (b)

Fig. 6. Convergence graphs on Breast [55] (a) The number of selected features over iterations (b) Classification error over iterations



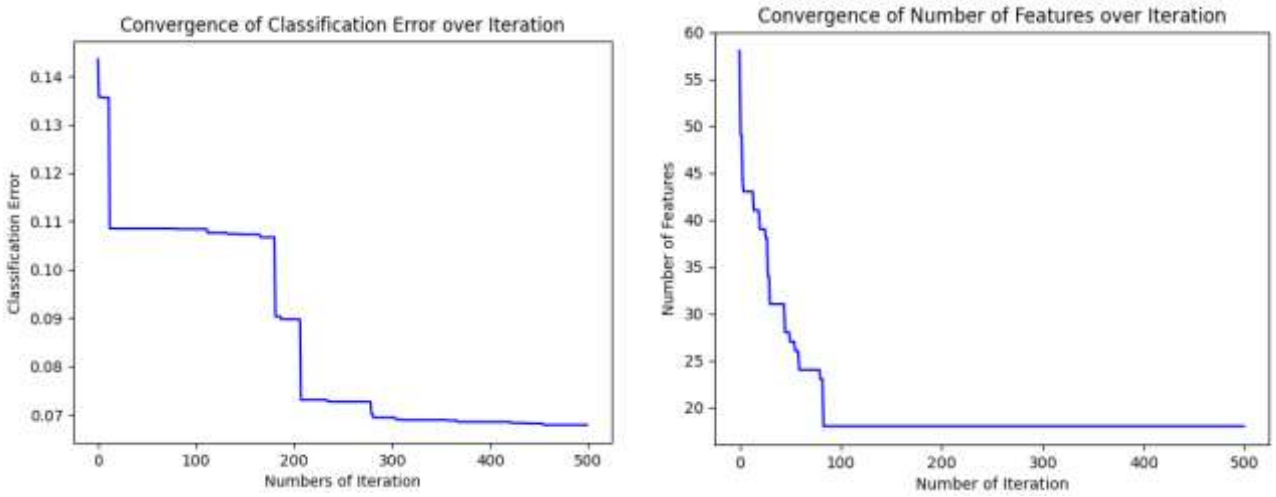
(a) (b)

Fig. 7. Convergence graphs on CNS [55] (a) The number of selected features over iterations (b) Classification error over iterations



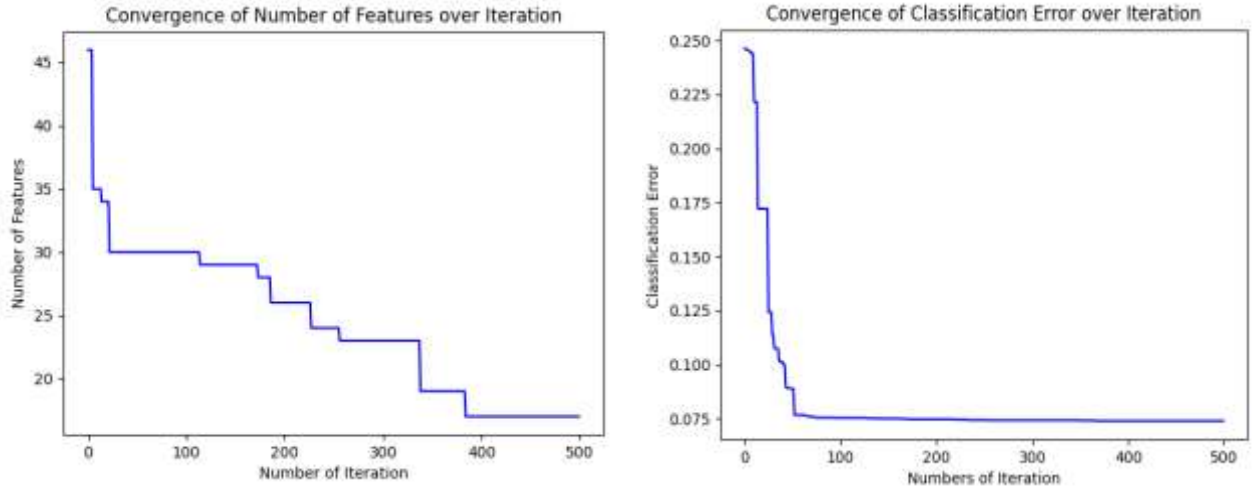
(a) (b)

Fig. 8. Convergence graphs on ALL-LMS-3C [55] (a) The number of selected features over iterations (b) Classification error over iterations



