Mohammed MAIZA, Samira CHOURAQUI, Chahira CHERIF, Abdelmalik TALEB-AHMED

DOI: 10.15199/48.2025.04.14

Cancer classification through the selection of genes extracted from microarray data

Klasyfikacja nowotworów poprzez selekcję genów wyodrębnionych na podstawie danych z mikromacierzy

Abstract: *In this work we propose to compare tree methods for feature selection in the binary classification context. We focus on the case where the number of variables is very large and much more important than the sample size, as is the case in most microarray data. Four classification algorithms were selected: Decision Tree (DT), k-Nearest Neighbors (K-NN), Neural Networks (NN), and Support Vector Machines (SVM), along with three filter-based feature selection criteria using mutual information: MIM (Mutual Information Maximization), JMI (Joint Mutual Information), and MRMR (Max-Relevance Min-Redundancy).First, we applied these classification algorithms to the microarray datasets without any preprocessing or feature selection, allowing us to establish a baseline for assessing the impact of preprocessing and feature selection on improving classification performance. The second method involved classification after data preprocessing but without feature selection, which enabled us to evaluate the impact of preprocessing on classification performance. Finally, the last method applied classification after both preprocessing and feature selection, allowing us to measure the combined impact of preprocessing and feature selection on improving classification performance.*

Streszczenie: *W niniejszym badaniu proponujemy porównanie trzech metod selekcji cech w kontekście klasyfikacji binarnej. Skupiamy się na przypadkach, w których liczba zmiennych jest wyjątkowo duża i znacznie przewyższa rozmiar próbki, co często występuje w danych z mikromacierzy. Wybraliśmy cztery algorytmy klasyfikacji: drzewo decyzyjne (DT), k-najbliższych sąsiadów (K-NN), sieci neuronowe (NN) oraz maszyny wektorów nośnych (SVM), a także trzy kryteria selekcji oparte na filtrach wykorzystujące informację̧ wzajemną: MIM (Maksymalizacja Informacji Wzajemnej), JMI (Wspólna Informacja Wzajemna) oraz MRMR (Maksymalna Trafność, Minimalna Redundancja). Początkowo zastosowaliśmy te algorytmy klasyfikacji bezpośrednio do zbiorów danych z mikromacierzy bez przetwarzania wstępnego ani selekcji cech, aby ustanowić punkt odniesienia do oceny wpływu przetwarzania wstępnego i selekcji cech na wydajność klasyfikacji. Drugie podejście polegało na klasyfikacji po przetworzeniu danych, ale bez selekcji cech, co pozwoliło nam ocenić wpływ przetwarzania wstępnego na wyniki klasyfikacji. W trzecim podejściu klasyfikację przeprowadzono po przetwarzaniu wstępnym i selekcji cech, co umożliwiło ocenę łącznego wpływu tych kroków na poprawę wydajności klasyfikacji.*

Keywords: Microarray Data, Gene Selection,Decision Tree,Neural Networks,SVM,k-NN, Mutual Information
Słowa kluczowe: Dane mikromacierzy, wybór genów, drzewo decyzyjne, sieci neuronowe, SVM, K-NN, informacja wzajemna

Introduction

Advances in molecular biology have revealed that genetic alterations, such as mutations or chromosomal rearrangements, can disrupt mechanisms regulating cell growth and division, leading to the formation of uncontrolled cancerous cells. Identifying the genes involved in these processes is crucial for early diagnosis and the development of targeted therapies. However, the abundance of genomic data generated by sequencing technologies presents a major challenge for selecting the relevant genes that distinguish cancerous

samples from healthy ones [1].

In this context, gene selection plays a crucial role in cancer detection and diagnosis. Despite the tens of thousands of genes in the human genome, only a fraction is truly relevant for characterizing and differentiating cancerous tissues from healthy ones [2]. However, the noise and redundancy inherent in genomic data complicate this task. The central issue of this study is therefore the following: how can we effectively

identify the most informative and relevant subset of genes for the accurate and reliable detection of different types of cancer ?

Cancer, as a serious disease characterized by the uncontrolled proliferation of cells that can invade healthy tissues, represents one of the primary causes of mortality. Its diversity is reflected in the numerous types of cancers classified by their tissue origin, such as breast cancer, lung cancer, or leukemia, and its multifactorial causes are rooted in genetic, environmental, and lifestyle factors. Early detection is of paramount importance, as it significantly increases the chances of recovery [3].

To resolve this issue, we proposed a structured process involving the application and comparison of classification algorithms on genetic datasets across three distinct scenarios. First, classification algorithms (DT, K-NN, NN, SVM) are directly applied to the raw data [4][5]. Next, the data is preprocessed (normalization, handling of missing values, noise reduction) before applying the same algorithms. Finally, the preprocessed data is optimized with gene selection techniques MMI, JMI and MRMR [6], before applying the classification

algorithms. The effectiveness of each method is evaluated by comparing classification accuracy and complementary performance metrics such as Precision, Recall, and F1-score [7]. This systematic approach aims to improve cancer detection accuracy and provide a comprehensive evaluation of the performance of different algorithms, thus offering an optimized methodology for identifying the most relevant genes.

