International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185
https://www.inderscience.com/ijcse

Minimum redundancy maximum relevance and VNS based gene selection for cancer classification in high-dimensional data

Ahmed Bir-Jmel, Sidi Mohamed Douiri, Souad Elbernoussi

DOI: 10.1504/IJCSE.2024.10061862

Article History:

Received:	23 May 2022
Last revised:	01 September 2022
Accepted:	20 September 2022
Published online:	25 January 2024

Minimum redundancy maximum relevance and VNS based gene selection for cancer classification in high-dimensional data

Ahmed Bir-Jmel*, Sidi Mohamed Douiri and Souad Elbernoussi

Department of Mathematics,
Faculty of Sciences,
Laboratory of Mathematics, Computer Science and Applications-Security of Information,
Mohammed V University in Rabat, Morocco
Email: birjmel@gmail.com
Email: douirisidimohamed@gmail.com
Email: souad.elbernoussi@gmail.com
*Corresponding author

Abstract: DNA microarray is a technique for measuring simultaneously the expression levels of a huge number of genes, these levels have a significant impact on cancer classification tasks. In DNA datasets, the number of genes exceeds the number of samples that make the presence of irrelevant or redundant genes possible. In this paper, two hybrid multivariate filters for gene selection, named VNSMI and VNSCor, are presented. These methods surpass the univariate filters by considering the possible interaction between genes through the search for a subset of genes that contains the minimum redundancy and the maximum relevance (MRMR). In the first stage of our approaches, we use a univariate filter by selecting the best-ranked genes. Then, we apply the variable neighbourhood search (VNS) metaheuristic coupled with an innovative stochastic local search (SLS) algorithm to find the final subset of genes. The experiments performed show that the proposed approaches are feasible and effective.

Keywords: gene selection; feature selection; cancer classification; VNS; stochastic local search; SLS; normalised mutual information; MRMR; DNA microarray.

Reference to this paper should be made as follows: Bir-Jmel, A., Douiri, S.M. and Elbernoussi, S. (2024) 'Minimum redundancy maximum relevance and VNS based gene selection for cancer classification in high-dimensional data', *Int. J. Computational Science and Engineering*, Vol. 27, No. 1, pp.78–89.

Biographical notes: Ahmed Bir-Jmel received his PhD degree in Applied Mathematics from Mohammed V University in Rabat, Faculty of Sciences, Morocco, in 2022. He received his Master degree in the field of Mathematical Engineering and Modelling (2016), from the same university. His research interests include machine learning, data mining, metaheuristic algorithms, optimisation and convex analysis.

Sidi Mohamed Douiri is a Full Professor in Department of Mathematics at the Faculty of Sciences, Rabat, Morocco, and a member of the French Society of Operational Research and Decision Support (ROADEF). He received his PhD in Applied Mathematics, Optimisation and Operational Research from Mohammed V. University, Rabat. He has participated in many research projects and published several papers in international journals and conferences. His research focuses on mathematical and computer optimisation, the solution of large-scale combinatorial search problems and on the design of effective heuristic and metaheuristic methods for practical applications.

Souad Elbernoussi is a full Professor at Mohammed-V University in Rabat, Morocco, and the Director of the Laboratory of Mathematics, Computing and Applications-Security of Information (LabMIA-SI). Her research interests include convex and non-convex optimisation, combinatorial optimisation, interior point methods, semidefinite programming, multi-criteria systems, heuristic and metaheuristic methods, graph colouring, localisation in graphs, localisation in sensor networks, transport problems, scheduling problems, energy efficiency, cryptography, classification, credit scoring, data mining, machine learning.

1 Introduction

In recent years, microarray technology has become one of the most important advances in collecting data on biological systems. This technology allowed researchers to diagnose whether a sample came from cancerous, benign tissue, or distinguish between different types of cancerous tumours. This diagnisis is performed by simultaneously measuring a huge number of genes in a biological sample (Schena et al., 1995; Peng et al., 2010). Microarray datasets are presented as a matrix of M rows (samples) and N columns (genes). While the information on the classes of the samples is already available (Bir-jmel et al., 2019), the analysis of DNA microarray for the cancer classification is formulated in our work as a supervised classification problem (Kumari and Swarnkar, 2011). Moreover, microarray datasets used in the context of cancer classification are characterised by a very large number of genes, while that of samples is small, and only a few genes are sufficient for classification (Li and Yang, 2002; Xiong et al., 2001). For that, Golub et al. (2012) show that 50 genes are generally sufficient to model a 38-sample binary classification problem.

Dimensionality reduction is therefore an essential step in the data preprocessing process. A reduction method consists in finding a representation of the initial data in a smaller space, there are mainly two different reduction approaches: feature extraction, and feature selection or gene selection which consists of selecting the genes most relevant from the dataset (Ma and Zhu, 2013). The latter reduction consists of replacing the initial set of data with a new reduced set, built from the combinations of the initial features (Maafiri et al., 2021, 2022). From a biological point of view, for cancer classification, it is more efficient to select real genes than to create artificial features with uncertain biological significance. Therefore, selecting a small subset of genes containing the necessary information for a given disease is one of the main goals of microarray data analysis (Li et al., 2008). This selection allows a better interpretation of the biological relationship between the genes and the clinical result, also a better scientific understanding of the problem posed. Moreover, it overcomes the dimensionality problem to improve the quality of classifiers by increasing the accuracy of cancer prediction. Therefore, in our paper, we address the problem of gene selection in the context of microarray classification. In general, there are three selection methods common to many taxonomies depending on the evaluation criterion and their interaction with the classifier: filter models, wrapper models, and embedded models (Tang et al., 2014; Kohavi and John, 1997). For filters, the quality of a subset of genes is measured independently of any learning algorithm. Generally, any filtering method consists of ranking genes in order of importance with respect to a specific evaluation criterion, which can be either univariate or multivariate. In the univariate methods, each gene is scored independently and the selection is done by selecting the best-ranked genes. On the other hand, the multivariate scheme takes into consideration the interaction between genes and it is capable of handling redundant genes (Tabakhi and Moradi, 2015; Yu and Liu, 2003). Signal-to-noise-ratio (Mishra and Sahu, 2011), T-test (Jafari and Azuaje, 2006), information gain (IG) (Bonev et al., 2008), Fischer score (Gu et al., 2012), and Relief-F (Shreem et al., 2012) are the most representative algorithms of filter model. While in wrappers methods, the selection depends on the classifier used. Indeed, these approaches wrap gene selection around the learning algorithm, and often they use the classification accuracy as criteria for evaluating subsets of genes each of which will be used to train the classification model. Generally, these methods give better results compared to filter methods because they take into account the dependency between genes and the intrinsic bias of the classification algorithm. However, they are less general than filter methods because they must be re-executed if another learning algorithm is used. So there is no guarantee that the solution is optimal for other classifiers. For that, these methods require a significant computation time to reach convergences and can be insoluble for large problems (Bir-Jmel et al., 2019; Kohavi and John, 1997). In addition, the embedded methods where the gene selection procedure is performed during the construction of the classifier (Peng et al., 2010).