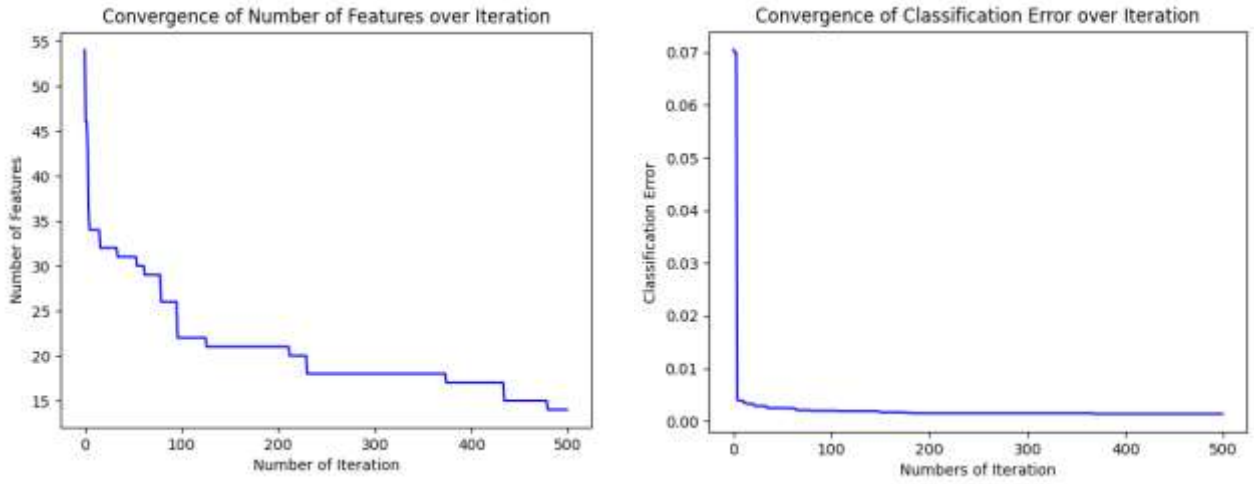
(a) (b)

Fig. 9. Convergence graphs on ALL-AML-4C [55] (a) The number of selected features over iterations (b) Classification error over iterations



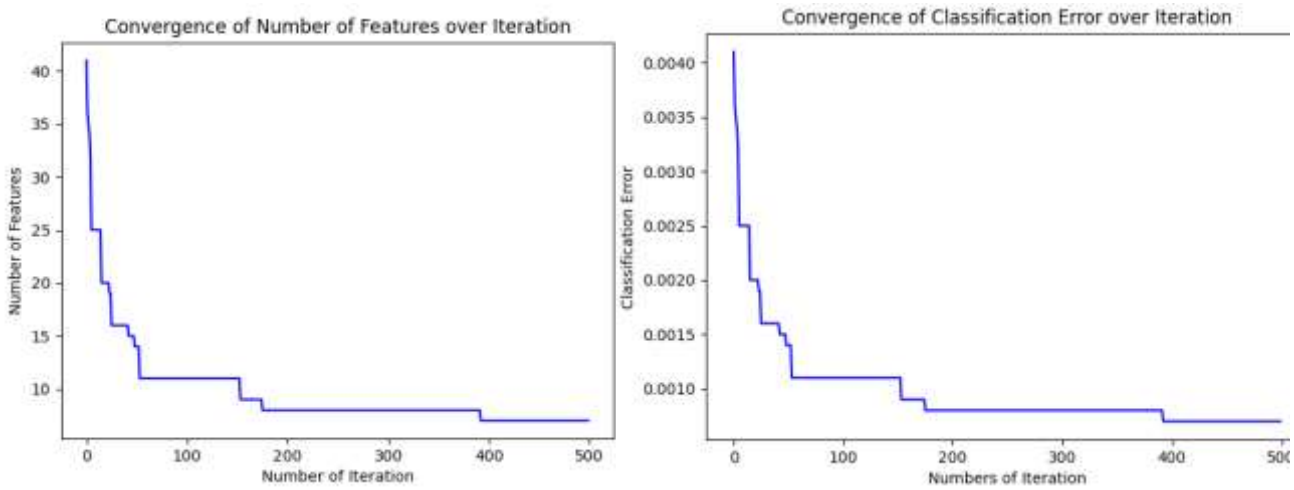
(a) (b)

Fig. 10. Convergence graphs on Lung Cancer II [55] (a) The number of selected features over iterations (b) Classification error over iterations



(a) (b)

Fig. 11. Convergence graphs on Leukaemia [56] (a) The number of selected features over iterations (b) Classification error over iterations



(a) (b)

Fig. 12. Convergence graphs on Lymphoma [57] (a) The number of selected features over iterations (b) Classification error over iterations

3.3 Number of Selected Genes

Most of the genes' list IDs are valid and can access via https://www.ncbi.nlm.nih.gov/ and https://david.ncifcrf.gov/conversion.jsp [49]. However, some genes don't have gene ID and gene symbol, such as Contig40252_RC, namely, 'UI-H-BI1-adt-g-06-0-UI.s1 NCI_CGAP_Sub3 Homo sapiens cDNA clone IMAGE:2718131 3', mRNA sequence' in breast cancer dataset [55]. This is because the provided record describes a cDNA clone without specifying a particular gene associated with it. Table 2 to Table 8 list the available selected genes from the datasets.