DNA microarray

Studying the entire transcriptome of a cell is of considerable importance in gaining a better understanding of the organism's functional mechanisms. In the past, biologists were able to measure the expression level (the number of transcripts) of a small number of genes at a time. Microarray technology now enables them to study thousands of genes simultaneously [8] an advance that will enable them to determine the complex relationships between genes.

Biochips have a wide range of applications. For example, it is now possible to understand the dynamics of the transcriptome and the genetic networks involved, and to classify tumors according to their molecular signature. Conversely, genes can be explored to determine the function of an unknown gene. This knowledge can then be used to better understand diseases and develop new medicine [9].

Gene expression data

Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein [10][11]. Microarray technologies provide the opportunity to compute the expression level of tens of thousands of genes in cells simultaneously [12]. We speak of gene expression when the information stored in our DNA is converted into instructions for the production of proteins or other molecules. Gene expression is the conversion of the DNA sequences into mRNA sequences by transcription then translated into amino acid sequences called proteins.

Microarray data classification

Microarray data are presented as gene expression matrices derived from image analysis, where genes are represented by rows and various samples such as tissues or experimental conditions are represented by columns. Each cell value indicates the expression level of a specific gene in a particular sample. The disease classification system based on microarray data utilizes labeled gene expression samples to generate a classifier model, which categorizes new data samples into predefined disease groups [13].

Supervised classification

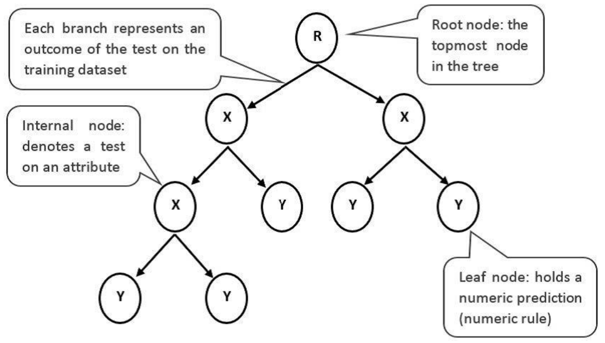
Supervised classification involves assigning a class or category to new observations based on a model learned from training data that includes observations with known classes. The goal is to predict the value of a categorical target variable by capturing the relationships between this variable and the descriptive attributes of the observations.

Various algorithms exist, based on statistical, geometric, neural, or margin theory approaches. We will detail the principles and characteristics of each method: DT, which recursively partitions the data space; the K-NN method, which classifies based on the nearest points, the NN capable of learning complex nonlinear relationships, the kernel methods that project the data into a feature space, and finally the SVM that determine the optimal hyperplane for class separation [14][15].

Decision tree (DT)

DTs are recognized for their robustness and proven performance in various industrial and research applications. Despite their significant history, decision tree construction algorithms remain largely unchanged, aiming to segment each node appropriately using a specific criterion and to construct the tree recursively from the root to the leaves [16].

A DT, in the field of computer science, is a classification method in the form of a tree, consisting of a root node, internal nodes representing tests on the features, branches representing the outcomes of the tests, and leaves representing the predicted values. The construction of the tree recursively divides the feature space into binary partitions, maximizing a class purity criterion at each step [17].



Each branch represents an outcome of the test on the training dataset

Root node: the topmost node in the tree

Internal node: denotes a test on an attribute

Leaf node: holds a numeric prediction (numeric rule)

Fig. 1. Decision Tree general structure

The construction of the tree is essential in this classification method. The algorithms descend during the tree construction, dividing the sample at each step into the most homogeneous subsets possible. This recursive procedure follows a process from the root to the leaves, selecting at each node the most relevant split based on the most discriminative test attribute [18]. Figure 1 illustrates DT general structure [19].

DT are constructed by following a very explicit set ofrules, which facilitates the user's understanding of the results. They generally require few resources, resulting in relatively short training and testing times.

The k-nearest neighbors algorithm (K-NN)

K-NN algorithm is one of the simplest classification methods [20]. It operates by using a set of labeled samples, where an unknown sample is assigned to the class represented by the majority of its k most similar samples. The principle is as follows: given a new instance x described by p attributes, the algorithm identifies the k nearest instances to x within the training set. The class of x is then determined by the majority class among these k nearest neighbors. This depends on three parameters [21]:

- The number of neighbors k.
- The distance measure between examples, typically Euclidean for numerical attributes.
- The decision rule: the majority class among the k neighbors.

Neural networks (NN)

A NN is an information processing system inspired by the functioning of biological neurons. It consists of interconnected artificial neurons that process information in a parallel and distributed manner [22].

The network learns by adjusting the weights of the connections between neurons, called synapses, which store knowledge similarly to biological synapses.

NN have several interesting characteristics. Their nonlinearity comes from the neurons, allowing for complex modeling suited to challenging problems. Adaptive learning occurs by adjusting synaptic weights in response to examples, improving the network over time.

Fault tolerance is another key quality, as knowledge is distributed in a way that makes the network robust to the failure of individual neurons.

Finally, parallel processing capability allows the network to handle multiple pieces of information simultaneously, increasing efficiency in data processing.

