EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

# NNFSRR: Nearest Neighbor Feature Selection and Redundancy Removal Method for Nearest Neighbor Search in Microarray Gene Expression Data

Rupali Bhartiya[1, *], Gend Lal Prajapati [2]

[1] Department of Computer Science & Engineering, Shri Vaishav Institute of Information Technology, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
[2]Department of Computer Engineering, Institute of Engineering & Technology, Devi Ahilya University, Indore-452001, India

## Abstract

INTRODUCTION: Gene expression data analysis is a critical aspect of disease prediction and classification, playing a pivotal role in the field of bioinformatics and biomedical research. High-dimensional gene expression datasets hold a wealth of information, but their effective utilization is hindered by the presence of irrelevant dimensions and noise. The challenge lies in extracting meaningful features from these datasets to enhance the accuracy of disease prediction and classification while maintaining computational efficiency.
Feature selection is a crucial step in addressing these challenges, as it aims to identify and retain only the most informative characteristics from large high-dimensional microarray datasets. In the context of microarray gene expression data, characterized by its substantial dimensionality, selecting relevant features is essential for efficient nearest neighbor search, a fundamental component of various analytical tasks in bioinformatics and data mining.
Existing feature selection methods in high-dimensional data often face issues related to the trade-off between search accuracy and computational efficiency. This paper introduces a novel approach, the Nearest Neighbor Feature Selection with Symmetrical Uncertainty-based Redundancy Removal (NNFSRR) method, designed to enhance the classification of microarray gene expression data through feature selection. The NNFSRR method focuses on reducing the dimensionality of the dataset by identifying and removing redundant features, allowing subsequent searches to operate solely on relevant dimensions.
OBJECTIVES: The primary goal is to evaluate the NNFSRR method's effectiveness in improving nearest neighbor search in microarray gene expression datasets by reducing dimensionality. This method utilizes Symmetrical Uncertainty-based correlation between dimensions for feature selection and aims to enhance accuracy and efficiency compared to existing methods.
METHODS: The NNFSRR method uses Symmetrical Uncertainty to identify and remove redundant features from microarray gene expression datasets. Reduced datasets are used for nearest neighbor search, improving accuracy and efficiency. Experiments are conducted using real-world datasets, and comparisons with existing methods are made based on search time and accuracy.
RESULTS: The NNFSRR method demonstrates improved nearest neighbor search performance, outperforming basic brute force methods and existing feature selection techniques. Selected feature sets exhibit strong class associations while minimizing feature correlations, enhancing classification precision.
CONCLUSION: In conclusion, the NNFSRR method presents a promising approach to address the challenges posed by high-dimensional gene expression data. It effectively reduces dimensionality, improves search accuracy, and enhances the efficiency of nearest neighbor search. Our experimental results demonstrate that this method outperforms existing techniques in terms of search time and accuracy, making it a valuable tool for applications in bioinformatics, data mining, pattern recognition, and biological information retrieval. The NNFSRR method holds the potential to advance our understanding of complex biological processes and support more accurate disease prediction and classification.

**Keywords:** High dimension**,** Feature, gene expressions, Symmetric Uncertainty

Received on 29 May 2023, accepted on 03 September 2023, published on 19 September 2023

doi: 10.4108/eetpht.9.3910

*Corresponding author. Email: rupalibhartiya@gmail.com

## 1. Introduction

The application of gene expression data for data analytics has become very popular recently. It is also used for various bioinformatics and detection of various types of diseases, drug generation etc. [11]. In this paper we consider feature selection for the purpose of data classification in high dimensions. An example of classification task on high dimensional data like gene expression data where feature selection plays an important role is the diagnosis of tumour associated cells [1, 2]. It can also be used for diagnosis of expressions of genes and are categorized into different classes so that it will improve the underlying biological process understanding [18].

Objective of gene expression classification is to make a suitable model that find the genes and used for the class prediction of other unidentified samples. Problem with the finding of efficient gene classification techniques is very large number of dimensions or features as compared to the number of records and the change in expression level by different available techniques.

Genetic expression data is high dimensional in nature. It usually consists of very high number of dimensions and few numbers of records. Processing on this high dimensional data require much effort in other applications also like pattern mining, Geographical Information System (GIS), web mining, DNA matching, copyright violation detection, bioinformatics and many more [2]. Many of the algorithms available for searching in high dimensions are not suitable for bioinformatics applications. In gene expression data, level of expression of normal people and patients with some diseases are different. For example, it is possible to diagnose cancer patients by classifying expression level in different groups. As number of dimensions in gene expression data is very high and generally it consists of noise, missing data, mislabelled data, redundant data, cross platform comparison and biasing issues. In addition to that, it possibly consists of many irrelevant dimensions that are not related with the disease diagnosis, but complete data set is used for diagnosis [3, 4]. Missing and noisy data generally removed during pre-processing but huge number of dimensions describing the data can be reduced by feature selection and reduction techniques. Objective of feature selection in high dimensions is to reduce the computation time of algorithm and accuracy of search result. In high dimensional data, dimension selection technique is generally applied during pre-processing. Many feature selection methods are available but problem of feature selection in high dimensions is still a big issue for automatic detection of specified search results [3]. This immensity creates “curse of dimensionality” in high dimensional applications in terms of scalability and performance [18].

For efficient classification of high dimensional data like gene expression data, feature selection based on feature ranking is used. This selection is a function returns a relevance index $I(D|D',C)$ for data mining task C (like classification or clustering) on the data D by finding how much relativity of selected feature set D’ is relevant for the task C. In this process of feature selection, data D and mining task C remains fixed except subset D’ which varies as per the function J(D’). Computation of relevance index for each feature *Di* ($i$=1, 2,....$N$) generate ranking of features $J(D_{i1}) \leq J(D_{i2}) \ldots\ldots \leq J(D_{iN})$. Features having low index may be eliminated. If features are not independent, then there may be redundancy in most of the features especially in case of high dimensional data. Hence it is not sufficient to eliminate features on the basis of indexing only. Method which finds the best subset of features may use redundant feature filter for competent feature subset.