Table 2
Details of selected genes of SRBCT dataset [50]

No.	Gene ID	Gene Symbol	Official Full Name
51	6628	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1
137	6566	SLC16A1	solute carrier family 16 (monocarboxylic acid transporters), member 1
205	47	ACLY	ATP citrate lyase
544	1152	CKB	creatine kinase B
906	51741	WWOX	WW domain-containing oxidoreductase
1093	4609	MYC	MYC proto-oncogene, bHLH transcription factor
1167	1537	CYC1	cytochrome c-1
1424	6392	SDHD	succinate dehydrogenase complex subunit D
1924	3566	IL4R	interleukin 4 receptor
1932	3159	HMGA1	high-mobility group (nonhistone chromosomal) protein isoforms I and Y

Table 3
Details of selected genes of Breast dataset [55]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
2542	NM_012067	22977	AKR7A3	Homo sapiens aldo-keto reductase family 7 member A3 (AKR7A3), mRNA
6214	NM_012429	23541	SEC14L2	Homo sapiens SEC14-like lipid binding 2 (SEC14L2), transcript variant 1, mRNA
8776	NM_006115	23532	PRAME	Homo sapiens PRAME nuclear receptor transcriptional regulator (PRAME), transcript variant 1, mRNA
11465	M29873	1556	CYP2B7P	Cytochrome P450 family 2 subfamily B member 7, pseudogene
12413	D86976	23526	ARHGAP45	Rho GTPase activating protein 45
13152	NM_014668	9687	GREB1	Homo sapiens growth regulating estrogen receptor binding 1 (GREB1), transcript variant a, mRNA
13760	NM_015417	25876	SPEF1	Homo sapiens sperm flagellar 1 (SPEF1), mRNA
18998	NM_000157	2629	GBA1	Homo sapiens glucosylceramidase beta 1 (GBA1), transcript variant 1, mRNA
19939	NM_017680	54829	ASPN	Homo sapiens asporin (ASPN), transcript variant 1, mRNA
21356	NM_001168	332	BIRC5	Homo sapiens baculoviral IAP repeat containing 5 (BIRC5), transcript variant 1, mRNA

Table 4
Details of selected genes of CNS dataset [55]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
1057	J02854_at	10398	MYL9	myosin light chain 9
1112	J04444_at	1537	CYC1	cytochrome c1
1151	J05459_at	2947	GSTM3	glutathione S-transferase mu 3
1480	L33842_rna1_at	3615	IMPDH2	inosine monophosphate dehydrogenase 2
1719	M14200_rna1_at	1622	DBI	diazepam binding inhibitor, acyl-CoA binding protein
3210	U43328_at	1404	HAPLN1	hyaluronan and proteoglycan link protein 1
3587	U66619_at	6604	SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3
3605	U67963_at	11343	MGLL	monoglyceride lipase
4028	X02152_at	3939	LDHA	lactate dehydrogenase A
4246	X52966_at	6165	RPL35A	ribosomal protein L35a
4247	X53331_at	4256	MGP	matrix Gla protein
4778	X90840_at	547	KIF1A	kinesin family member 1A
4794	X91504_at	10139	ARFRP1	ADP ribosylation factor related protein 1
5057	Z12830_at	6745	SSR1	signal sequence receptor subunit 1
5096	Z25749_rna1_at	6201	RPS7	ribosomal protein S7

5956	J03077_s_at	5660	PSAP	prosaposin
6035	Z31560_s_at	6657	SOX2	SRY-box transcription factor 2
7098	M31667_f_at	1544	CYP1A2	cytochrome P450 family 1 subfamily A member 2

Table 5
Details of selected genes of ALL-AML-3C dataset [55]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
332	D28423_at	6428	SRSF3	serine and arginine rich splicing factor 3
1207	L05148_at	7535	ZAP70	zeta chain of T cell receptor associated protein kinase 70
1350	L16991_at	1841	DTYMK	deoxythymidylate kinase
1843	M24351_cds2_at	5744	PTHLH	parathyroid hormone like hormone
2131	M63488_at	6117	RPA1	replication protein A1
2156	M64595_at	5880	RAC2	Rac family small GTPase 2
2297	M85220_at	3492	IGH	immunoglobulin heavy locus
2456	S66793_at	407	ARR3	arrestin 3
2666	U07358_at	7786	MAP3K12	mitogen-activated protein kinase kinase kinase 12
2907	U20979_at	10036	CHAF1A	chromatin assembly factor 1 subunit A
3096	U35451_at	10951	CBX1	chromobox 1
3454	U58970_at	10953	TOMM34	translocase of outer mitochondrial membrane 34
3896	U87972_at	3420	IDH3B	isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit beta
4161	X15218_at	6497	SKI	SKI proto-oncogene