In order to solve the gene selection problem, we search for a 'good' subset of genes according to a fitness function, this potentially requires looking at $2^N - 1$ subsets, and the number of genes N which is large enough in microarrays, then evolutionary algorithms have attracted much attention to solve this problem which is considered NP-Hard (Amaldi and Kann, 1998). The common goal of metaheuristic algorithms is to find nearly optimal solutions for a given problem (Douiri and El Bernoussi, 2012), for example, variable neighbourhood search (VNS) is considered a metaheuristic that can be used for combinatorial optimisation in which an optimal solution is searched on a discrete search space. Especially the gene selection problem, while there is no better algorithm in data mining. Several researchers have used metaheuristic methods to solve this problem including VNS (Pacheco et al., 2007; Garcıa-Torres et al., 2004, 2016; Pentelas et al., 2019), ACO (Chiang et al., 2008; Li et al., 2013; Sharbaf et al., 2016; Tabakhi and Moradi, 2015; Bir-Jmel et al., 2019), PSO (Li et al., 2008; Bir-Jmel et al., 2019; Chuang et al., 2008), genetic algorithm (Li et al., 2008; Lee and Leu, 2011), and bat-inspired algorithm (BA) (Alomari et al., 2017).

Recently, with the continuous progress in combinatorial optimisation, intelligent algorithms are proposed and applied for gene selection. For that, Dabba et al. (2021) proposed a modified moth flame algorithm (mMFA), combined with mutual information maximisation (MIM) to solve gene selection problems. In the same context, a hybrid gene selection method based on a novel multi-filter ensemble technique and simplified swarm optimisation (SSO) is proposed in Lai and Huang (2021). Zhang et al. (2021) combined the IG and a modified fruit fly optimisation algorithm (FOA) to choose the relevant genes for improving classification performances. In another

method called GI-SVM-RFE, a hybrid gene selection was proposed, which combines the Gini index and SVM with recursive gene elimination (Almutiri and Saeed, 2022). While in Rostami et al. (2022) the authors proposed an innovative multi-objective graph theoretic-based method for gene selection for microarray data classification. Some works applied VNS to solve the feature selection problem and obtained promising results compared with the state-of-the-art for gene selection for cancer classification (Pacheco et al., 2007; Garcıa-Torres et al., 2004, 2016; Pentelas et al., 2022; Helder et al., 2022). Therefore, the goal is to adapt VNS as an optimisation technique to solve the gene selection problem, while the fitness function used is maximum minimum redundancy relevance (MRMR) which is introduced in Ding and Peng (2005). On the other hand, a stochastic local search (SLS) algorithm is adopted to refine the solutions and avoid premature convergence. The use of MRMR allows the selection of genes that have the greatest relevance for the target class and not redundant ones. Thus, the disadvantages of classical filter methods (univariate) is solved by taking into consideration possible interactions between genes as well as low-scoring genes. For VNS (Mladenovic and Hansen, 1997; Hansen and Mladenovic, 2001) which is a recent metaheuristic designed to solve combinatorial optimisation problems, we start with an initial solution and we try to improve it by exploring the search space based on multiple neighbourhood structures (shake procedure) via wellchosen local search method (improvement procedure) to escape the local optima. In order to solve the gene selection problem two hybrid approaches are proposed. The proposed approach start by using univariate filter to select the p-best genes based on their relevance. As second step, a hybrid method based on the VNS metaheuristic and a SLS algorithm are applied to the p-best genes selected in the previous step. The first proposed method (VNSMI) is based on some concepts of information theory. While the second (VNSCor) is based on the Pearson correlation coefficient (PCC). The contributions in this paper can be summarised as follows:

- new two-stage method for gene selection combining univariate and multivariate schemes
- the multivariate scheme, VNS metaheuristic is adopted using SLS algorithm to find near optimal subset of genes for cancer classification
- the mutual information and the PCC are used to construct the fitness function
- the experiments results evaluated on several datasets while the obtained results are effective in terms of classification accuracy and number of genes picked.

The paper sections are organised as follows: In the first section, we describe the methods. The second section analyses the proposed method in details. And finally, in the last section, some experiments on six well-replicated microarray datasets are presented.

2 Methods

2.1 Notations

In this study, $D = (x_1; x_2; \ldots ; x_M)$, $x_i \in \mathbb{R}^N$ presents dataset of M samples. While $\{g_1, g_2 \ldots g_N\}$, $g_i \in \mathbb{R}^M$ denote the N genes vectors and $C = (y_1; y_2; ...; y_M)$ denote the class labels.

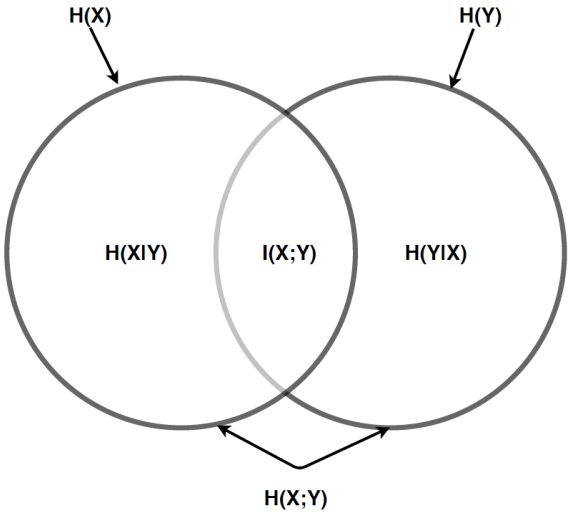
2.2 Mutual information

The mutual information is a useful metric in gene selection problem for cancer classification (Bonev et al., 2008; Ding and Peng, 2005; Lai et al., 2016). For that, based on the information theory we proposed the first proposed *VNSMI* method while the columns of our matrix D (genes and class label) can be considered as random variables. Indeed, the statistical dependence of two random variables X, Y is the mutual information between them. It is often measured in bits, and is defined as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X \mid Y) - H(Y \mid X). \end{aligned}$$

where $H(Y)$ and H(X) denotes the measurement of uncertainty of Y and X respectively. The conditional entropies of X given Y and Y given X, respectively, are $H(X|Y)$ and $H(Y|X)$, while the joint entropy in X and Y is $H(X, Y)$. The relationship between the entropy and the mutual information can be interpreted in Figure 1.

Figure 1 The association between the entropy and the mutual information



In the discrete context, according to Shannon's description of entropy, we have:

$$H(X) = -\sum_{x \in X} p(x) \log(p(x))$$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log(p(x,\ y))$$

where $p(x, y)$ denotes the probabilistic joint distribution of the variables X and Y, and $(p(x), p(y))$ denotes the respective marginal probabilities.