In gene expression data, row represents the genes and column represents various sample values like tissues and experimental conditions. Dimensions or features are gene expressions and combination of these gene expressions represents presence or absence of any disease [5, 6]. Values of gene dimensions are represented in numbers which represents the expression level of an individual gene in any specific gene case. In feature selection, dimensions selected among the complete set of dimensions makes combination for objective of making a small subset of dimensions that represent the problem with minimum or no degradation and may generate better results as compared to solution with complete dimension set. For example, in a high dimensional dataset, there is large number of features and most of them are irrelevant and redundant for the application. In this condition, feature selection is very important to handle this problem in task with high dimensional data. Feature selection has several advantages like improved storage overhead, data understanding, data visualization, cost effectiveness and reduced feature sets. All these advantages of feature selection lead to increase in speed, improved performance of the algorithm and possibility of simpler models for use.

Feature selection is categorized into two classes namely filter and wrapper [19]. Wrapper methods, as depicted in Fig 1, select dimensions based on some algorithm for classifying the objects and then evaluate its classification performance, whereas filter methods, as depicted in Fig 2, work on the classification criteria based on the information content in the features. Among these two classes filter method is generally used for handling high dimensional data as these methods are based on weight and subset of features to find the best feature individually or in a subset. Weight of these methods is calculated by information content of features like distance between classes, entropy, information gain or other statistical values. Wrapper methods are complex as these methods tries to optimize the classification learning

algorithm which increases the complexity of algorithm as there is need to train data with large number of dimensions. This issue also tried to solve using various heuristic algorithms like forward and backward selection but there is no guarantee of efficient optimization because of its heuristic nature. There are various feature selection methods available for low dimensional data and very few for high dimensions. Each method will generate different result set of selected dimensions from same dataset of complete dimensions and perform differently for different types of applications.

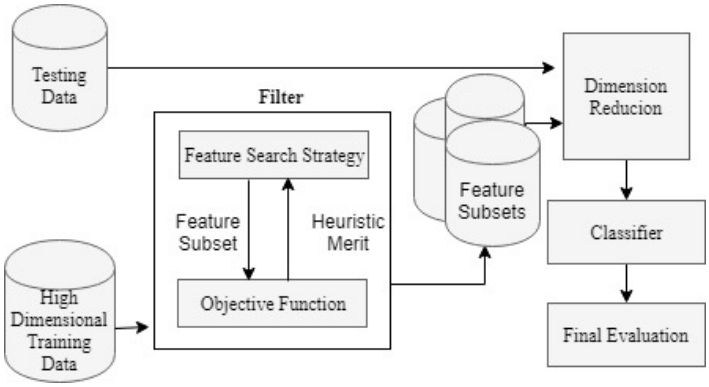


**Figure 1.**: Filter Method for Feature Selection

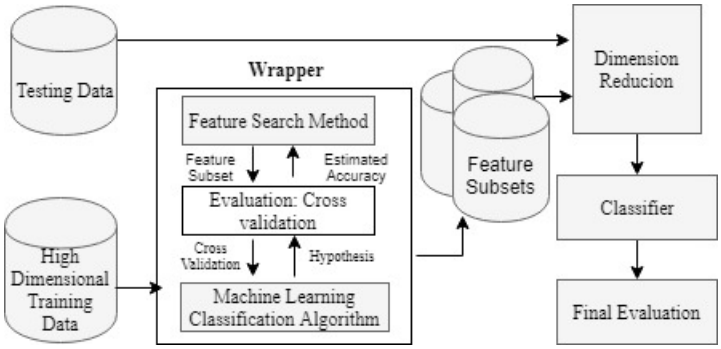


**Figure 2:** Wrapper Method of Feature Selection

In case of gene expression data, feature selection plays important role in finding the nearest matching pattern from the existing query [9, 13]. If efficient feature selection technique will not be applied on dataset, then it will definitely give inefficient result. For this challenging task of finding nearest neighbour in high dimensional data, if more than one feature selection technique will be combined then result would be enhanced. There are algorithm exists which works in direction of hybrid dimension selection technique in high dimensional data but finding nearest neighbor in high dimensions using feature selection is not much explored.

Dimension selection is still one of the major challenges in high dimensions for statistical machine learning and used in many applications where nearest neighbour search results are required [9, 12]. This paper is concerned about various issues related to existing methods, building a prediction model using dimension selection for efficiently finding the nearest neighbors in high dimensional data, implementation

of the feature selection algorithm, computational complexity of the proposed method and efficiency comparison with the existing solutions. Algorithm is based on the concept of first dividing the dimensions of gene expression data into different subsets i.e. groups, then apply proposed dimension selection algorithm on different subsets of data based on finding the of the feature in the cluster. Correlation between features is found at cluster level and then at global level to remove redundancy of clusters. Here we use Pearson's coefficient function to find correlation between features.

In the following, first we review the methods of feature selection in high dimensions, and then propose a novel approach of nearest neighbour search for selecting features by grouping result of feature selection based on given query and the correlation between the features. The The proposed feature selection algorithm will address the issues with the existing feature selection methods used for finding nearest matching data in high dimensional data of gene expressions and their classification. Proposed algorithm is based on the concept of feature selection based on correlation between features by computing Symmetric Uncertainty (SU) between features and then removing redundant features by comparing each feature with other features in the group and class. We present a framework for our feature subset selection algorithms, which can help with classification and clustering. This requires some aspect, e.g.,

- To choose significant feature for classification or clustering precision.
- To find that which traits are relevant for a specific class and which ones aren't.
- To decide that attribute is redundant out of all the ones that are relevant or not.
- To find out that two attributes are strongly related or not.