Table 6
Details of selected genes of ALL-AML-4C dataset [55]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
1339	L15388_at	2869	GRK5	G protein-coupled receptor kinase 5
1365	L19267_at	1762	DMWD	DM1 locus, WD repeat containing
1489	L34409_at	7468	NSD2	nuclear receptor binding SET domain protein 2
1805	M21389_at	3852	KRT5	keratin 5
2720	U09770_at	1396	CRIP1	cysteine rich protein 1
3469	U59878_at	10981	RAB32	RAB32, member RAS oncogene family
3969	U93049_at	2533	FYB1	FYN binding protein 1
4076	X05299_at	1059	CENPB	centromere protein B
4208	X51417_at	2103	ESRRB	estrogen related receptor beta
4847	X95735_at	7791	ZYX	zyxin
4882	X98172_at	841	CASP8	caspase 8
4928	Y00062_at	5788	PTPRC	protein tyrosine phosphatase receptor type C
4943	Y00796_at	3683	ITGAL	integrin subunit alpha L

Table 7
Details of selected genes of Lung cancer II dataset [55]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
2841	37070_at	4359	MPZ	myelin protein zero
3191	37848_at	8496	PPFIBP1	PPFIA binding protein 1
4793	32655_s_at	10902	BRD8	bromodomain containing 8
6779	39430_at	8658	TNKS	tankyrase
6944	40074_at	10797	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
7149	40518_at	5788	PTPRC	protein tyrosine phosphatase receptor type C
8014	34372_at	10075	HUWE1	HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1
8084	34801_at	9924	PAN2	poly(A) specific ribonuclease subunit PAN2
8531	36207_at	6397	SEC14L1	SEC14 like lipid binding 1
9134	38368_at	10902	BRD8	bromodomain containing 8
9366	38839_at	5217	PFN2	profilin 2

9937	40607_at	1808	DPYSL2	dihydropyrimidinase like 2
11877	835_at	11333	PDAP1	PDGFA associated protein 1
12374	318_at	8971	H1-10	H1.10 linker histone

Table 8
Details of selected genes of Leukaemia dataset [56]

No.	List ID	Gene ID	Gene Symbol	Official Full Name
485	D50913_at	23203	PMPCA	peptidase, mitochondrial processing subunit alpha
1058	J02874_at	2167	FABP4	fatty acid binding protein 4
1144	J05243_at	6709	SPTAN1	spectrin alpha, non-erythrocytic 1
1540	L38616_at	9577	BABAM2	BRISC and BRCA1 A complex member 2
1687	M11749_at	7070	THY1	Thy-1 cell surface antigen
2290	M84711_at	6189	RPS3A	ribosomal protein S3A
2707	U09411_at	7691	ZNF132	zinc finger protein 132
3036	U31628_at	3601	IL15RA	interleukin 15 receptor subunit alpha
3085	U34879_rna1_at	3292	HSD17B1	hydroxysteroid 17-beta dehydrogenase 1
3993	U96113_at	11059	WWP1	WW domain containing E3 ubiquitin protein ligase 1
4023	X01059_at	2796	GNRH1	gonadotropin releasing hormone 1
5091	Z24680_at	2615	LRRC32	leucine rich repeat containing 32
5372	S69369_at	5077	PAX3	paired box 3

Therefore, no gene ID or gene symbol can be derived from this information and these genes will be excluded from the tables. For the Lymphoma dataset [57], the gene's information was failed to be retrieved due to invalid gene list IDs which are GENE2538X, GENE2356X, GENE717X, GENE1611X, GENE1625X, GENE1519X and GENE30X.

4. Discussions

The objective of feature selection in DNA microarray analysis is to attain a high level of classification accuracy while simultaneously reducing the number of features chosen from the dataset. Thus, a high-performance feature selection method could accurately identify the most informative genes that significantly contribute to the classification performance while removing the irrelevant genes.

From Table 9, the proposed method, Pasi Luukka-GA + SVM, has a higher classification accuracy (0.9000) and a smaller number of selected features (18), compared to the current method, SU-HSA + NB hybrid method [58], at (0.8439 and 25) when performed in Breast cancer dataset [55]. The proposed method also performs better in the CNS dataset [55] than the SU-HSA + NB method [58] in terms of accuracy (0.9167) and the number of selected features (20).

For the Leukaemia dataset [56], the Pasi Luukka-GA + SVM and the Pasi Luukka-AltWOA + SVM [29] methods have reached the same accuracy (1.0000), but the proposed method has a significantly smaller number of features of 14 compared to 30. In comparison to ICA-ABC + NB [59] in the Lung Cancer II dataset [55], the proposed method in this study selected a smaller set of features (15) compared to 24, while achieving a similar classification accuracy of around 0.92.