The summation is replaced by a definite double integrals in the continuous case:

$$I(X;\ Y) = \int_Y \int_x p(x,\ y) \log\left(\frac{p(x,\ y)p(x,\ y)}{p(x)p(y)} \right) dxdy$$

The properties of mutual information are as follows:

- $I(X; X) = H(X)$
- $I(X; Y) = 0$ if and only if X and Y are independent
- The mutual information is positive or null $I(X;\ Y) \geq 0$
- The mutual information is symmetric $I(X; Y) = I(Y; X)$.

The normalised mutual information (NI) between g_i and g_s $NI\ (g_i; g_s)$, is defined as (Estévez et al., 2009).

$$NI\left(g_i;\ g_s\right) = \frac{I\left(g_i; g_s\right)}{\min\left\{H(g_i), H(g_s)\right\}.}$$

Thus, we assign a relevance score of each gene g_i to the class label C, this metric $SNMI_i$ is defined as:

$$SNMI_i = \frac{NI\left(g_i; C\right)}{\max_{j\in 1,..,N} NI\left(g_j; C\right)}.$$

This score of relevance is inspired from the IG algorithm (Lai et al., 2016).

The calculation of the mutual information is based on the estimation of the probability density and joint probability of the variables. In our research, we use Peng's mutual information MATLAB toolbox (Peng, 2022) to do this estimation.

2.3 Pearson correlation coefficient

Our second proposed method *VNSCor* is based on the correlation coefficient, which is a measure of the strength of the linear relationship between two variables (genes). Let g_1 and g_2 be two random variables, the coefficient of correlation between g_1 and g_2 is defined as follows:

$$\rho_{g_1, g_2} = \frac{\operatorname{cov}\left(g_1,\ g_2\right)}{\sigma_{g_1}\sigma_{g_2}} \tag{1}$$

where $cov(g_1, g_2)$ is the covariance between g_1 and g_2, σ_{g_1} is the standard deviation of g_1 and σ_{g_2} the standard deviation of g_2.

The correlation value may take on a range of values from −1 to 0 to +1, let ($r_{ij} = |\rho g_i, g_j|$) the absolute value of the correlation between g_i and g_j.

Based on this notion, we can define a measure of relevance SR_i of each gene g_i to the class label by

$$SR_i = \frac{\left|\rho_{g_i}, C\right|}{\max_{j\in 1,..,N}\left|\rho_{g_j}, C\right|}.$$

3 VNS

VNS is a new metaheuristic for solving combinatorial and global optimisation problems, presented by Mladenovi and Hansen in 1997 (Mladenovic and Hansen, 1997; Hansen and Mladenovic, 2001). This method can be broken down into two stages: a stage in which a local search converges towards a local optimum, and another stage that makes it possible to escape from it.

This metaheuristic uses the following concepts:

- Systematic change of neighbourhood to extract local optimums.
- Jump from one solution to another if only if there is an improvement.
- Using local search to get a local optimum.

Among the recent works in which the VNS metaheuristic has been proposed to solve feature selection problems, we find (Pacheco et al., 2007; Garcıa-Torres et al., 2004, 2016; Pentelas et al., 2022).

The process of VNS starts with an initial solution and a predefined finite number of neighbourhoods N_k, k =1, …, k_{max}, at each iteration, a new candidate solution (Shaking step) is generated from the current neighbourhood, and then the current solution is refined and improved using a local search algorithm. If this new solution s0 is better than the previous one, the process resumes with the first neighbourhood, in the opposite case the same steps are repeated but passing to the next neighbourhood.

3.1 Basic VNS [Hansen and Mladenovic (2001), Figure 2] initialisation

Select an initial candidate solution s, and build the neighbourhood structures phase N_k, k = 1, …, k_{max}, that will be used in our search phase.

Select a stopping criterion.

Set k = 1

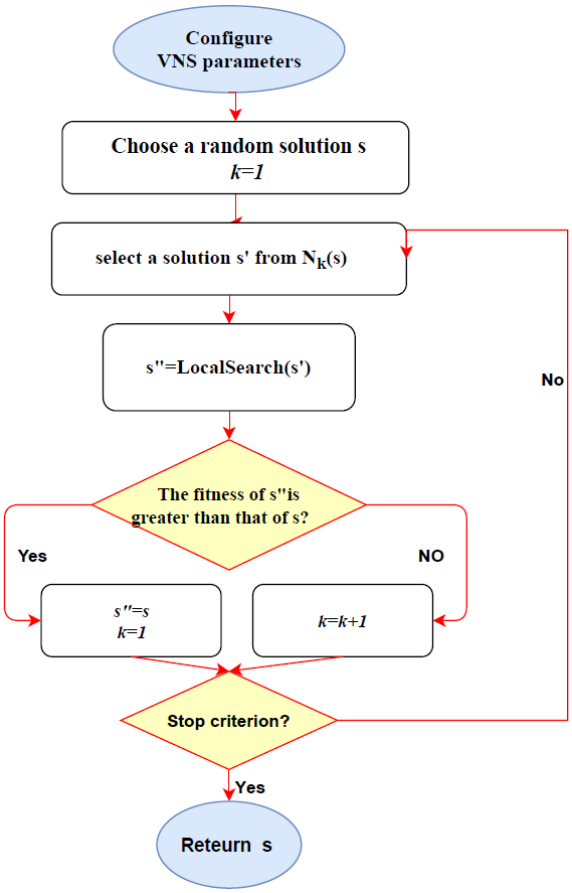
Until the stop condition is met repeat:

1 Shaking: generate a random solution s' from the neighbourhood $N_k(s)$ of the current solution s'.

2 Local search: in order to select the best solution from the neighbourhood $N_k(s)$, we apply a local search algorithm on the perturbed solution s', the resulting local optimum is denoted by s''.

3 Move or not: if s'' is better than s, move there (s ← s''), and set k = 1, otherwise, the solution s remains unchanged, and set $k \leftarrow k + 1$;

Figure 2 Flowchart of the basic VNS (see online version for colours)



3.2 VNS for gene selection

In this section, we provide a new method of selecting genes for the classification of cancer, the aim of which is to extract a highly relevant subset of genes that can improve the quality of classifiers. Given the fact that this problem is NP-hard (Amaldi and Kann, 1998), neighbourhood-based metaheuristics seem to be well suited to deal with this combinatorial problem.

- Formalisation of the problem Gene selection's major purpose is to locate a subset of relevant genes from a high-dimensional microarray dataset, this allows us to obtain a high classification of cancer, and to have a better interpretation of the biological relationship between the genes and the clinical outcome considered, then a better scientific understanding of the given problem.
- The search space: defined by the possible subsets of selected genes.