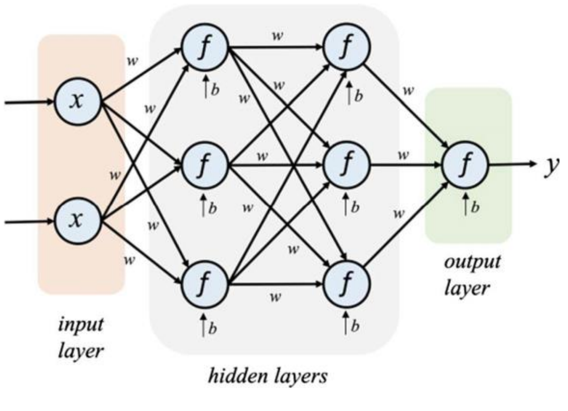


Fig. 2. Neural networks

In Figure 2 [22], the circles represent neurons arranged in the form of layers. This network has three layers: the input layer, which receives information $x_1, \dots, x_n$ through neurons via weights wi. The output layer containing neurons, providing the results of internal calculations.

Kernel Methods and Support Vector Machines (SVM)

Support Vector Machines (SVM), introduced in the 1990s, are classifiers based on two key concepts: the notion of maximum margin and the use of kernel functions. The principle is to identify the separating hyperplane between positive and negative examples, the goal is to maximize the margin, which is the distance between the decision boundary and the closest data points, known as support vectors. This is framed as a quadratic optimization problem [23].

Since data are not always linearly separable, SVMs employ a technique that projects data into a higher-dimensional space, where it is more likely that a linear separator exists. This projection is accomplished using a kernel function, which indirectly maps the data into a high-dimensional space without explicitly computing the coordinates of the points in that space [24]. Thus, SVMs are powerful classifiers, leveraging the kernel-transformed space to achieve linear separation of complex data.

Feature Selection

Feature selection is a crucial step in data mining, aimed at eliminating redundant data to improve classification algorithms. It is essential for reducing training times and preventing overfitting. Widely used across various fields, this practice requires in-depth domain knowledge and can be performed manually or with tools. Two main approaches are commonly used: *wrapper* methods and *filter* methods. Current research aims to identify an optimal subset of features that satisfies multiple objectives [25].

Let's also retain the definition presented in the context of a feature set $F = f_1, f_2, \dots, f_n$ the objective is to determine a subset F' that optimizes the performance of the learning algorithm. Formally, F' must maximize a score function v, in the following manner:

(1) $$F' = argmax_{G \in \Gamma} \{v(G)\}$$

This represents the subset F′ that maximizes the score function v over the set of candidate subsets Γ.

Feature selection techniques preserve the original representations of the features; instead, they select a subset from them. These approaches preserve the initial semantics of the features, thereby facilitating easier interpretation by domain experts. In theory, the objective is to find the optimal subset of features that maximizes the previously mentioned score function.

It is crucial that feature selection is performed solely on the training data, while the test set is subsequently used to evaluate the quality of the selected features (subsets).

Proposed approach for cancer classification

Cancer poses a major challenge to global public health, necessitating innovative approaches for its diagnosis and classification.

The proposed approach for cancer classification is based on a sequence of steps. First, a dataset containing

genes categorized by cancer types is gathered. This data is then subjected to preprocessing to prepare it for analysis.

In the next step, classification algorithms are applied to the data without utilizing detection algorithms. Subsequently, the process is repeated, but this time detection algorithms are emploprior to the classification stage. This du yed al approach aims to enhance the accuracy of cancer type classification by leveraging both classification and detection algorithms in an integrated manner. The proposed approach is illustrated in Figure 3.

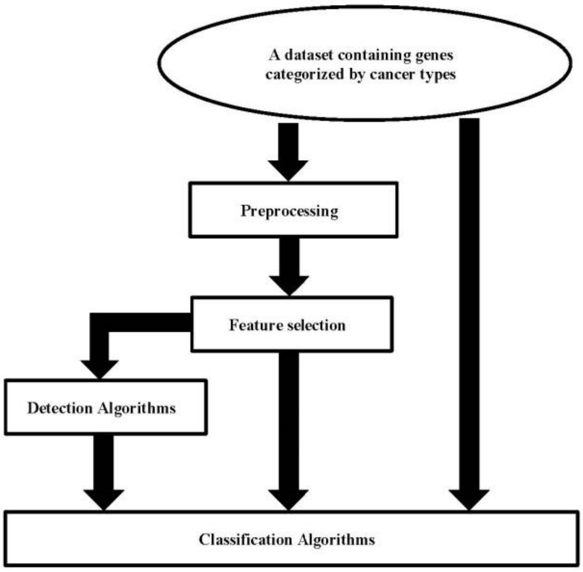

A dataset containing genes categorized by cancer types
Preprocessing
Feature selection
Detection Algorithms
Classification Algorithms

Fig. 3. Proposed approach

We needed to restrict our selection to a limited subset of our data in order to target the most relevant and informative features among the genetic data, thus developing accurate and reliable classification models. When working with genomic or gene expression data, the presence of a large number of features (genes) can make the training of classification models inefficient and computationally expensive, with an increased risk of overfitting.

This is why the application of feature selection techniques is crucial for reducing data complexity. The goal is to retain only a subset of the most informative and discriminative genes for cancer type classification. These methods enable the identification of the most significant genes for distinguishing between different cancer classes while eliminating redundant or irrelevant genes.