For multiple class datasets such as ALL-AML-3C [55], when comparing SU-HSA + NB [58] to the proposed method in this work, research results show that the method used in this work has a lower accuracy (0.9333) than SU-HSA + NB (1.0000), and has a slightly lower number of selected features (23 compared to 25). The proposed method also has achieved a lower number of the selected features of 19 than 22 of the SU-HSA + NB method [58], but the classification accuracy is lower in the ALL-AML-4C dataset [55].

However, the Pasi Luukka-GA + SVM significantly performs in the lymphoma dataset [57] among the 4 methods, having the highest accuracy of 100% and the least number of selected features of 7. Lastly, the proposed method in this study achieved 100% accuracy and select 15 features from the SRBCT dataset [50], which has the best performance among the compared methods.

Overall, the Pasi Luukka-GA + SVM method performs well in the binary-class datasets, in terms of both classification accuracy and the number of selected features. However, on some multiple-class datasets, the Pasi Luukka-GA + SVM method may have slightly lower accuracy, but it generally performs well in selecting the optimal subset of genes. For all datasets, the Pasi Luukka has achieved more than 90% and even obtained 100% for leukaemia, lymphoma, and SRBCT datasets. The Pasi Luukka-GA + SVM method shows superior performance in reducing the number of features, as it selects the least number of features compared to all other methods on all datasets.

The proposed hybrid feature selection method can be improved in the filter method used. The creator of the Pasi Luukka method has proposed an advanced method on the base of the Pasi Luukka method or feature selection with a similarity classifier and entropy measures, namely, Fuzzy Similarity and Entropy measure (FSAE) feature selection [41]. The FSAE method calculates the scaled entropy based on the scaling factor determined from the difference between class means. The introduced method has overcome the deficiency and higher accuracy. Pasi Luukka-GA + SVM is implemented on the datasets mentioned in section 3.1. As a result, a comparison of the accuracy and number of selected feature subsets is presented in Table 9.

Table 9
Comparison of Pasi Luukka-GA + SVM with existing methods

Datasets used	Reference	Filter	Wrapper	Classifier	Training Accuracy	Classification Accuracy	No. of selected genes
SRBCT [50]	Jain *et al.,* [61]	Correlation-based Feature Selection (CFS)	Particle Swarm Optimization (PSO)	NB	-	1.0000	37
	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	0.9989	34
	Proposed Method	Pasi Luukka	GA	SVM	1.0000	1.0000	15
Breast [55]	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	0.8439	25
	Proposed Method	Pasi Luukka	Genetic Algorithm (GA)	SVM	0.9091	0.9000	18
CNS [55]	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	0.8244	32
	Proposed Method	Pasi Luukka	GA	SVM	0.8542	0.9167	20
ALL-AML-3C [55]	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	1.0000	25
	Proposed Method	Pasi Luukka	GA	SVM	0.9474	0.9333	23
ALL-AML-4C [55]	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	0.9743	22
	Proposed Method	Pasi Luukka	GA	SVM	0.9474	0.9333	19
Lung cancer II [55]	Aziz *et al.,* [59]	Independent component analysis (ICA)	Artificial Bee Colony (ABC)	NB	-	0.9245	24

	Proposed method	Pasi Luukka	GA	SVM	0.9383	0.9268	15
Leukaemia [56]	Kundu *et al.,* [29]	Pasi Luukka	Altruistic Whale Optimization Algorithm (AltWOA)	SVM	-	1.0000	30
	Proposed method	Pasi Luukka	GA	SVM	1.0000	1.0000	14
Lymphoma [57]	Moradi *et al.,* [60]	Probabilistic random function	Particle Swarm Optimization	KNN	-	0.8771	50
	Jain *et al.,* [61]	Correlation-based Feature Selection (CFS)	Particle Swarm Optimization (PSO)	NB	-	1.0000	24
	Shreem *et al.,* [58]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	-	1.0000	10
	Proposed Method	Pasi Luukka	GA	SVM	1.0000	1.0000	7

5. Conclusions

In conclusion, the Pasi Luukka-GA + SVM is a newly proposed hybrid feature selection method that outperforms all the compared methods on all datasets. This study shows that the Pasi Luukka-GA + SVM is a promising approach in the feature selection process in DNA microarray analysis. The method only focused on the mention method and parameter. Therefore, this study provides a room/improvement for future research, enabling further advancements and refinement in the field by the scientific community/other researchers.

Acknowledgement
This research was supported by Universiti Teknikal Malaysia Melaka.