For the purpose of developing an algorithm or model, a fitness function known as SU can be utilised. The presented methodology can be used to identify the most important characteristics for Data Mining tasks. It is based on high-dimensional supervised datasets for attribute correlation analysis. We can find the answers to these questions when that happens. Efficiency of algorithm using feature selection is compared with the existing feature selection methods used for high dimensional data. Classification result of high dimensional gene expression data with full features is compared with classification result with proposed feature selection.

## 2. Literature Review

All available dimension selection techniques have goal of searching for a best subset of dimensions. Some techniques use a criterion function to decide whether any dimension should be selected or not. Some techniques try all combination of dimensions and then decide the criterion function which gives the highest value of function. This exhaustive search method is useful for less dimensionality since it is time consuming in high dimensions. Function is computed for every set of dimension subsets. Then number of computations to find the best set of dimensions will be $d/k$ where $d$ is number of dimensions and $k$ is required number of dimensions. This brute force method of selecting best dimension subset will increase time complexity for higher number of dimensions and make it impractical to use for high dimensions.

Branch and bound methods generate tree structure where root of the tree consists of all dimensions and their children consists of all combination with one dimension removed. From each of the child node, next child nodes are generated where another dimension has been removed and so on. If $f$ number of dimensions is required to retain among $d$ number of dimensions, then height of the tree will be $d$-$f$. Leaf nodes consist of combination of selected dimensions where its criterion function is computed. This is stored as bound and use for future branch evaluation and branch will not be generated next if the criterion function value is below the bound value. For another leaf node if the value of criterion function is larger than bound, bound value will be updated, and combination of dimension subset is best so far. Due to its monotonicity property, branch and bound based dimension selection is not practical for higher dimensions.

In Relaxed Branch and Bound method, monotonicity principle is used but it is relaxed. This method is also complex as other criteria for margin is used here, so this is also not efficient for high dimensional data. Method of selecting best individual features is simple and comparatively fast as there is no need to evaluate different combinations of dimension subsets. But individually selected subset of dimensions does not give efficient result as they are incapable of removing redundant dimensions because redundant dimensions are likely to have similar ranking, i.e. these methods does not consider dimension interdependency.

Basically, dimension selection strategy is based on two methods: one is statistical and other is information based. Many dimension selection techniques were tested for high dimensional data like Information Gain, Information Ratio, towing rule, sum minority, max minority, gain index, sum of variance, SVM, t-statistics, Pearson-Correlation and correlation measurement, Euclidean distance, distributed feature selection. Some techniques use embedded methods feature selection with SVM, kernel mutation information, Fuzzy Correlation, PLS-based local recursive feature elimination. Information based methods are based on redundancy and relevance characteristics of each dimension [16].

Distance measure in full dimensions can be unreliable and finding such subsets may be useless as dimensions may consist of noise or irrelevant dimensions. For this purpose of

selecting key set of dimensions that best represent the class or eliminating irrelevant dimensions that are not used for classifying a particular cluster or subset. Hybrid techniques use these information as well as statistical measurement i.e. which combine filtration and evaluation of best selection of dimensions were also applied on genetic data for analysing patterns. As gene expression data consists of many thousands of dimensions, therefore it is a long-term process. Feature selection on gene expression data generally works and give better performance and approximately similar accuracy as all dimensions do not participate in exactly defining a class. Performance of feature selection on microarray data is evaluated by estimating the error in classification [6, 10, 14, and 15]. Effective Range Based Selection (ERBS) [23] is one of the methods in which dimensions are selected based on their weight and those that separate the class.

For general application on high dimensional data, no method is superior to the other however its effectiveness is judged on factors as well as data population, size of dimension, data type and type of problem. The dimension selection technique is advantageous in high dimensional micro-array dataset like modelling of a better classifier and identification of irrelevant genes.

In case of high dimensionality, number of training dataset should be large of data for getting higher accuracy of classification. By reducing the dimensionality, time of training and accuracy of classification will be increased. But this is possible if dimensions removed are not discriminatory features. This ensures that the classification accuracy does not decrease. To some extent given number of dimensions for training data will leads to increase in classification accuracy, but after that accuracy will be decreased [5, 6, 7, 8].

## 3. Dimension Selection

Selecting the properties or features that clearly identify a problem, rather than those that are irrelevant or superfluous is the hallmark of dimension selection. Many aspects of dimension selection have been examined by researchers. To find the best subset of a dimension, one characteristic is to evaluate how effective each dimension subset is. Filters, wrappers, and the embedded technique, which is a combination of the filter and wrapper methods, are the standard classifications for the various feature selection approaches.

High-dimensional data handling necessitates dimension selection in order to save training time and memory usage while also ensuring the accuracy of the results. For high-dimensional data, filtering methods are simple and fast because they don't rely on an algorithm for selection. Chi-square, Information Gain (IG), correlation measurement, distance dependency, information and consistency-based filtration are some of the metrics used for filtration in filter methods. Unsupervised (Information Gain, t-test feature selection, Correlation-based feature selection, etc.) and supervised (Bayesian network) filter methods are also used to categorise filter algorithms. Correlation Based Filter (FCBF), Minimum Redundancy Maximum Relevance (MRMR), MD filter and more filter methods are available based on these measurements. As a result of their separation from the classifier, filter methods have an advantage over wrapper methods in many situations. By engaging with the classifier, wrapper techniques employ selection algorithms that take into account dependencies across dimensions. Uses algorithms as an assessment function, and tests the model developed by the learning algorithm before generating a rank based on their effectiveness. Each dimension is interacted with by the classifier in order to accomplish this goal. The number of folds used in cross-validation for determining the model's quality can be used as an analogy for this interaction. Using a learning technique for feature subset evaluation raises the computational complexity and the problem of overfitting. Option 3 is known as Embedded or Hybrid and it integrates the feature selection and knowledge acquisition processes. A feature selection and a machine learning algorithm can be improved by using this method. Feature Selection-Perceptron, Recursive Feature Elimination for Support Vector Machines, are some of the ways for feature selection using embedded approach. It is less expensive to run than the wrapper method. SVM Method of recursive feature reduction, Least Absolute Shrinkage and Selection Operator (LASSO) are some of the incorporated methods.