As a result, the performance of the models is significantly enhanced in terms of accuracy, recall, and generalization. In this work, we use mutual information as a measure of attribute relevance.

Data Subset Selection

We needed to restrict our selection to a limited subset of our data to target the most relevant and informative features within the genetic data, thereby developing precise and reliable classification models. When working with genomic or gene expression data, the large number of features (genes) can make training classification models inefficient and computationally expensive, with an increased risk of overfitting. Therefore, applying feature selection techniques is essential to reduce data complexity.

The goal is to retain only a subset of the most informative and discriminative genes for cancer type classification. These methods help identify the most significant genes for distinguishing between various cancer classes, while eliminating redundant or irrelevant genes. Consequently, model performance is significantly improved in terms of accuracy, recall, and generalization.

In this work, we use mutual information as a measure of attribute relevance.

Mutual Information

Mutual information is a reliable indicator of the relevance between variables, making it a widely used measure in various feature selection algorithms [26]. However, its calculation can be complex, and the effectiveness of a feature selection algorithm is closely tied to the accuracy of the computed mutual information [27].

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{2}$$

where:

• *p(x,y)* is the joint probability distribution of X and Y.

• *p(x)* and *p(y)* are the marginal probability distributions of X and Y, respectively.

This equation measures the amount of information that X and Y share, quantifying their mutual dependence.

Mutual information is zero when the variables X and Y are statistically independent, which means:

$$I(X;Y) = 0 \quad \text{if} \quad p(x,y) = p(x) \cdot p(y) \tag{3}$$

This means that if the joint probability distribution of X and Y equals the product of their marginal distributions, then the mutual information is zero, indicating no dependency between the two variables.

Mutual information is linearly related to the entropies of the variables according to the following equations:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

where:

• *H(X)* is the entropy of variable X.

• *H(Y)* is the entropy of variable Y .

• *H(X,Y)* is the joint entropy of variables X and Y.

This relationship shows that mutual information can be understood as the reduction in uncertainty about one variable given knowledge of the other.

Feature selection based on mutual information filter

Mutual information quantifies the dependency relationship between two random variables. In the context of feature selection, we assess the mutual information between the target variable (the variable to be predicted) and each explanatory variable (feature) to measure their degree of mutual dependence. Features exhibiting high mutual information with the target variable are considered the most informative and relevant, and will therefore be selected for inclusion in the predictive model.

Several feature selection criteria based on mutual information have been proposed in the scientific literature. For this study, we have selected three specific criteria, which will be detailed in the following sections :

- **Mutual Information Maximization (MIM)** is a technique used in feature selection and representation learning, aiming to maximize the mutual information between the input features and the target variable. The objective is to select features that provide the most information about the target variable, thereby improving model performance [28].
  This maximization process helps in identifying the most relevant features that contribute significantly to predicting the target variable. The formulation for MIM can be expressed as:

  $$\max_{F' \subseteq F} I(X;Y) \tag{5}$$

  where:

  • F' is the subset of features selected from the original feature set F.

  • *I(X; Y)* is the mutual information between the selected features X and the target variable Y .
- **Joint Mutual Information (JMI)** is a feature selection technique that seeks to select a subset of features that maximizes the joint mutual information between the selected features and the target variable. JMI considers the collective information provided by a set of features rather than evaluating them individually [29]. In practice, JMI aims to retain features that not only provide information about the target variable but also maintain strong interdependencies among themselves, enhancing overall model performance. The formulation for Joint Mutual Information can be expressed as:

  $$\max_{F' \subseteq F} I(F';Y) \tag{6}$$

  where:

  • *I(F′; Y)* is the mutual information between the selected features F' and the target variable Y .
- **Max-Relevance Min-Redundancy (MRMR)** is a feature selection criterion that seeks choose features that are highly relevant to the target variable while minimizing redundancy among the selected features. This approach is particularly useful in high-dimensional datasets, where reducing redundancy can lead to more efficient models without sacrificing performance. This formulation emphasizes maximizing the relevance of the features to the target while penalizing the inclusion of highly correlated features, thereby promoting diversity in the selected feature set [30]. The MRMR criterion can be formulated as:

  $$\max_{F' \subseteq F} \left(I(F';Y) - \frac{1}{|F'|^2} \sum_{f_i, f_j \in F'} I(f_i; f_j) \right) \tag{7}$$

  where:

  • *I(fi; fj)* is the mutual information between the features fi and fj .

  • | F' | is the number of features in the subset F'.

Classification

Once the attributes are extracted and the feature selection based on mutual information filtering is applied, the final step is to perform the classification. In machine learning, classification consists of two phases: training and testing. During the training phase, a predictive model is constructed from the training data [31]. In the testing phase, this model is evaluated to determine whether it is accurate enough to be deployed on new data. In this work, we opted for supervised learning algorithms for the classification task, specifically k-NN, SVM, DT, and NN.

Results and discussion

We applied our method to four microarray datasets, each dataset was divided into two parts, namely training data and testing data, where training data was used for the learning process, and testing data was used in the testing process of the model obtained [32].