The objective function:

In this paper, the objective function J is used to measure the quality of a given subset of genes S. More precisely, the fitness function J seek to minimise the redundancy and maximise the relevance into S using a minimal number of genes. Indeed, we were inspired from the objective function defined in Ding and Peng (2005):

Minimum redundancy ensures the selection of a set of genes that contains uncorrelated genes for a better representation of the original dataset. For a subset of genes S, the minimum redundancy condition is:

$$\min W_I(S),\ W_I(S) = \frac{1}{|S|^2} \sum_{g_i, g_j \in S} NI(g_i, g_j) \tag{2}$$

On the other hand, the maximum relevance condition ensures the selection of a subset of genes with high relevance to the target class, for a given subset S we seek to:

$$\max\ V_I(S),\ V_I = \frac{1}{|S|} \sum_{g_i \in S} SNMI_i \tag{3}$$

The goal of maximising the relevance and minimising the redundancy is to optimise simultaneously the conditions equations (2) and (3), This creates a multi-objective optimisation problem, which involves merging both into a single objective function. We believe in treating both conditions equally, and the easiest way to do that is:

$$\max\left(V_I(S) - W_I(S)\right)$$

The way of combining relevance and redundancy leads to the selection criterion of a new subset of genes, called *MID*: mutual information difference criterion (Ding and Peng, 2005). Thus, our objective function is defined as follows: for a given subset of genes S, we have

$$J(S) = V_I(S) - W_I(S)$$

We can remark that the said optimisation problem takes into consideration the possible interaction between genes.

The use of evolutionary methods to solve the gene selection problem directly is inefficient due to the large dimensionality of microarray data (Bir-Jmel et al., 2019). To get beyond this obstacle, our proposed methods start with a pre-treatment stage, where we use a filter method. More precisely, at this stage, the p genes which have a high score are selected.

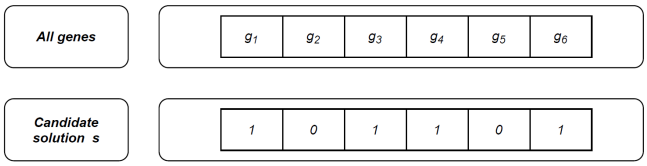
The p genes chosen in the first stage are fed into the second stage, which employs a VNS in conjunction with a local search algorithm to choose a subset of genes that contains maximum relevance and minimum redundancy:

$$\max\left(V_I - W_I\right)$$

- Representation of solutions In order to apply VNS to the problem of gene selection, one must first represent the search space in a suitable. Consider the set of p genes selected in the preprocessing step, a subset of genes S (i.e., a potential solution s) is represented by a p- binary string (i.e., $s = [s_1, s_2, \ldots, s_p]$). In s a value of '$s_i = 0$' indicates that g_i is not selected in S, whereas a bit $s_i = 1$ shows that the associated gene is selected in S.

In the Figure 3, the candidate solution s means that among 6 genes, g_1; g_3, g_4 and g_5 are selected $S = \{g_1; g_3; g_4; g_6\}$.

Figure 3 Representation of a candidate solution



- Neighbourhood structure (shaking step): The neighbourhood of a solution s corresponds to the set of solutions accessible from s using an elementary movement, thus, the Neighbourhood Structure is an essential notion in the VNS algorithm which requires to define the distance between two candidate solutions (sets of genes), in our work we use the Hamming distance.

Given two solutions according to the encoding used $s = [s_1, s_2, \ldots, s_p]$ and $s' = [s'_1, s'_2, \ldots, s'_p]$ the distance between s and s' is:

$$d(s, s') = \sum_{i=1}^{p} |s_i - s'|$$

where $d(s, s')$ is the minimum number of flip moves required to change one solution s into the other s'.

Therefore the k neighbourhood of s noted $N_k(s)$ is:

$$Nk(s) = \{s' : d(s, s') \leq k\}$$

- Neighbourhood of a solution Let s be a current solution, we fix the maximum number of neighbourhoods kmax

N_k, $k = 1, \ldots, k_{max}$ if we want to perform a perturbation (shaking step) of s to obtain a new solution s' of order m (from a given neighbourhood m, $N_m(s)$) such that $m \in \{1, \ldots, k_{max}\}$, we repeat the following process m times: we randomly choose a number $j \in \{1, 2, ...p\}$, then we exchange the associated bit in s (ie., $s_j \rightarrow |s_j - 1|$). The follow algorithm show how the perturbation of order m work.

Algorithm 1 Shake method

Function: Shake method (s, m)

Input: $s = [s_1, s_2, \ldots, s_p]$: candidate solution; m: neighbourhood order

Output: s^m perturbed solution from the neighbourhood $N_m(s)$.

Begin

$s^m = s$;

$k = 0$;

while $k < m$ **do**

$j = \lfloor rand * (p+1) \rfloor \% \lfloor . \rfloor$ is the floor function.

$s_j^m = |s_j^m - 1|$

$k = k + 1$

end while

Return s^m.

- Initialisation: The initial solution s_{init} is generated completely randomly from the p selected genes in the filter stage. sinit is considered as an initial solution to our basic VNS.
- SLS: the main role of local search methods is to improve the solution built in the shaking step and and deliver good results in a fair amount of time. For this purpose, we use a SLS inspired from Boughaci and Alkhawaldeh (2018), this method (SLS) starts with the solution generated in the shaking phase, and then executes a series of operations that combine diversification and intensification tactics to achieve good results. In fact, we define two operations:
- The diversification phase: selecting a solution from N_1 randomly.
- The intensification phase: picking the best neighbour solution in N_1.

The intensification phase has a fixed probability of $ip > 0$, while the diversification phase has a probability of $1-ip$.

As long as the initial solution is improved, the process is repeated. This is how our SLS looks.

Algorithm 2 Stochastic local search for gene selection

Function: SLS(s, ip)

Input: s = a candidate solution ; ip = the probability of doing the intensification phase

Output: s_{best} better than s.

Begin

k = 0;

while $k < 1$ **do**

if ($rand < ip$) then

s' = pick the best solution from the neighbour $N_1(s)$ (The intensification phase).

else

s' = pick a random solution from the neighbour $N_1(s)$ (The diversification phase).

end if

if $J(s') \geq J(s)$ then

$s = s'$

$k = -1\%$ continue the local search.

end if

k=k+1

end while

$s_{best} = s$

Return s_{best}.

3.3 Proposed methods for gene selection

We propose two gene selection methods: the first one is *VNSMI* that based on normalised mutual information, while the second one is VNSCor is based on the PCC. We first present the VNSMI approach, concerning the VNSCor is

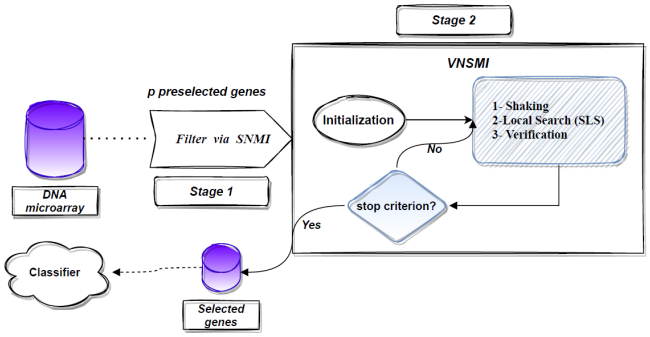
derived from the first one just we replace the normalised mutual information by the absolute correlation coefficient and SNMI score by SR coefficient.