## 4. Symmetric Uncertainty

First thing to consider while calculating the correlation between features for feature selection is:
C-Correlation: How to decide that whether the feature is relevant to class of not.
F-Correlation: How to decide that whether the selected relevant feature is redundant or not with respect to other features.
Features are relevant if their values are correlated with the class, otherwise they are irrelevant [21, 22]. As in eq. (1), a dimension $d_i$ is relevant iff there exists some dimension $d$ and class C for which probability $p(D_i = d_i) > 0$ such that

$$p(C = c \mid D = d) \neq p(C = c) \quad (1)$$

Let D be a full set of dimensions. $d_i \in D$ is a dimension, and $S_i = D - \{d_i\}$. Then Definition of dimension relevance categorized relevance into three categories:

Definition 1 (Strong Relevance): As represented in eq. (2), a dimension $d_i$ is strongly relevant if and only if

$$p(C|d_i, S_i) \neq p(C|S_i) \quad (2)$$

Definition 2 (Weak Relevance): As in eq. (3), a feature $d_i$ is weakly relevant if and only if

$$p(C|d_i, S_i) \neq p(C|S_i) \quad (3)$$

$^a$ $c_i' \subset S_i$, such that $p(C|d_i, c_i') \neq p(C|c_i')$

Definition 3 (Irrelevance): As in eq. (4), a feature $d_i$ is irrelevant if and only if

$$\text{For all } S_i \in C,\ p(C|d_i, c_i') = p(C|c_i') \quad (4)$$

Strongly relevance feature shows that feature is strongly required for optimal set of features. Weakly relevance feature shows that it is not always required but may require for optimality in some cases irrelevant feature shows that feature is not required at all. Symmetric Uncertainty (SU) is the measure of fitness of any dimension for selection between dimensions and the target class. The dimension having higher SU value has importance than other dimensions.

Symmetric uncertainty measure, as represented in eq.(5), calculate the correlation between features using formula

$$SU = \frac{2 * IG(A|B)}{H(A) + H(B)} \quad (5)$$

Where H(A) and H(B) is the entropy of feature A and B respectively. H(A) is represented as in eq.(6)

$$H(A) = -p(A) * \log_2 p(X) \quad (6)$$

Where IG(A,B) is the Information Gain between features A and B. IG is the amount by which the entropy of B decreases. It represents the additional information about B provided by X. Information Gain, as in eq. (7) is represented by

$$IG(A|B) = H(A) - H(A|B) \quad (7)$$

Where H(A|B) represents the uncertainty of A given that the value of B is known. H(A|B) is represented as in eq. (8)

$$H(A|B) = -p(B) * p(A|B) \log 2(p(A|B)) \quad (8)$$

Where $p$(B) is the prior probability for all values of feature B.

IG is symmetrical measure as order of features A and B has no effect on the value of IG. It is the measure of correlation between two dimensions where its 0 value represents that two dimensions are independent of each other. It is defined as the difference between the sum of the marginal entropies and their joint entropy.

## 5. Feature Redundancy Estimation

Features after eliminating irrelevant features are grouped into non overlapped clusters for identifying redundant features. Correlation Coefficient (CCE) between features is used for this purpose of redundancy removal, because it can be applied on numerical data also [20, 21]. CCE is used to measure linear relationship between variables. CE gives value between -1 and 1 and its absolute value gives the correlation level between features. CCE between features a and b is calculated using equation (9).

$$CCE(A,B) = \left| \frac{\sum_{i=1}^{N} A_i B_i - N\bar{A}\bar{B}}{\sqrt{\sum_{i=1}^{N} A_i^2 - N\bar{A}^2}\ \sqrt{\sum_{i=1}^{N} B_i^2 - N\bar{B}^2}} \right| \quad (9)$$

CCE gives the value between 0 and 1. High CCE value represents that the feature is relevant. High dimensional dataset like gene expression consists of unbalanced data also in each class. These data also contain noise, so discretisation may be applied for noise removal. For this, individual feature is standardised to have 0 mean and 1 variance. These values are discretised into -1, 0 and 1 which represents three states.

Redundancy is finding out by calculating the Mutual Information or Symmetric Uncertainty (SU). If SU is high, then two features are redundant. SU can be combined with minimum spanning tree (MST) for clustering features. After removing the irrelevant features by comparing criteria for feature elimination based on redundancy are based on the concepts discussed in previous section. Redundancy is defined as

Given a set of features $S$ subset of $D$, a feature $d_i$ | ($d_i$ belongs to $D$ and $d_i$ not belongs to $S$) as in eq. (10), is non-redundant with respect to $S$ iff:

$$P(p(d_i|S) = p(d_i)) = 1 \quad (10)$$

Redundancy of any feature is evaluated by measuring the relevance between feature and its class. Redundancy of any dimension is evaluated by its relevancy. $R_{ac}$ is relevance between feature $a$ and class $c$. Relevance $R_{ab}$ is relevance between feature $a$ and $b$, where $a \neq b$. There can be following probability of relevance value between features.

If $R_{ac}$ = Large, it means that the feature '$a$' consist of more information about the class.

If $R_{ab}$ is large, it means that correlation between feature $a$ and $b$ is strong. If $R_{ab}$ =1, it means that features '$a$' and '$b$' are completely related, which means that feature a is redundant. If $R_{ab} \neq 1$, it means that it is not predictable to specify that feaure $a$ is redundant or not.

If $R_{ab}$ is small, means that correlation between feature $a$ and $b$ is weak. It means that feature $a$ is not redundant.

If $R_{ac}$ is small, it means that feature 'a' consists of less information about the class $C$.