- Colon cancer: This dataset is of Colorectal Cancer (CRC), caused from the epithelial cells lining the

colon or rectum of the gastroin-testinal tract. It contains information of 36 patients of which 18 are positive samples, while are other 18 negative samples [33].
- Prostate Tumor: This dataset contains 102 samples and 12600 genes, out of which 52 are prostate-tumour samples and 50 are non-tumour prostate samples [34].
- Leukemia: The dataset consists of 47 samples from Acute -Lymphoblastic-Leukemia (ALL) patients and 25 cases of Acute-Myeloid-Leukemia (AML) [35].
- The Lymphoma dataset consists of 96 samples from both normal and cancerous populations of human lymphocytes, with each sample measured across 4026 genes [36].

Description of the gene expression datasets used are summarized in Table 1 .

Evaluating the performance of a predictive classification model in machine learning is crucial for selecting the optimal model.

This step allows for an assessment of the quality and reliability of the predictions made by the trained model. To conduct this evaluation phase rigorously, a key tool is utilized: the confusion matrix (see Table 2).

Table 1. Brief description of the datasets

Dataset	Genes	Training data	Testing data	Observations +1/-1
Colon cancer	2000	62	-	22/40
Prostate Tumor	12600	102	-	52/50
Leukemia	7129	38	34	27/11 - 20/14
Lymphoma	4026	60	36	45/15 - 27/9

This matrix provides a detailed visual representation that contrasts the model’s predictions with the true classes of the observations. A thorough analysis of this matrix enables the calculation and interpretation of various performance metrics (Precision, Recall, F1-score, Accuracy).

Based on these quantitative indicators, the expert can objectively compare different candidate models and select the one demonstrating the best predictive capabilities on new data [37].

Table 2. Confusion matrix

Class	Y	$\overline{Y}$
Y	TP	FP
$\overline{Y}$	FN	TN

It is a table that displays various predictions and test results [38], comparing them with real values where :
- TP (True Positive) : Number of well-predicted processes in class Y.
- FP (False Positive) : Number of processes predicted to be in class Y when they should not be.
- FN (False Negative) : Number of processes are predicted to be of the Y class when in fact they are not.
- TN (True Negative) : Number of correctly predicted processes in the Y class.
- Precision is a metric used to evaluate the accuracy of positive predictions. It measures the proportion of true positive predictions out of all positive predictions made by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

- Recall, also known as sensitivity, measures the proportion of positive observations correctly predicted relative to all actual positive observations.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

- F1-score is a metric used to evaluate the balance between Precision and Recall. It is especially useful with imbalanced datasets. The formula for the F1-score is as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

- Accuracy measures the proportion of correct predictions made by the model out of all predictions.it represents the ratio between the number of correct predictions and the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

For the classification task, we used the Python programming language (version 3.10.9)[1], which includes most of the commonly used classification algorithms.

The classification report is a performance evaluation tool for a classification model. It provides details on Precision, Recall, F1-score, and Support for each class. This report can be generated from the confusion matrix, presenting the results for each class as well as their average, either weighted or unweighted, depending on the selected option.

In Table 3, the classification report is presented after considering two classes: "Normal" and "Cancer".

Table 3. classification report

	Precision	Recall	F1-score	Support
0	0.50	0.50	0.50	294
	0.53	0.53	0.53	315
Accuracy			0.52	609
Macro avg	0.52	0.52	0.52	609
Weighted avg	0.52	0.52	0.52	609

For each class, Precision, Recall, and F1-score are calculated. Additionally, an extra row in this table provides the weighted average of the metrics, where the weight is determined by the total number of examples in each class. Finally, the "Support" column indicates the number of samples present for each of the two classes.

Table 4 presents the classification results obtained from the direct application of classification algorithms (DT, NN, SVM, K-NN), to cancer genomic datasets. This analysis was conducted without any prior data preprocessing or feature selection steps.

We used the raw datasets datasets, incorporating all available genes (attributes), without eliminating redundancies, addressing missing values, or implementing normalization techniques. Additionally, we refrained from reducing the high dimensionality of the data by selecting a pertinent subset of genes. The aim was to evaluate the unprocessed performance of various classifiers on these complex, highdimensional genomic datasets without any preliminary processing.

Table 4. Classification results without data preprocessing or feature selection

	DT	NN	SVM	K-NN
Colon cancer	0.73	0.86	0.80	0.73
Prostate Tumor	0.85	0.78	0.78	0.69
Leukemia	0.73	0.84	0.82	0.77
Lymphoma	0.88	0.85	0.84	0.86

Table 5 presents the classification results obtained after applying the classification algorithms following data preprocessing, but without the use of feature selection algorithms.

We first applied preprocessing techniques to the raw datasets. This included data cleaning to address missing values and normalization to scale all variables consistently. At this stage, we did not reduce the high dimensionality of the data by selecting a relevant subset of genes (variables); all available genes were retained.

Table 5. Classification results with data preprocessing and without feature selection

	DT	NN	SVM	K-NN
Colon cancer	0.83	0.86	0.89	0.83
Prostate Tumor	0.81	0.82	0.88	0.75
Leukemia	0.79	0.91	0.94	0.91
Lymphoma	0.90	0.90	0.93	0.87

Table 6 shows the classification results achieved after performing full data preprocessing, followed by applying different relevant feature selection methods before training and evaluating the classification algorithms.