3.3.1 Structure of the proposed approach VNSMI

Our first proposed method presented in Algorithm 3 for gene selection *VNSMI* is based on a mixture of two filter methods (one univariate and the other multivariate). The univariate scheme is based on the SNMI score by selecting the p top ranked genes, while the multivariate scheme is done through a VNS coupled with a SLS algorithm called SLS. And finally, the quality of the resultant gene subset is measured using a specific classifier through the cross-validation (LOOCV). The general procedure of this approach is characterised by a sequential process in two stages:

- Stage 1: The goal of this stage of pre-processing is to reduce the size of the initial DNA microarray data by filtering out the non-informative genes using the SNMI score; the number of pre-selected genes is set to p. This phase creates a group of p genes that are sorted in order of relevancy based on the SNMI score (univariate scheme)
- Stage 2: This step consists in applying the VNS method, in which V NSMI generates candidate subsets of genes from the genes kept in the previous step, and this is done using the VNS concepts in conjunction with the SLS algorithm. *J* is the employed objective function.

Figure 4 Flowchart of the proposed method (*V NSMI*) (see online version for colours)



Algorithm 3 Variable neighbourhood search approach for gene selection

Function: *V NSMI* (D, k_{max}, it_{max}, p, ip)

Input: X = Initial microarray data, k_{max}: The total number of neighbourhood structures, it_{max}: The maximum iterations number. p: the number of genes to select in the first stage, ip = the probability of doing the intensification phase.

Output: s_{Best}: The best subset of genes.

Stage 1: Filter

Step 1: Calculate $SNMI_i$ for each gene g_i.

Step 2: Select the p best ranked genes for use in the Stage 2.

Stage 2: The application of VNS.

Calculate the normalised mutual information between all genes.

Generate an initial solution s.

$l = 0$;

while $l < it_{max}$ **do**

  $k = 1$;

  while $k \leq k_{max}$ **do**

    $s' =$ Shake method(s, k);

    $s'' = SLS\ (s', ip)$;

    if $J\ (s'') \geq J\ (s)$ **then**

  $s = s''$;

    $k = k_{max} + 1\%$ Return to the first neighbour.

  else

    $k = k + 1$;

  end if

  end while

  $l = l + 1$;

end while

$s_{best} = s$

Return s_{best}.

- VNSCor: *VNSCor* is The second approach which is derived from *VNSMI*, while the only two differences are: the *SMI* score is replaced by the SR score, and the normalised mutual information between the genes is replaced by the absolute correlation.

4 Experiments

The implementation of the proposed approaches is performed using MATLAB R2016a. To assess the proposed techniques, we measure the accuracy (LOOCV) of the output subset of genes using the Linear SVM and 1NN classifiers.

4.1 Environment and datasets

- Datasets: The datasets utilised to evaluate our methods are all related to cancer classification problems (DNA microarray), include binary classifications, and have been used in several state-of-the-art papers Table 1. And since our approaches are designed to deal with high-dimensional problems, then we have datasets characterised by a high number of genes (ranging between 2,000 and 15,154 genes) and a small number of samples (varying between 62 and 253 samples).
- Preparation of datasets: The genes in the original data are continuous and had different scales. Therefore, we classified them in the first time using SVM, 1NN. So, the data is then scaled to give each gene a mean value of zero and unit variance. This values is used in the *VNSCor* technique. On the contrary, in order to

approximate the mutual information values between genes in *V NSMI*, the data is discretised using the following method: for each gene in the data we calculate the respective σ(standard deviation) and μ(mean), then the data of the gene vector which has a value greater than $\mu + \sigma/2$ have been transformed to 1; values between $\mu - \sigma/2$ and $\mu + \sigma/2$ have been transformed into 0; and all values less than $\mu - \sigma/2$ has been transformed into −1.

Table 1 Characteristics of the datasets used

Dataset	*# genes*	*# samples*	*# classes*
Colon (Alon et al., 1999)	2,000	62	2
DLBCL (Statnikov et al., 2004)	5469	77	2
Leukemia (Golub et al., 1999)	7,129	72	2
Ovarian (Petricoin et al., 2002)	15,154	253	2
Prostate Tumor Statnikov et al., 2004)	10,509	102	2
Prostate (Singh et al., 2002)	12,600	102	2

4.2 Parameters

In order to evaluate the proposed methods, the impemetations have been performed on Acer laptop with an Intel Core I5 2.30GHz processor and 8GB RAM.

Several experiments were conducted in order to acquire a suitable parameterisation. For that, we fixed an initial values for the parameters, then we select one parameter, and at each run we change it until we get the best result. The procedure of adjusting each parameter is continued until the solutions could no longer be improved. The parameters of the proposed approaches are listed in Table 2.

Table 2 Parameters used for experiments.

Parameter	*Value*
The total number of chosen genes in the stage 1 (Filter) p	100
The total number of neighbourhood structures (VNS) k_{max}	15
The maximum iterations number (VNS) it_{max}	200
The probability of the intensification in SLS ip	0, 5

5 Results and comparisons

In order to evaluate the proposed methods, six datasets of (DNA microarray) have been used. Due to the nondeterministic nature of our methods, each dataset is subjected to 50 independent runs in order to generate more credible results. Table 3 illustrates the outcomes of gene selection using *V NSMI* and subsequently utilising the *V NSCor* approach. In all these methods, the classification accuracy is calculated using the 1NN and SVM (linear) classifier. To show the utility of this selection we have compared them to the SVM, 1NN classifiers without selection. The obtained results have been analysed based on the accuracy, number of genes employed, and execution time of our proposed approaches.

From the Table 3, it is clear that VNSCor method combined with the SVM classifier shows the best performance among all methods on all datasets. On the other hand, the average of the selected genes using VNSMI is inferior to that using VNSCor for the majority of datasets except Ovarian and Colon datasets. The effectiveness of the presented methods is derived from the significant increase in classification accuracy (Table 3 Figures 5 and 6). Indeed, The VNSMI and VNSCor approaches minimise the number of genes used in all datasets while improving their accuracy compared to the use of the all subset of genes. In addition, our methods achieve perfect accuracy for 3 datasets (DLBCL, Leukimia and Ovarian) with less than 13 genes.