Based on these probable values of relevance, criteria for redundancy can be defined as:

Redundant: if $R_{ab}$ is large
Not redundant: if $R_{ab}$ is small

The correlation between a feature $F_a$ ( $F_a \in S$ and the class $C$ is strong iff:

$SU_{i,c} > \lambda$ (Threshold), and
$\forall F_b \in S'$ $(b \neq a)$, and
There exists no $F_b$ such that $SU_{b,a} \geq SU_{a,c}$

## 6. Gene Expression Data

It is possible to analyse hundreds of genes simultaneously using microarray gene expression data. A gene's mRNA expression profile indicates whether or not that gene is active. Early detection of disease can be achieved by studying the molecular state of the cell using this genetic expression data.
Microarray data can be analysed with the help of a gene expression matrix, which is a table in which each column represents a gene and each row a sample record, each with the label *ca*. To indicate the measurement of the *i*th gene for a *j*th record, the expression levels for all of the genes in that record are represented as $d_{ij}$.
In other words, consider D to be a three-dimensional feature set as represented in Fig. 3. Selecting an ideal subset of features D' is the goal in feature selection, as long as P(C|D') equals or exceeds P(C|D).

| Gene1 | Gene2 | . | . | Gene n | Class |
|---|---|---|---|---|---|
| $d_{11}$ | $d_{12}$ | . | . | $d_{1n}$ | $c_1$ |
| $d_{21}$ | $d_{22}$ | . | . | $d_{2n}$ | $c_2$ |
| . | . | . | . | . | . |
| $d_{m1}$ | $d_{m1}$ | . | . | $d_{mn}$ | $c_m$ |

**Figure 3:** Gene expression data representation

Objective of classification genetic expression is to predict the class of the unknown records by developing a model for identification of genes. Major difficulty in gene expression data is the high dimensionality of the data and limited number of records available. Classification of microarray involves selection of features using SU and removal of redundancy by comparing relevance between features and class.
Performance Evaluation of feature selection can be measured by classification accuracy is measure of performance of the algorithm using number of selected features. These two performance factors as represented in eq. (11) and eq. (12) are calculated as

$$\text{Performance} = w * \text{Accuracy} + w2 * (1 - n / N) \quad (11)$$

Where $w_a$ and $w_b$ are weight coefficients used to represent the performance function.
Accuracy of the algorithm for classification is as follows:

$$\text{Accuracy} = Nc * 100\% / Nc + NI \quad (12)$$

Where Nc and NI are number of correctly classified features and number of incorrectly classified features respectively.

## 7. Proposed NNFSRR Method

In developing an algorithm for nearest neighbor search using approximation method of dimension selection and relevance and redundancy analysis, primary objective is to select a feature subset that represents the optimal subset and then search within the dataset of selected features. Method uses filter method for feature selection and wrapper method for classification. “Table 1” represent some preliminaries and notations which are used in algorithm.

Table 1: Preliminaries and Notations

| Symbol | Represents |
|---|---|
| $S$ | Set of $d$ number of features |
| $S'$ | Set of selected features |
| $S''$ | Set of selected non redundant features |
| $d_1, d_2, \ldots d_n$ | Set of $n$ number of dimensions |
| $Q$ | Query |
| $X_1, X_2, \ldots X_m$ | $M$ number of clusters of data |

| | |
|---|---|
| $C_m$ | Centroid of cluster $X_m$ |
| *KNN* | *K* number of nearest neighbors |
| *FLAG* | To continue or stop the search |
| $d(q, C_m)$ | Distance between query and centroid of *m* number of cluster |
| $\lambda$ | Threshold to select *SU* |
| $SU_d$ | Symmetric Uncertainty |

Proposed NNFSRR method consists of sub procedures for nearest neighbor search, feature selection, redundancy removal and classification of high dimensional data. Complete process is represented through flow chart as in Fig 4. Sub procedures are described as follows:

The K nearest neighbours of Cluster x are located in the array S' by employing the KNNSearch(x, S', K) algorithm. This is a straightforward application of the K-Nearest Neighbor (KNN) classification function, which sorts occurrences according to the degree of similarity between them. The function is merely approximated locally in this form of lazy learning, and the whole computation is deferred until classification takes place. When classifying an item, the majority rule is the one that is utilised. At every point in time, K is a positive integer. The items that will serve as neighbours are selected from a list of things for which the appropriate classification is already known. Procedure is better represented in the proposed framework as in Fig. 5 for classification of selected genetic data for nearest neighbour search.

Fselection(D) - This function takes gene expression data with full dimensional features and returns the features selected by the Symmetric Uncertainty method based on the threshold value.

Rremoval( ) - This function compares the SU value of feature with the other features for removal of redundant feature having higher SU value. This function obtains multiple feature subsets, where each partial subset consists of non-redundant features. In this procedure all features are sorted in descending order based on their SU value with the class. Feature having large value of SU represents higher distinction as compared to other features in the class and thus is more useful for classification. After SU computation, threshold value is set, which is them compared with the SU value. A feature is selected if its value of SU is higher than threshold, otherwise feature is not selected. All the selected features are then stored in subset S' which is the subset of all relevant features derived from the original set S.

GClassify(Class) – This function assigns the class of gene expression data using the Naive Bayes classification.