Table 6. Evaluation results of classification algorithms after feature selection

Dataset	Feature selection	DT	NN	SVM	K-NN
Colon cancer	MIM JMI MRMR	0.89 0.98 1.00	1.00 0.98 0.77	0.90 1.00 0.80	0.92 0.93 0.90
Prostate Tumor	MIM JMI MRMR	0.75 0.80 0.48	0.67 0.73 0.89	0.75 0.75 0.69	0.88 0.90 0.88
Leukemia	MIM JMI MRMR	1.00 0.55 0.84	0.82 0.88 0.55	0.84 0.66 0.71	0.91 0.89 0.98
Lymphoma	MIM JMI MRMR	0.88 0.92 0.97	0.86 0.90 1.00	0.85 0.87 1.00	0.89 0.91 0.94

We first prepared the raw data by performing rigorous preprocessing, including imputing missing values and normalizing the variables to bring them to a consistent scale.These steps are essential to ensure data quality and consistency before applying the learning algorithms.

Table 7. Results of the DT following feature selection, presenting comparative values for Precision, Recall, F1-score, and Accuracy

Dataset	Feature selection	Precision	Recall	DT F1-score	Accuracy
Colon cancer	MIM JMI MRMR	0.74 0.78 0.80	0.72 0.85 0.87	0.78 0.78 0.80	0.78 0.71 0.78
Prostate Tumor	MIM JMI MRMR	0.75 0.88 0.89	0.87 0.79 0.91	0.85 0.85 0.79	0.82 0.78 0.71
Leukemia	MIM JMI MRMR	0.87 0.85 0.84	0.82 0.89 0.75	0.84 0.76 0.71	0.88 0.89 0.88
Lymphoma	MIM JMI MRMR	0.85 0.87 0.91	0.88 0.86 0.89	0.84 0.88 0.87	0.78 0.81 0.90

Table 8. Results of the NN following feature selection, presenting comparative values for Precision, Recall, F1-score, and Accuracy

Dataset	Feature selection	Precision	Recall	NN F1-score	Accuracy
Colon cancer	MIM JMI MRMR	0.74 0.98 0.90	0.89 0.95 0.87	0.98 0.77 0.80	0.92 0.85 0.89
Prostate Tumor	MIM JMI MRMR	0.85 0.80 0.88	0.87 0.73 0.89	0.65 0.75 0.97	0.73 0.95 0.92
Leukemia	MIM JMI MRMR	0.80 0.85 0.84	0.82 0.88 0.75	0.84 0.66 0.71	0.79 0.87 0.88
Lymphoma	MIM JMI MRMR	0.95 0.97 0.90	0.98 0.96 0.89	0.94 0.98 0.91	0.88 0.88 0.95

Table 9. Results of the SVM following feature selection, presenting comparative values for Precision, Recall, F1-score, and Accuracy

Dataset	Feature selection	Precision	Recall	SVM F1-score	Accuracy
Colon cancer	MIM JMI MRMR	0.84 0.98 1.00	1.00 0.95 0.97	0.98 0.91 0.95	0.98 0.93 0.98
Prostate Tumor	MIM JMI MRMR	0.15 0.80 0.99	0.17 0.73 1.00	0.25 0.75 0.97	0.87 0.78 0.95
Leukemia	MIM JMI MRMR	0.79 0.87 1.00	0.82 0.81 0.95	0.84 0.89 0.91	0.90 0.92 0.96

Lymphoma	MIM	0.95	0.98	0.94	0.88
	JMI	0.97	0.96	0.98	0.92
	MRMR	1.00	0.99	1.00	0.98

Next, we applied three different feature selection methods: MIM (Maximum Mutual Information), JMI (Joint Mutual Information), and MRMR (Maximum Relevance Minimum Redundancy). These algorithms identify the most relevant and informative genes (attributes) for the classification task while eliminating redundancies. This process significantly reduced the initial dimensionality of the genomic data. For each dataset and feature selection method, we trained and evaluated a range of classification algorithms.

Table 10. Results of the K-NN following feature selection, presenting comparative values for Precision, Recall, F1-score, and Accuracy

Dataset	Feature selection	Precision	Recall	K-NN F1-score	Accuracy
Colon cancer	MIM	0.86	0.72	0.88	0.86
	JMI	0.89	0.91	0.88	0.90
	MRMR	0.89	0.97	0.80	0.92
Prostate Tumor	MIM	0.85	0.87	0.85	0.77
	JMI	0.80	0.73	0.75	0.82
	MRMR	0.90	0.89	0.79	0.89
Leukemia	MIM	0.73	0.82	0.84	0.74
	JMI	0.75	0.88	0.66	0.71
	MRMR	0.84	0.75	0.71	0.79
Lymphoma	MIM	0.75	0.78	0.84	0.77
	JMI	0.87	0.86	0.98	0.78
	MRMR	0.91	0.94	0.90	0.93

The results of our experiments initially demonstrated the performance of the classification algorithms, represented by the values of the evaluation metrics: Precision, Recall, and F1-score with feature selection (see Tables 7, 8, 9, and 10). These tables provide an evaluation of the influence of relevant feature selection, along with preprocessing, on the classification performance relative to the number of features retained.