References
[1] National Cancer Institute Institute. "What Is Cancer?," *National Cancer Institute*. (2021). https://www.cancer.gov/about-cancer/understanding/what-is-cancer
[2] World Health Organization. "Cancer." *World Health Organization*, https://www.who.int/health-topics/cancer
[3] Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. "Cancer statistics, 2022." *CA: a cancer journal for clinicians* 72, no. 1 (2022): 7-33. https://doi.org/10.3322/caac.21708
[4] Gupta, Surbhi, Manoj K. Gupta, Mohammad Shabaz, and Ashutosh Sharma. "Deep learning techniques for cancer classification using microarray gene expression data." *Frontiers in physiology* 13 (2022): 952709. https://doi.org/10.3389/fphys.2022.952709
[5] Kumar, Awanish, Satish Chandra Pandey, and Mukesh Samant. "DNA-based microarray studies in visceral leishmaniasis: identification of biomarkers for diagnostic, prognostic and drug target for treatment." *Acta Tropica* 208 (2020): 105512. https://doi.org/10.1016/j.actatropica.2020.105512
[6] Necsulea, Anamaria, and Henrik Kaessmann. "Evolutionary dynamics of coding and non-coding transcriptomes." *Nature Reviews Genetics* 15, no. 11 (2014): 734-748. https://doi.org/10.1038/nrg3802
[7] Vajdi, Amir, Nurit Haspel, and Hadi Banaee. "A new dp algorithm for comparing gene expression data using geometric similarity." In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1157-1161. IEEE, 2015. https://doi.org/10.1109/BIBM.2015.7359846
[8] Narrandes, Shavira, and Wayne Xu. "Gene expression detection assay for cancer clinical use." *Journal of Cancer* 9, no. 13 (2018): 2249. https://doi.org/10.7150/jca.24744
[9] Parthasarathy, S. "Bioinformatics: Application to genomics." *Plant Biology and Biotechnology: Volume II: Plant Genomics and Biotechnology* (2015): 279-300. https://doi.org/10.1007/978-81-322-2283-5_13
[10] Russo, Giuseppe, Charles Zegar, and Antonio Giordano. "Advantages and limitations of microarray technology in human cancer." *Oncogene* 22, no. 42 (2003): 6497-6507. https://doi.org/10.1038/sj.onc.1206865

[11] Almugren, Nada, and Hala Alshamlan. "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification." *IEEE access* 7 (2019): 78533-78548. https://doi.org/10.1109/ACCESS.2019.2922987
[12] Guyon, Isabelle, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh, eds. *Feature extraction: foundations and applications*. Vol. 207. Springer, 2008.
[13] Bolón-Canedo, Verónica, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. "A review of microarray datasets and applied feature selection methods." *Information sciences* 282 (2014): 111-135. https://doi.org/10.1016/j.ins.2014.05.042
[14] Vogelstein, Joshua T., Eric W. Bridgeford, Minh Tang, Da Zheng, Christopher Douville, Randal Burns, and Mauro Maggioni. "Supervised dimensionality reduction for big data." *Nature communications* 12, no. 1 (2021): 2872. https://doi.org/10.1038/s41467-021-23102-2
[15] Guo, Shun, Donghui Guo, Lifei Chen, and Qingshan Jiang. "A L1-regularized feature selection method for local dimension reduction on microarray data." *Computational biology and chemistry* 67 (2017): 92-101. https://doi.org/10.1016/j.compbiolchem.2016.12.010
[16] Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." *Advances in bioinformatics* 2015, no. 1 (2015): 198363. https://doi.org/10.1155/2015/198363
[17] Saeys, Yvan, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." *bioinformatics* 23, no. 19 (2007): 2507-2517. https://doi.org/10.1093/bioinformatics/btm344
[18] Wei, Guangfen, Jie Zhao, Yanli Feng, Aixiang He, and Jun Yu. "A novel hybrid feature selection method based on dynamic feature importance." *Applied Soft Computing* 93 (2020): 106337. https://doi.org/10.1016/j.asoc.2020.106337
[19] Barbieri, Matheus Cezimbra, Bruno Iochins Grisci, and Márcio Dorn. "Analysis and comparison of feature selection methods towards performance and stability." *Expert Systems with Applications* (2024): 123667. https://doi.org/10.1016/j.eswa.2024.123667
[20] Jahangiri, Sonia, Masoud Abdollahi, Ehsan Rashedi, and Nasibeh Azadeh-Fard. "A machine learning model to predict heart failure readmission: toward optimal feature set." *Frontiers in Artificial Intelligence* 7 (2024): 1363226. https://doi.org/10.3389/frai.2024.1363226
[21] Arram, Anas, Masri Ayob, Musatafa Abbas Abbood Albadr, Dheeb Albashish, and Alaa Sulaiman. "A hybrid of an automated multi-filter with a spatial bound particle swarm optimization for gene selection and cancer classification." *Heliyon* 11, no. 5 (2025). https://doi.org/10.1016/j.heliyon.2025.e42544
[22] Ergul Aydin, Zeliha, and Zehra Kamisli Ozturk. "Filter-based feature selection methods in the presence of missing data for medical prediction models." *Multimedia Tools and Applications* 83, no. 8 (2024): 24187-24216. https://doi.org/10.1007/s11042-023-15917-6
[23] Lazar, Cosmin, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. "A survey on filter techniques for feature selection in gene expression microarray analysis." *IEEE/ACM transactions on computational biology and bioinformatics* 9, no. 4 (2012): 1106-1119. https://doi.org/10.1109/TCBB.2012.33
[24] Bommert, Andrea, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. "Benchmark for filter methods for feature selection in high-dimensional classification data." *Computational Statistics & Data Analysis* 143 (2020): 106839. https://doi.org/10.1016/j.csda.2019.106839
[25] Vijayasarveswari, V., Norfadila Mahrom, Rafikha Aliana A. Raof, Phak Len Al Eh Kan, Muhammad Amiruddin Ab Razak, Bavanraj Punniya Silan, Ahmad Ashraf Abdul Halim, Mohd Wafi Nasrudin, Nuraminah Ramli, and Yusnita Rahayu. "Development of Statistically Modelled Feature Selection Method for Microwave Breast Cancer Detection." In *The 2nd International Conference on Engineering and Technology (ICoEngTech 2023)*. 2023.
[26] Luukka, Pasi. "Feature selection using fuzzy entropy measures with similarity classifier." *Expert Systems with Applications* 38, no. 4 (2011): 4600-4607. https://doi.org/10.1016/j.eswa.2010.09.133
[27] Pudjihartono, Nicholas, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O'Sullivan. "A review of feature selection methods for machine learning-based disease risk prediction." *Frontiers in Bioinformatics* 2 (2022): 927312. https://doi.org/10.3389/fbinf.2022.927312
[28] Pudjihartono, N., T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan. "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction." *Frontiers in Bioinformatics* 2 (2022): 927312-927312. https://doi.org/10.3389/fbinf.2022.927312
[29] Kundu, Rohit, Soham Chattopadhyay, Erik Cuevas, and Ram Sarkar. "AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets." *Computers in biology and medicine* 144 (2022): 105349. https://doi.org/10.1016/j.compbiomed.2022.105349
[30] Zhao, Hong, and Shenglong Yu. "Cost-sensitive feature selection via the ℓ2, 1-norm." *International Journal of Approximate Reasoning* 104 (2019): 25-37. https://doi.org/10.1016/j.ijar.2018.10.017