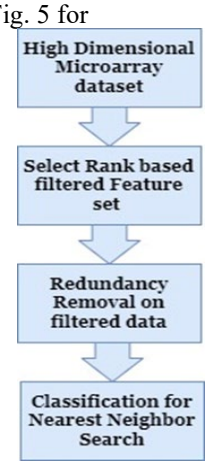


**Figure 4:** Flow Chart of Microarray data classification

**NNFSRR Algorithm:**

Pseudo-code of the proposed NNFSRR Search algorithm using feature selection and redundancy removal is as follows:

Input: Dataset $S= \{d_1, d_2, ..., d_n C\}$, Threshold $\lambda$, Query *q*.
Output: *K* Nearest Neighbors

1. Begin
2. Initialize array *S*'=NULL,
3. For *i*=1 to *n*
4. *S'* = Fselection(*S*, $\lambda$)
5. *S''* = Rremoval(*S'*)
6. K= KNNSearch(*S''*, *q*)

**Fselection(S, $\lambda$)**

1. Begin
2. Initialize *i*=0, $S_{rel}$ = NULL,
3. Compute $SU_i$ for each feature $d_i$
4. If $SU_i >= \lambda$ then
5. Insert $d_i$ into $S_{rel}$
6. Return $S_{rel}$
7. End

**Rremoval(*S'*)**

1. Begin
2. If ($d_i$ < >NULL)
3. S''=S'
4. If $SU_i$>= $SU_{i+1}$ \\Compare $SU_i$ with $SU_{i+1}$
5. $S''$= $d_i$ else $S''$= $d_{i+1}$
6. Return $S''$

**KNNSearch(*S"*, *q*)**

1. Set FLAG=1, *KNN*=NULL
2. Calculate centroid of each cluster of elements in each class.
3. Select $C_n$ having minimum value
4. Search *KNN* in cluster $C_n$
5. Return array KNN.

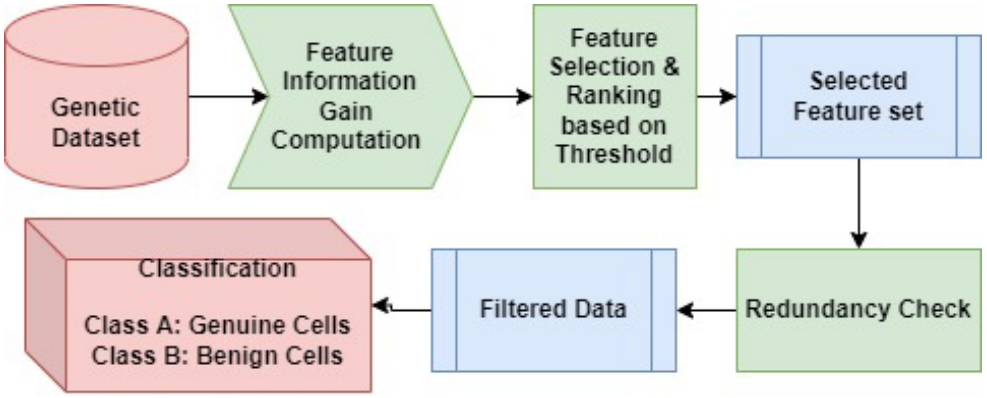


**Figure 5**: Proposed framework of Feature Selection and Redundancy Removal of Genetic Data Classification

Proposed framework of feature selection and redundancy removal of genetic data classification is one of the major key contributions in research. A two-stage framework for detection of class to gain better classification is represented as in Figure 5. Framework has a novel aspect of selecting the genes based on information gain and symmetric uncertainty between features in first stage and redundancy removal in second stage of the framework and then accurately computes the classification task.

## 8. Data Description and Experimental Result

The benchmarks and genetic data set from UCI machine learning repository is used in the study are summarised in Table 2. More than 2,000 characteristics and as few as 1,500 approximate samples can be found in the data sets. After then, the samples are randomly divided between the training and testing datasets. The training set includes 75% of the recordings, while the test set includes just 30% of the recordings. For each data set, we ran five separate tests to see how accurate the final categorisation was. Five rounds of cross validation are used in training. Cross-validation method known as "leave one out" method was used on data sets having less than 100 examples. One sample is excluded from the training set for the purpose of leave-one-out cross-validation; this process is done $n$ times with a different training set size each time ($n$-1). Additional to this, the fuzzy K-NN classifier is used to avoid ties when there are more than two categories in the data set.

The proposed technique given in this research was evaluated using data from four different datasets: the lymphoma dataset, the Ova Colon dataset, colon dataset and the Golub dataset. All 62 samples in the lymphoma dataset have 4026 features. The Ova colon dataset has 15,451 samples and 10,937 features. Two groups were formed based on the numerical representations of the tissue samples. This dataset was obtained from the mldata machine learning dataset repository and is called the Ova-colon. This is a cancer of gene expression. A total of 801 individuals and 20,531 traits are included in the RNA-Seq dataset. The ilumina HiSeq platform was used to measure the RNA-Seq gene expression levels, which are unique to each sample. Cancer genome atlas pan-cancer analysis yielded the original dataset.

Table 2: Dataset description

| Dataset | #Attributes | #Instances | #Classes |
|---|---|---|---|
| OVA-Colon | 10,937 | 1545 | 2 |
| Lymphoma | 4027 | 66 | 3 |
| Colon Tumor | 2001 | 61 | 2 |
| Golub data | 7129 | 72 | 2 |

On these dataset, Threshold value is set to 0.

Numerous resources that are accessible to the general public have already put the techniques outlined in the previous sentence into practise. These tools are accessible in a variety of programming languages, including Java, MATLAB, and R programming, among others. WEKA machine learning environment and related library functions are used for experimentation of the investigations which is a set of machine learning algorithms for data mining tasks written in Java, is used in all compared studies with the other approaches as the tool provides accurate analytics result in high dimensions. Methods for analysing high-dimensional data, such as classification and feature selection, are included in WEKA. The experiment is carried out on a platform that has an Intel(R) Core(TM) i7-2400CPU operating at 3.10GHz, and it also has Microsoft The results of a comparison of the classification error rates (in percentage) produced by the suggested NN Search method, which makes use of Feature Selection and SU, and other classifiers for various data sets are presented in Table 3. One of the result after application of feature selection computed using tool is represented as in Figure 6. All of the data sets have acquired a greater level of accuracy as a result of the combination of feature selection and symmetric uncertainty computation. Classification accuracy of different methods on Lymphoma dataset is represented as in Table 3. Based on above result classification accuracy of other methods is also calculated. According to the findings, it is abundantly evident that the proposed strategy for feature selection is superior to a great deal of well-established classifiers. As a result, the proposed strategy has attained the best performance in three out of the four data sets. The suggested technique achieves the highest performance for the Colon Tumor dataset, surpassing both the Admenn and k-NN classifiers. The proposed strategy has not yet attained the highest level of performance in the Colon Tumor data sets. Due to the fact that just a restricted number of features were chosen, this data set is unable to combine the benefits of feature selection and SU. These features may not be sufficient enough to classify comparably. The average classification error rate is displayed in Table 4, throughout the course of the five distinct splits of each data set for the various algorithms. As can be seen from the error rate, the suggested classifier achieves significantly better results than its competitors in terms of the error rate.