Figure 5 Comparison of the average accuracy using (1NN) (see online version for colours)

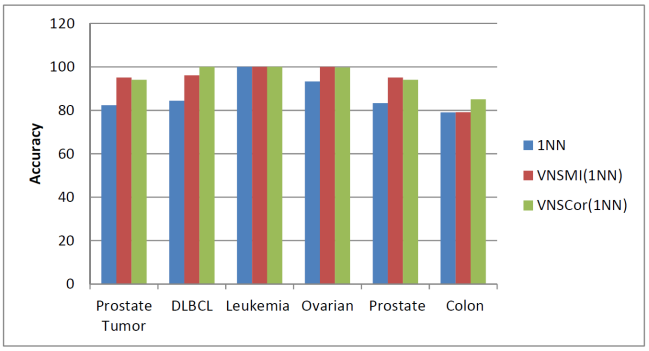
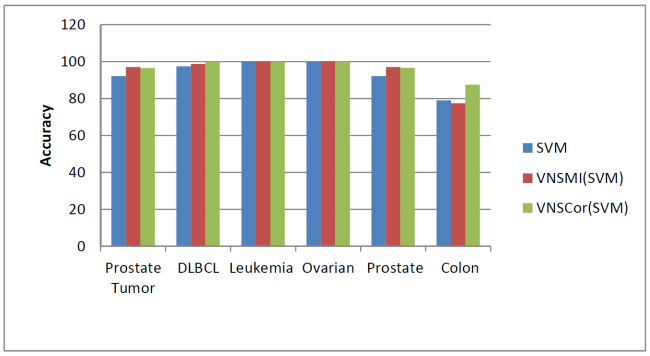


Figure 6 Comparison of the average classification accuracy using (SVM) (see online version for colours)



The obtained results show that VNSMI and VNSCor perform similarly for the majority of datasets, while VNSCor having a little accuracy edge. In addition, the combination between our proposed methods and the SVM classifier surpasses the 1NN classifier. For all datasets, the VNSMI and VNSCor approaches achieved high accuracies using SVM and 1NN classifiers. While a classification of more than 95.1% using only less than 20 genes for five datasets.

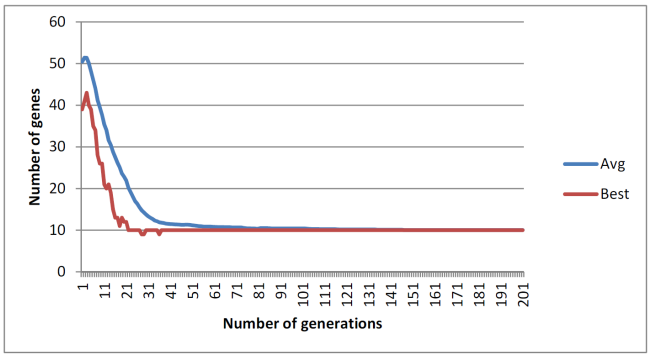
Table 3 Comparison of SVM, 1NN, VNSMI and VNSCor (LOOCV)

Datasets	*Performances*	*1NN*	*SVM*	*VNSMI(1NN)*		*VNSMI(SVM)*		*VNSCor(1NN)*		*VNSCor(SVM)*	
				best	*Avg.*	*best*	*Avg.*	*best*	*Avg.*	*best*	*Avg.*
Prostate tumour	Accuracy	82.35	92.16	95.1	95.1	97.06	97.06	95.1	94.04	98.04	96.53
	Genes	10,509	10,509	9	9	9	9	13	14.38	13	14.38
	Time of execution(s)	-	-		4.89				5.01		
DLBCL	Accuracy	84.42	97.4	96.1	96.1	98.7	98.7	100	100	100	99.97
	Genes	5,469	5,469	10	10	10	10	13	16.64	13	16.64
	Time of execution	-	-		4.27	-	-	-	4.4		
Leukemia	Accuracy	100	100	100	100	100	100	100	100	100	100
	Genes	7,129	7,129	11	11	11	11	12	15.44	12	15.44
	Time of execution(s)	-	-		4.41				4.48		
Ovarian	Accuracy	93.28	100	100	100	100	100	100	99.76	100	100
	Genes	15,154	15,154	20	20.94	20	20.94	8	12.54	8	12.54
	time of execution(s)	-	-	-	5.49				5.63		
Prostate	Accuracy	83.33	92.16	95.1	95.1	97.06	97.06	95.1	94.04	98.04	96.63
	Genes	12,600	12,600	9	9	9	9	13	14.56	13	14.56
	Time of execution(s)				4.82				4.94		
Colon	Accuracy	79.03	79.03	82.26	79.1	79.03	77.45	88.71	85.03	88.71	87.52
	Genes	2,000	2,000	13	10.06	13	10.06	7	8,06	7	8.06
	Time of execution(s)				4.41				4,52		

Notes: Remark: the classification in SVM, 1NN is done in a single run. Accuracy: (leave one out cross validation 'LOOCV'). Genes: The number of genes used. Avg.: The average of the 50 runs. SVM: The support vector machine classifier using a linear kernel. 1NN: The 1−nearest neighbour classifier. Time of execution: The average of execution time in seconds.

We can find also from the Table 3 and Figures 5 and 6 that our gene selection methods are not specific just for one classifier, but also to wrapper approaches. Another point that makes our methods effective is the time of execution, while we can achieve a reasonable classification in a few seconds.

Figure 7 Comparison of the number of genes used over generations for 'Colon' using *VNSMI* (see online version for colours)



Figures 7, 8, 9, and 10 show the effectiveness of the proposed approaches in minimising the number of genes, and in maximising the fitness function based on 'Colon' dataset. We can notice that the results obtained by the 50 runs are quite close, while the differences between the best and average solutions are a bit similar. The figures also demonstrate the importance of our multivariate filter approaches VNS-based in reducing the number of genes while maintaining high accuracy rates, this shows how the number of genes is combined in the objective function *J*.

Explained cancer prediction models are developed by integrating gene selection and 1NN and SVM classifiers. With the gene selection strategy developed, we seek a set of genes whose relevance is maximum and redundancy is minimum using the VNS methaheureusitic, a SLS algorithm and a well-chosen objective function. The performances of our approaches are measured in terms of accuracy, number of selected genes and running time. The developed methods had high accuracy and the number of selected genes are less compared to the initial number. Form the presented results we can see that our gene selection approaches are well-founded based on the tests we conducted. Indeed, our approaches achieved good classification accuracy in the six

datasets studied. And since gene selection is a special case of feature selection, then the proposed approaches can be used for other tasks, namely: face recognition (Yang et al., 2007), industry (Fernandes et al., 2019; Dai et al., 2015) and others (Jovic et al., 2015).

Figure 8 Comparison of the number of genes used over generations for 'Colon' using *VNSCor* (see online version for colours)

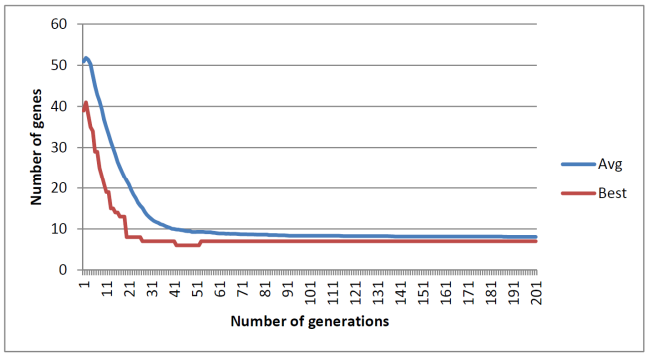


Figure 9 Comparison of the objective function's evolution J for 'Colon' using *VNSMI* (see online version for colours)