The detailed evaluation of the models revealed that SVMs and neural networks provide excellent performance following relevant feature selection, while decision trees are less effective. The critical importance of the preprocessing and feature selection steps was underscored. This paves the way for the development of hybrid approaches and their potential application in oncology for personalized early cancer diagnosis.

Conclusion

In this study, our primary objective was to address the critical challenge of identifying the most informative and relevant genes for the accurate and reliable detection of various types of cancer. To achieve this, we developed a structured three-step process, with each step designed to evaluate the effectiveness of classification algorithms under different conditions.

Initially, we directly applied classification algorithms to the raw data without any prior manipulation. Subsequently, we enhanced the data quality through preprocessing, which included normalization, handling missing values, and reducing noise, before reapplying the same algorithms. Finally, we optimized the preprocessed data by selecting the most relevant genes using specific techniques before submitting them to the classification algorithms.

The results indicated that the SVM algorithm was the most effective, achieving a classification accuracy of 100% on the majority of the datasets following the application of feature selection techniques. NN also demonstrated promising performance. In contrast, the performance of DT and KNN was generally lower.

The proposed solution, which involves a systematic comparison of algorithms across various preprocessing and feature selection scenarios, offers several notable advantages. Firstly, it facilitates a comprehensive evaluation of the performance of different classification algorithms on genomic data, thereby providing a thorough insight into their respective strengths and weaknesses. Additionally, by incorporating preprocessing and feature selection steps, this approach enhances the quality of the input data by reducing noise and redundancy, thereby increasing the accuracy of the resulting models.

However, despite its numerous advantages, the proposed solution also has certain limitations to consider. One of the main constraints is the increased computational complexity resulting from the application of multiple algorithms and preprocessing techniques on large genomic datasets. Furthermore, although feature selection techniques such as MIM, MRMR, and JMI have proven effective, they may not always capture certain complex interactions between genes, necessitating the exploration of alternative or hybrid methods. Finally, since our study focused on specific datasets, the generalizability of our results to other contexts may be limited.

***Authors**: Dr. Mohammed Maiza, Faculty of Mathematics and Computer Science,University of Sciences and Technology of Oran, Algeria, Faculty of Exact and applied Sciences, Ahmed Benbella Oran 1 University, Algeria, email: mohammed.maiza@univusto.dz, maiza.mohammed@univ-oran1.dz, Prof. Samira Chouraqui, Faculty of Mathematics and Computer Science, University of Sciences and Technology of Oran, Algeria, email: s.chouraqui@yahoo.fr, Dr. Chahira Cherif, Faculty of Exact and applied Sciences, Ahmed Benbella Oran 1 University, Algeria, email: cherif.chahira@univ-oran1.dz, Prof. Abdelmalik Taleb-Ahmed, Institute of Electronics, Microelectronics and Nanotechnology (IEMN), Université Polytechnique des Hauts-de-France, Université de Lille, Valenciennes, France, email: Abdelmalik-Taleb.Ahmed@uphf.fr*

REFERENCES

[1] Torgovnick A., Schumacher B. DNA repair mechanisms in cancer development and therapy, 6(157),Front Genet,Apr 23, 2015.

[2] Moshood A. H., Tinuke O. O., Kayode S. A., Microarray cancer feature selection: Review, challenges and research directions, International Journal of Cognitive Computing in Engineering, Volume 1, pp 78-97, 2020.

[3] Brown J.S., Amend S.R., Austin R.H., Gatenby R.A., Hammarlund E.U., Pienta K.J., Updating the Definition of Cancer. Mol Cancer Res. 21(11),pp 1142-1147, Nov 1, 2023.

[4] Bansal M., Goyal A., Choudhary A., A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning, Decision Analytics Journal, Volume 3, 100071, 2022.

[5] Danades A., Pratama D., Anggraini D. and Anggriani D., Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status, 6th International Conference on System Engineering and Technology (ICSET), Bandung, Indonesia, pp 137-141, 2016.

[6] Fernàndez L. M., Simple but not simplistic: reducing the complexity of machine learning methods, Doctoral Thesis,2020.

[7] Amaar, A., Aljedaani, W., Rustam, F. et al. Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. Neural Process Lett 54, pp. 2219–2247, 2022.

[8] Ghorai S., Mukherjee A., Sengupta S. and Dutta P. K., Cancer Classification from Gene Expression Data by NPPC Ensemble, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 3, pp 659-671, May-June, 2011.

[9] Supplitt S., Karpinski P., Sasiadek M., Laczmanska I., Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. Int J Mol Sci. Jan 31;22(3):1422, 2021.

[10] Huang D.S., Gupta P., Wang L. , Gromiha M., Emerging Intelligent Computing Technology and Applications , 9th International Conference, ICIC 2013, Nanning, China, July 28-31, 2013.

[11] LATKOWSKI T.,OSOWSKI S., Feature selection methods in application to gene expression: autism data, PRZEGLĄD ELEKTROTECHNICZNY, pp 199-204, 2014.

[12] Li D. and Wang H., A Markov chain model-based method for cancer classification, 8th International Conference on Natural Computation, Chongqing, China, pp 1064-1068, 2012.