Table 3: Classification accuracies (%) of different feature selection methods with NN on Lymphoma database.

| Dimensions Data | 10 | 30 | 50 | 70 | 90 | 110 | 130 | 94.79 |
|---|---|---|---|---|---|---|---|---|
| Adamenn | 89.58 | 96.88 | 94.79 | 95.83 | 97.92 | 97.92 | 96.88 | 94.79 |
| C4.5 | 68.75 | 84.38 | 86.46 | 88.54 | 87.50 | 85.42 | 88.54 | 94.79 |
| CDW | 88.54 | 95.83 | 94.79 | 94.79 | 95.83 | 96.88 | 96.88 | 94.79 |
| Dann | 88.54 | 91.67 | 93.75 | 93.75 | 93.75 | 88.54 | 88.54 | 94.79 |
| K-NN | 89.58 | 94.79 | 95.83 | 97.92 | 95.83 | 97.92 | 97.92 | 94.79 |
| Naïve Bayes | 94.79 | 94.79 | 96.88 | 96.88 | 97.92 | 97.92 | 97.92 | 94.79 |
| Proposed NNFSRR | 96.12 | 95.67 | 96.28 | 97.95 | 97.95 | 97.95 | 97.95 | 94.79 |

Table 4: Average classification error rate (in %)

| **Dataset** | **Features selected using Proposed method** | **Adamenn** | **C4.5** | **CDW** | **Dann** | **k-NN** | **N-Bayes** | **Proposed Method (SD)** |
|---|---|---|---|---|---|---|---|---|
| **OVA-Colon** | 1020 | 30.7 | 28.2 | - | 24.4 | 24.1 | 28.2 | 10.4 |
| **Lymphoma** | 1785 | 25.2 | 20.7 | 24.1 | 23.2 | 22.4 | 25.5 | 8.9 |
| **Colon Tumor** | 135 | 26.5 | 24.3 | - | 19.6 | 25.9 | 24.5 | 25.7 |
| **Golub data set** | 1967 | 17.8 | 16.9 | 20.1 | 16.5 | 22.8 | 21.1 | 11.2 |

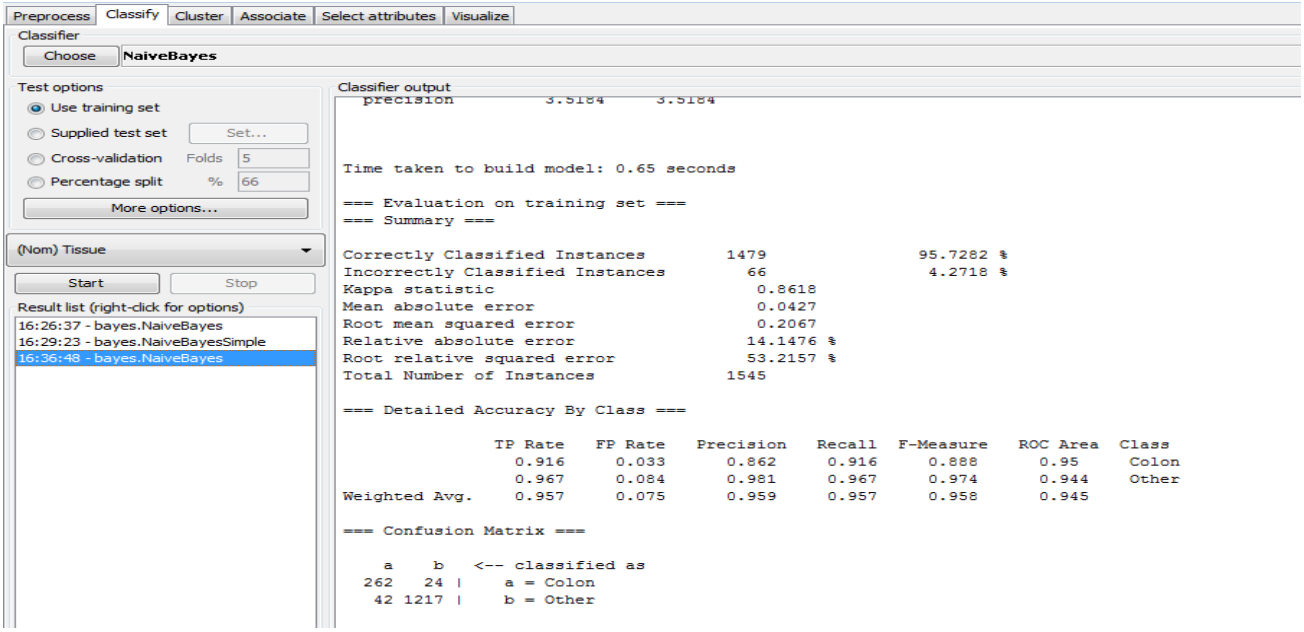


**Figure 6:** Screenshot of result after applying feature selection using Naïve Bayes Algorithm

## 9. Conclusion

As part of this research, we present a method of nearest neighbor search selecting feature subsets from high-dimensional datasets using feature selection and ranking based on symmetric uncertainty and redundancy removal using correlation coefficient. When working with numerical data, it is found that this technique performs extremely well; we can also adapt this strategy to work with mixed categories of data (such as nominal and categorical) without first normalising the data into discrete values. For high-dimensional datasets and biological datasets with millions of features, this method might be helpful in solving the problem of feature selection. In the near future, we hope to extend the application of this method to extremely high dimensional data, so that it can be tested on data with a very high dimensionality.