[31] Mohd Ali, Nursabillilah, Rosli Besar, and Nor Azlina Ab. Aziz. "Hybrid feature selection of breast cancer gene expression microarray data based on metaheuristic methods: A comprehensive review." *Symmetry* 14, no. 10 (2022): 1955. https://doi.org/10.3390/sym14101955

[32] Ali, Nursabillilah Mohd, Ainain Nur Hanafi, Mohd Safirin Karis, Nur Hazahsha Shamsudin, Ezreen Farina Shair, and Nor Hidayati Abdul Aziz. "Hybrid feature selection of microarray prostate cancer diagnostic system." *Indonesian Journal of Electrical Engineering and Computer Science* 36, no. 3 (2024): 1884-1894. https://doi.org/10.11591/ijeecs.v36.i3.pp1884-1894

[33] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." *Journal of machine learning research* 3, no. Mar (2003): 1157-1182.

[34] Solorio-Fernández, Saúl, J. Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. "A new hybrid filter–wrapper feature selection method for clustering based on ranking." *Neurocomputing* 214 (2016): 866-880. https://doi.org/10.1016/j.neucom.2016.07.026

[35] Taheri, Nooshin, and Hossein Nezamabadi-pour. "A hybrid feature selection method for high-dimensional data." In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 141-145. IEEE, 2014. https://doi.org/10.1109/ICCKE.2014.6993381

[36] Yilin, T. U., Tsuyoshi Inoue, Shota Yabui, Keiichi Katayama, and Shigeyuki Tomimatsu. "Hybrid feature selection method for SVM classification and its application for fault diagnosis of wear and peeling in journal bearing with a little muddy water using long-term real data." *Journal of Low Frequency Noise, Vibration and Active Control* 42, no. 1 (2023): 231-252. https://doi.org/10.1177/14613484221118997

[37] Pirgazi, Jamshid, Mohsen Alimoradi, Tahereh Esmaeili Abharian, and Mohammad Hossein Olyaee. "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets." *Scientific reports* 9, no. 1 (2019): 18580. https://doi.org/10.1038/s41598-019-54987-1

[38] Got, Adel, Abdelouahab Moussaoui, and Djaafar Zouache. "Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach." *Expert Systems with Applications* 183 (2021): 115312. https://doi.org/10.1016/j.eswa.2021.115312