[13] Ganesh Kumar P. , Ammu V. and Victoire T. A. A., Building Decision Rules Using a Novel Data Driven Method for Microarray Data Classification, International Conference on Process Automation, Control and Computing, Coimbatore, India, pp 1-6, 2011.

[14] Alqahtani, A., Alsubai S., Sha M., Vilcekova L., Javed T., Cardiovascular Disease Detection using Ensemble Learning, Computational Intelligence and Neuroscience, 5267498, 9 pages, 2022.

[15] Sun X., Park J., Kang K. and Hur J., Novel hybrid CNN-SVM model for recognition of functional magnetic resonance images, IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, pp 1001-1006, 2017.

[16] Durgalakshmi B., Vijayakumar V., Feature selection and classification using support vector machine and decision tree. Computational Intelligence, 36, pp 1480–1492, 2020.

[17] Ijaz, M.F.; Alfian, G.; Syafrudin, M.; Rhee, J. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCANBased Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. Appl. Sci., 8, 1325, 2018.

[18] Maceika, A.; Bugajev, A.; Šostak, O.R.; Vilutien ˙ e, T. Decision Tree and AHP Methods Application for Projects Assessment: A Case Study. Sustainability, 13, 5502, 2021.

[19] Dwaraka Srihith, P. Vijaya Lakshmi, A. David Donald, T. Aditya Sai Srinivas, & G. Thippanna, A Forest of Possibilities: Decision Trees and Beyond. Journal of Advancement in Parallel Computing, 6(3), pp 29–37, 2023.

[20] Mahamdi, Yassine & Boubakeur, Ahmed & Mekhaldi, Abdelouahab & Benmahamed, Youcef, Power Transformer Fault

[21] Prediction using Naive Bayes and Decision tree based on Dissolved Gas Analysis, ENP Engineering Science Journal, 2, pp 1-5, 2022.

[22] Mustaqeem S., Anwar M.,Majid M. and Khan A.R. , Wrapper method for feature selection to classify cardiac arrhythmia, 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea (South), pp 3656-3659, 2017.

[23] M. Injadat, Optimized Machine Learning Models Towards Intelligent Systems, phdthesis,2020.

[24] Ünal H. & Ba¸sçiftçi F., Evolutionary design of neural network architectures: a review of three decades of research. Artificial Intelligence Review, 55, 2022.

[25] Grosan C. & Abraham A., Artificial Neural Networks, 17,pp 281-323, 2011.

[26] Arifin N. A.,Tiun S., Predicting Malay Prominent Syllable Using Support Vector Machine, Procedia Technology, Volume 11, pp 861-869, 2013.

[27] Nalepa, J., Kawulok, M., Selecting training sets for support vector machines: a review. Artif Intell Rev 52,pp 857–900, 2019.

[28] Anosh B. P. S., Annavarapu C. S. R., Dara S., Clusteringbased hybrid feature selection approach for high dimensional microarray data, Chemometrics and Intelligent Laboratory Systems, Volume 213,104305, 2021.

[29] B. Li, P. Zhang, S. Liang and G. Ren, Feature extraction and selection for fault diagnosis of gear using wavelet entropy and mutual information, 9th International Conference on Signal Processing, Beijing, China, pp 2846-2850, 2008.

[30] Sulaiman M. A. and Labadin J., Feature selection based on mutual information, 9th International Conference on IT in Asia (CITA), Sarawak, Malaysia, pp. 1-6, 2015.

[31] Jalali-Najafabadi F., Stadler M., Dand N., et al, Application of information theoretic feature selection and machine learning methods for the development of genetic risk prediction models, Sci Rep.11(1):23335, Dec 2, 2021.

[32] Khumukcham R., Urikhimbam B.C., Nazrul H., Dhruba K. B., JoMIC: A joint MI-based filter feature selection method, Journal of Computational Mathematics and Data Science, Volume 6, 100075, 2023.

[33] Jain, P.K., Jain, M. & Pamula, R. Explaining and predicting employees'attrition: a machine learning approach. SN Appl. Sci. 2, 757, 2020.

[34] Ginny Y. Wong, Frank H.F. Leung, Sai-Ho Ling, A hybrid evolutionary preprocessing method for imbalanced datasets, Information Sciences,Volumes 454–455,pp 161-177, 2018.

[35] Xinteng G., Xinggao L., A novel effective diagnosis model based on optimized least squares support machine for gene microarray, Applied Soft Computing, Volume 66, pp 50-59,2018.

[36] Alba E. , Garcia-Nieto J. ,Jourdan L. and Talbi E. -G., Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms,IEEE Congress on Evolutionary Computation, Singapore, pp. 284-290, 2007.

[37] Mylavarapu S. and Kaban A., Random projections versus random selection of features for classification of high dimensional data, 13th UK Workshop on Computational Intelligence (UKCI), Guildford, UK, pp 305-312, 2013.

[38] Cherif C., Abdi M.K., Maiza M.: Predictive approach to the degree of business process change, International Journal of Computing and Digital Systems, 14(1), pp. 10505-10513, Dec. 2023.

[39] Kou L. , Yuan Y., Sun J. and Lin Y., Prediction of Cancer Based on Mobile Cloud Computing and SVM, International Conference on Dependable Systems and Their Applications (DSA), Beijing, China, pp. 73-76, 2017.