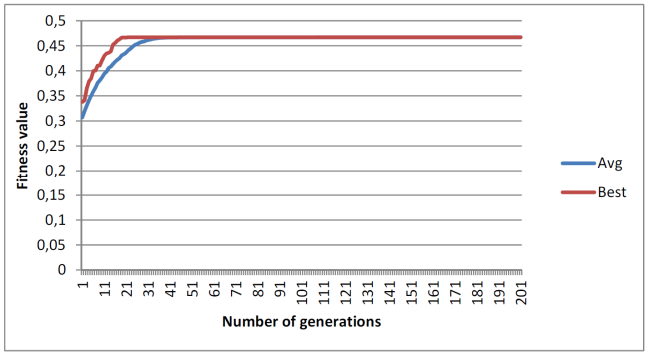
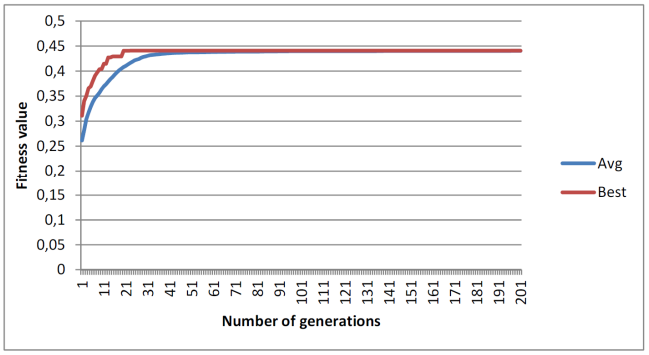


Figure 10 Comparison of the objective function's evolution *J* for 'Colon' using *VNSCor* (see online version for colours)



6 Conclusions

In this work, two hybrid multivariate filter approaches for the gene selection in DNA microarray data are presented. The proposed approaches include a pre-selection phase that is accomplished using a univariate scheme filter. The univariate scheme is used to select a preliminary small subset of genes, then a VNS metaheuristic coupled with a SLS algorithm is used to get the best subset of genes. The proposed methods are designed to select a small subset of relevant, non-redundant genes from the original dataset. The experimental results reveal that the proposed methods are quite effective in terms of classification accuracy and number of genes selected.

Acknowledgements

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the National Centre for Scientific and Technical Research (CNRST) under Project ALKHAWARIZMI/2020/19.

References

Almutiri, T. and Saeed, F. (2022) 'A hybrid feature selection method combining Gini index and support vector machine with recursive feature elimination for gene expression classification', *International Journal of Data Mining, Modelling and Management*, Vol. 14, No. 1, pp.41–62.

Alomari, O.A., Khader, A.T., Al-Betar, M.A. and Abualigah, L.M. (2017) 'Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm', *International Journal of Data Mining and Bioinformatics*, Vol. 19, No. 1, pp.32–51.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proceedings of the National Academy of Sciences*, Vol. 96, No. 12, pp.6745–6750.

Amaldi, E. and Kann, V. (1998) 'On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems', *Theoretical Computer Science*, Vol. 209, Nos. 1–2, pp.237–260.

Bir-Jmel, A., Douiri, S.M. and Elbernoussi, S. (2019) 'Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data', *Computational and Mathematical Methods in Medicine*, Vol. 2019, Article ID 7828590, 20pp., https://doi.org/10.1155/2019/7828590.

Bir-Jmel, A., Douiri, S.M. and Elbernoussi, S. (2019) 'Gene selection via BPSO and Backward generation for cancer classification', *RAIRO-Operations Research*, Vol. 53, No. 1, pp.269–288.

Bonev, B., Escolano, F. and Cazorla, M. (2008) 'Feature selection, mutual information, and the classification of highdimensional patterns', *Pattern Analysis and Applications*, Vol. 11, Nos. 3–4, pp.309–319.

Boughaci, D. and Alkhawaldeh, A.A.S. (2018) 'Three local search-based methods for feature selection in credit scoring', *Vietnam Journal of Computer Science*, Vol. 5, No. 2, pp.107–121.

Chiang, Y.M., Chiang, H.M. and Lin, S.Y. (2008) 'The application of ant colony optimization for gene selection in microarray-based cancer classification', In *Machine Learning and Cybernetics, 2008 International Conference on*, July, Vol. 7, pp.4001–4006, IEEE.

Chuang, L.Y., Chang, H.W., Tu, C.J. and Yang, C.H. (2008) 'Improved binary PSO for feature selection using gene expression dat', *Computational Biology and Chemistry*, Vol. 32, No. 1, pp.29–38.

Dabba, A., Tari, A., Meftali, S. and Mokhtari, R. (2021) 'Gene selection and classification of microarray data method based on mutual information and moth flame algorithm', *Expert Systems with Applications*, Vol. 166, p.114012, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.114012.

Dai, Q., Cheng, J.H., Sun, D.W. and Zeng, X.A. (2015) 'Advances in feature selection methods for hyperspectral image processing in food industry applications: a review', *Critical Reviews in Food Science and Nutrition*, Vol. 55, No. 10, pp.1368–1382.

Ding, C. and Peng, H. (2005) 'Minimum redundancy feature selection from microarray gene expression data', *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, pp.185–205.

Douiri, S.M. and El Bernoussi, S. (2012) 'An ant algorithm for the sum coloring problem', *International Journal of Applied Mathematics and Statistics*, Vol. 27, No. 3, pp.102–110.

Estévez, P.A., Tesmer, M., Perez, C.A. and Zurada, J.M. (2009) 'Normalized mutual information feature selection', in *IEEE Transactions on Neural Networks*, February, Vol. 20, No. 2, pp.189–201, doi: 10.1109/TNN.2008.2005601.

Fernandes, M., Canito, A., Bolón-Canedo, V., Conceição, L., Praça, I. and Marreiros, G. (2019) 'Data analysis and feature selection for predictive maintenance: a case-study in the metallurgic industry', *International Journal of Information Management*, Vol. 46, pp.252–262, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.10.006.

García-Torres, V., García-López, F.C., Melián-Batista, B., Moreno-Pérez, J.A. and Moreno-Vega, J.M. (2004) 'Solving feature subset selection problem by a hybrid metaheuristic', in *First International Workshop in Hybrid Metaheuristics at ECAI 2004 (HM 2004)*, pp.59–69.

García-Torres, M., Gómez-Vela, F., Melián-Batista, B. and Moreno-Vega, J.M. (2016) 'High-dimensional feature selection via feature grouping: a variable neighborhood search approach', *Information Sciences*, Vol. 326, pp.102–118, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2015.07.041.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., ... and Bloomfield, C.D. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, No. 5439, pp.531–537, DOI: 10.1126/science.286.5439.531.

Gu, Q., Li, Z. and Han, J. (2012) 'Generalized fisher score for feature selection', arXiv preprint arXiv:1202.3725.

Hansen, P. and Mladenovic, N. (2001) 'Variable neighborhood search: Principles and applications', *European Journal of Operational Research*, Vol. 130, No. 3, pp.449–467.

Helder, V.G., Filomena, T.P., Ferreira, L. and Kirch, G. (2022) 'Application of the VNS heuristic for feature selection in credit scoring problems', *Machine Learning with Applications*, Vol. 9, p.100349, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100349.