[39] Duque, Jorge, Antonio Godinho, Jose Moreira, and José Vasconcelos. "Data Science with Data Mining and Machine Learning A design science research approach." *Procedia Computer Science* 237 (2024): 245-252. https://doi.org/10.1016/j.procs.2024.05.102

[40] Ranjan, Rajesh, and Jitender Kumar Chhabra. "A Hybrid Feature Selection Approach on Medical Dataset." *IETE Journal of Research* (2025): 1-10. https://doi.org/10.1080/03772063.2025.2493784

[41] Lohrmann, Christoph, Pasi Luukka, Matylda Jablonska-Sabuka, and Tuomo Kauranne. "A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection." *Expert Systems with Applications* 110 (2018): 216-236. https://doi.org/10.1016/j.eswa.2018.06.002

[42] Chopard, Bastien, and Marco Tomassini. *An introduction to metaheuristics for optimization*. Vol. 226. Cham, Switzerland: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-93073-2

[43] Raidl, Günther R., Jakob Puchinger, and Christian Blum. "Metaheuristic hybrids." In *Handbook of metaheuristics*, pp. 385-417. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-91086-4_12

[44] Nssibi, Maha, Ghaith Manita, and Ouajdi Korbaa. "Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey." *Computer Science Review* 49 (2023): 100559. https://doi.org/10.1016/j.cosrev.2023.100559

[45] Katoch, Sourabh, Sumit Singh Chauhan, and Vijay Kumar. "A review on genetic algorithm: past, present, and future." *Multimedia tools and applications* 80 (2021): 8091-8126. https://doi.org/10.1007/s11042-020-10139-6

[46] Schmitt, Lothar M. "Theory of genetic algorithms." *Theoretical Computer Science* 259, no. 1-2 (2001): 1-61. https://doi.org/10.1016/S0304-3975(00)00406-0

[47] McCall, John. "Genetic algorithms for modelling and optimisation." *Journal of computational and Applied Mathematics* 184, no. 1 (2005): 205-222. https://doi.org/10.1016/j.cam.2004.07.034

[48] Noreen, Shahzadi, Aamir Shahzad, Safa Akhtar, and Farah Deeba. "Employing an integrated bioinformatics and systems biology approach to uncover key genes and drug targets for ovarian cancer." *Human Gene* (2025): 201408. https://doi.org/10.1016/j.humgen.2025.201408

[49] Sherman, Brad T., Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W. Baseler, H. Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)." *Nucleic acids research* 50, no. W1 (2022): W216-W221. https://doi.org/10.1093/nar/gkac194

[50] Khan, Javed, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold *et al.,* "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature medicine* 7, no. 6 (2001): 673-679. https://doi.org/10.1038/89044

[51] Lee, Yoonkyung, and Cheol-Koo Lee. "Classification of multiple cancer types by multicategory support vector machines using gene expression data." *Bioinformatics* 19, no. 9 (2003): 1132-1139. https://doi.org/10.1093/bioinformatics/btg102

[52] Zhang, Yongli. "Support vector machine classification algorithm and its application." In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3*, pp. 179-186. Springer Berlin Heidelberg, 2012.

[53] Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. "A comprehensive survey on support vector machine classification: Applications, challenges and trends." *Neurocomputing* 408 (2020): 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[54] Amami, Rimah, Dorra Ben Ayed, and Noureddine Ellouze. "An Empirical comparison of SVM and some supervised learning algorithms for vowel recognition." *arXiv preprint arXiv:1507.06021* (2015).

[55] Zhu, Zexuan, Yew-Soon Ong, and Manoranjan Dash. "Markov blanket-embedded genetic algorithm for gene selection." *Pattern Recognition* 40, no. 11 (2007): 3236-3248. https://doi.org/10.1016/j.patcog.2007.02.007

[56] Golub, Todd R., Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller *et al.,* "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286, no. 5439 (1999): 531-537. https://doi.org/10.1126/science.286.5439.531

[57] Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick *et al.,* "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403, no. 6769 (2000): 503-511. https://doi.org/10.1038/35000501

[58] Shreem, Salam Salameh, Salwani Abdullah, and Mohd Zakree Ahmad Nazri. "Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm." *International Journal of Systems Science* 47, no. 6 (2016): 1312-1329. https://doi.org/10.1080/00207721.2014.924600

[59] Aziz, Rabia, C. K. Verma, and Namita Srivastava. "A novel approach for dimension reduction of microarray." *Computational biology and chemistry* 71 (2017): 161-169. https://doi.org/10.1016/j.compbiolchem.2017.10.009

[60] Moradi, Parham, and Mozhgan Gholampour. "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy." *Applied soft computing* 43 (2016): 117-130. https://doi.org/10.1016/j.asoc.2016.01.044

[61] Jain, Indu, Vinod Kumar Jain, and Renu Jain. "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification." *Applied Soft Computing* 62 (2018): 203-215. https://doi.org/10.1016/j.asoc.2017.09.038