## Data Availability

The dataset used in this paper is available from the corresponding author upon request.

## Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

## References

[1] **Journal article:** Koul, Nimrita, and Sunilkumar S. Manvi. "Feature Selection from Gene Expression Data Using Simulated Annealing and Partial Least Squares Regression Coefficients." Global Transitions Proceedings (2022).

[2] **Journal article:** Hambali, Moshood A., Tinuke O. Oladele, and Kayode S. Adewole. "Microarray cancer feature selection: review, challenges and research directions." International Journal of Cognitive Computing in Engineering 1 (2020): 78-97.

[3] **Journal article:** P. E. Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany and X. Chen, "An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data," in IEEE Access, vol. 9, pp. 155353-155369, 2021, doi: 10.1109/ACCESS.2021.3123090.

[4] **Journal article:** Gumaei, Abdu, Rachid Sammouda, Mabrook Al-Rakhami, Hussain AlSalman, and Ali El-Zaart. "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression." Health Informatics Journal 27, no. 1 (2021): 1460458221989402.

[5] **Journal article:** Tripathy, Jogeswar, Rasmita Dash, Binod Kumar Pattanayak, Sambit Kumar Mishra, Tapas Kumar Mishra, and Deepak Puthal. "Combination of Reduction

Detection Using TOPSIS for Gene Expression Data Analysis." Big Data and Cognitive Computing 6, no. 1 (2022): 24.

[6] **Journal article:** Potharaju, Sai Prasad, and M. Sreedevi. "Distributed Feature Selection (DFS) Strategy for Microarray Gene Expression Data to Improve the Classification Performance." Clinical Epidemiology and Global Health, vol. 7, no. 2, June 2019, pp. 171–176, 10.1016/j.cegh.2018.04.001. Accessed 3 June 2020.

[7] **Journal article:** Tripathy, Jogeswar, Rasmita Dash, Binod K. Pattanayak, Sambit K. Mishra, Tapas K. Mishra, and Deepak Puthal. 2022. "Combination of Reduction Detection Using TOPSIS for Gene Expression Data Analysis" Big Data and Cognitive Computing 6, no. 1: 24. https://doi.org/10.3390/bdcc6010024

[8] **Journal article:** Chuang, Li-Yeh, Cheng-Huei Yang, and Cheng-Hong Yang. "Tabu search and binary particle swarm optimization for feature selection using microarray data." Journal of computational biology 16, no. 12 (2009): 1689-1703.

[9] **Journal article:** Buaba, Ruben, Abdollah Homaifar, William Hendrix, Seung Woo Son, Wei-keng Liao, and Alok Choudhary. "Randomized algorithm for approximate nearest neighbor search in high dimensions." Journal of Pattern Recognition Research 1 (2014): 111-122.

[10] **Journal article:** Fan, Bin, Qingqun Kong, Baoqian Zhang, Hongmin Liu, Chunhong Pan, and Jiwen Lu. "Efficient nearest neighbor search in high dimensional hamming space." Pattern Recognition 99 (2020): 107082.

[11] **Journal article:** F. Korn, B. Pagel, and C. Faloutsos, "On the dimensionality curse" and the self-similarity blessing"," IEEE Transaction Knowledge Data Engineering, vol. 13, no. 1, pp. 96–111, 2001.

[12] **Journal article:** G, Vasanthi. "Nearest Neighbors Search Algorithm for High Dimensional Data." Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP8, 30 July 2020, pp. 1215–1218, 10.5373/jardcs/v12sp8/20202636. Accessed 13 Nov. 2020.

[13] **Conference:** P. Zhu, X. Zhan and W. Qiu, "Efficient k-Nearest Neighbors Search in High Dimensions Using MapReduce," 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, 2015, pp. 23-30, doi: 10.1109/BDCloud.2015.51.

[14] **Journal article:** Kushilevitz, Eyal, Rafail Ostrovsky, and Yuval Rabani. "Efficient search for approximate nearest neighbor in high dimensional spaces." SIAM Journal on Computing 30, no. 2 (2000): 457-474.

[15] **Journal article:** Dubiner, Moshe. "A Heterogeneous High-Dimensional Approximate Nearest Neighbor Algorithm." IEEE Transactions on Information Theory, vol. 58, no. 10, Oct. 2012, pp. 6646–6658, 10.1109/tit.2012.2204169.

[16] **Journal article:** H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," Procedia Computer Science, vol. 23, pp. 5–14, 2013.

[17] **Journal article:** Liu, Yingfan, Hao Wei, and Hong Cheng. "Exploiting lower bounds to accelerate approximate nearest neighbor search on high-dimensional data." Information Sciences 465 (2018): 484-504.

[18] **Journal article:** Fathi H, AlSalman H, Gumaei A, Manhrawy IIM, Hussien AG, El-Kafrawy P. An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data. Comput Intell Neurosci. 2021;2021:7231126. Published 2021 Dec 29. doi:10.1155/2021/7231126

[19] **Conference:** D. Koller and M. Sahami. Hierachically classifying documents using very few words. In Machine Learning: Proceedings of the Fourteenth International Conference. Morgan Kaufmann, 1997.

[20] **Journal article:** Liu, Huiqing, Jinyan Li, and Limsoon Wong. "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns." Genome informatics 13 (2002): 51-60.

[21] **Journal article:** L Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Mach. Learn. Res., vol. 5, pp. 1205–1224, 2004.

[22] **Journal article:** S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review," in: C. Aggarwal and C. Reddy (eds.), Data Clustering: Algorithms and Applications, CRC Press, 2013.

[23] **Journal article:** Jianzhong Wang, Shuang Zhou, Yugen Yi, Jun Kong, "An Improved Feature Selection Based on Effective Range for Classification", The Scientific World Journal, vol. 2014, Article ID 972125, 8 pages, 2014. https://doi.org/10.1155/2014/972125