Jafari, P. and Azuaje, F. (2006) 'An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors', *BMC Medical Informatics and Decision Making*, Vol. 6, No. 1, p.27.

Jovic, A., Brkic, K. and Bogunovic, N. (2015) 'A review of feature selection methods with applications', In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May pp.1200–1205, IEEE.

Kohavi, R. and John, G.H. (1997) 'Wrappers for feature subset selection', *Artificial Intelligence*, Vol. 97, Nos. 1–2, pp.273–324.

Kumari, B. and Swarnkar, T. (2011) 'Filter versus wrapper feature subset selection in large dimensionality micro array: a review', *International Journal of Computer Science and Information Technologies*, Vol. 2, No. 3, pp.1048–1053.

Lai, C-M. and Huang, H-P. (2021) 'A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique', *Applied Soft Computing*, Vol. 100, p.106994, ISSN 1568–4946, https://doi.org/10.1016/j.asoc.2020.106994.

Lai, C-M., Yeh, W-C. and Chang C-Y. (2016) 'Gene selection using information gain and improved simplified swarm optimization', *Neurocomputing,* Vol. 19, No. 218, pp.331–338.

Lee, C.P. and Leu, Y. (2011) 'A novel hybrid feature selection method for microarray data analysis', *Applied Soft Computing*, Vol. 11, No. 1, pp.208–213.

Li, S., Wu, X. and Tan, M. (2008) 'Gene selection using hybrid particle swarm optimization and genetic algorithm', *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, Vol. 12, No. 11, pp.1039–1048.

Li, W. and Yang, Y. (2002) 'How many genes are needed for a discriminant microarray data analysis', in *Methods of Microarray Data Analysis,* pp.137–149, Springer, Boston, MA.

Li, Y., Wang, G., Chen, H., Shi, L. and Qin, L. (2013) 'An ant colony optimization based dimension reduction method for high-dimensional datasets', *Journal of Bionic Engineering*, Vol. 10, No. 2, pp.231–241.

Ma, Y. and Zhu, L. (2013) 'A review on dimension reduction', *International Statistical Review*, Vol. 81, No. 1, pp.134–150.

Maafiri, A., Bir-Jmel, A., Elharrouss, O., Khelifi, F. and Chougdali, K. (2022) 'LWKPCA: a new robust method for face recognition under adverse conditions', in *IEEE Access*, Vol. 10, pp.64819–64831, doi: 10.1109/ACCESS.2022.3184616.

Maafiri, A., Elharrouss, O., Rfifi, S., Al-Maadeed, S.A. and Chougdali, K. (2021) 'DeepWTPCA-L1: a new deep face recognition model based on WTPCA-L1 norm features', in *IEEE Access*, Vol. 9, pp.65091–65100, doi: 10.1109/ACCESS.2021.3076359.

Matijevic, L. (2022) 'Variable neighborhood search for multi-label feature selection', in *International Conference on Mathematical Optimization Theory and Operations Research*, pp.94–107, Springer, Cham.

Mishra, D. and Sahu, B. (2011) 'Feature selection for cancer classification: a signal-to-noise ratio approach', *International Journal of Scientific and Engineering Research*, Vol. 2, No. 4, pp.1–7.

Mladenovic, N. and Hansen, P. (1997) 'Variable neighborhood search', *Computers and Operations Research*, Vol. 24, No. 11, pp.1097–1100.

Pacheco, J., Casado, S. and Nuñez, L. (2007) 'Use of VNS and TS in classification: variable selection and determination of the linear discrimination function coefficients', *IMA Journal of Management Mathematics*, Vol. 18, No. 2, pp.191–206.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) 'Light-generated oligonucleotide arrays for rapid DNA sequence analysis', *Proc. Natl. Acad. Sci.*, USA, Vol. 91, pp.5022–5026.

Peng, H. (2022) *Mutual Infomation MATLAB Toolbox* [online] http://www.mathworks.com/matlabcentral/fileexchange/14888-mutualinformation-computation (accessed 15 October 2022).

Peng, Y., Wu, Z. and Jiang, J. (2010) 'A novel feature selection approach for biomedical data classification', *Journal of Biomedical Informatics*, Vol. 43, No. 1, pp.15–23.

Pentelas, A., Sifaleras, A. and Koloniari, G. (2019) 'A reduced variable neighborhood search approach for feature selection in cancer classification', in *International Conference on Variable Neighborhood Search*, October, pp.1–16, Springer, Cham.

Petricoin III, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., ... and Liotta, L.A. (2002) 'Use of proteomic patterns in serum to identify ovarian cancer', *The Lancet*, Vol. 359, No. 9306, pp.572–577.

Rostami, M., Forouzandeh, S., Berahmand, K., Soltani, M., Shahsavari, M. and Oussalah, M. (2022) 'Gene selection for microarray data classification via multi-objective graph theoretic-based method', *Artificial Intelligence in Medicine*, Vol. 123, p.102228, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2021.102228.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science*, Vol. 270, No. 5235, pp.467–470.

Sharbaf, F.V., Mosafer, S. and Moattar, M.H. (2016) 'A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization', *Genomics*, Vol. 107, No. 6, pp.231–238.

Shreem, S.S., Abdullah, S., Nazri, M.Z.A. and Alzaqebah, M. (2012) 'Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection', *J. Theor. Appl. Inf. Technol*., Vol. 46, No. 2, pp.1034–1039.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., ... and Lander, E.S. (2002) 'Gene expression correlates of clinical prostate cancer behavior', *Cancer cell*, Vol. 1, No. 2, pp.203–209.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D. and Levy, S. (2004) 'A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis', *Bioinformatics*, Vol. 21, No. 5, pp.631–643.

Tabakhi, S. and Moradi, P. (2015) 'Relevance–redundancy feature selection based on ant colony optimization', *Pattern Recognition*, Vol. 48, No. 9, pp.2798–2811.

Tang, J., Alelyani, S. and Liu, H. (2014) 'Feature selection for classification: a review', *Data Classification: Algorithms and Applications*, Vol. 37 [online] https://www.cvs.edu.in/upload/feature_selection_for_classification.pdf.

Xiong, M., Fang, X. and Zhao, J. (2001) 'Biomarker identification by feature wrappers', *Genome Research*, Vol. 11, No. 11, pp.1878–1887.

Yang, A.Y., Wright, J., Ma, Y. and Sastry, S.S. (2007) *Feature Selection in Face Recognition: A Sparse Representation Perspective*, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2007-99.

Yu, L. and Liu, H. (2003) 'Feature selection for high-dimensional data: A fast correlation-based filter solution', in *Proceedings of the 20th International Conference on Machine learning (ICML-03)*, pp.856–863.

Zhang, J., Feng, J., Yang, X. and Liu, J. (2021) 'Gene selection and classification combining information gain ratio with fruit fly optimisation algorithm for single-cell RNA-seq data', *International Journal of Computational Science and Engineering*, Vol. 24, No. 5, pp.495–504.