Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Bayesian Feature Selection Method for Multiclass Classification with Application to Cancer Gene Expression Data

Maher Emarly

b00097022@alumni.aus.edu

American University of Sharjah

Ayman Alzaatreh

American University of Sharjah

Luai Al-Labadi

University of Toronto

Research Article

Keywords: Bayesian statistics, Relative belief ratio, Feature selection, Filter method, Kullback-Leibler divergence, Cancer classification, Microarray data, Gene expression analysis

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9106857/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Bayesian Feature Selection Method for Multiclass Classification with Application to Cancer Gene Expression Data

Maher Emarly[1*], Ayman Alzaatreh[1], Luai Al-Labadi[1,2]

[1]Department of Mathematics, American University of Sharjah, Sharjah, UAE
[2]Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Ontario, Canada
[*]Correspondence: Maher Emarly, (email: b00097022@alumni.aus.edu)

Abstract

Feature selection is essential for high-dimensional datasets, as it facilitates analysis and classification by finding the most informative features. This particularly important in gene expression data, where disease prediction and diagnosis are critical. Cancer microarray and sequencing datasets are characterized by high dimensionality and limited sample sizes, which pose significant challenges to large-scale data-driven classification. Since not all genes (features) contribute equally to predictive performance, effective feature selection is crucial in reducing dimensionality, improving model interpretability, and enhancing computational efficiency. In this study, we introduce a multiclass classification method that employs the relative belief ratio as a filtering criterion for high dimensional data. The proposed method generates a strength measure that serves as an importance score for each feature and enables systematic assessment of feature relevance across multiple classes. The resulting scores are used to rank features according to their discriminative power with respect to the target variable. The effectiveness of the proposed approach is evaluated using synthetic data and eight multiclass cancer gene expression datasets, and it is compared with several baseline filter-based feature selection methods. Four popular classifiers: logistic regression, random forest, K-nearest neighbors, and support vector machine were employed for classification. Experimental results demonstrate that the proposed approach is capable of identifying the relevant genes associated with various types of cancer (leukemia, colon tumor, lung cancer, breast cancer, lymphoma, and central nervous system tumor) and achieves competitive or superior classification accuracy while maintaining interpretability and low dimensional representations, highlighting its potential for biomedical research and large-scale machine learning and data analytics applications.

Keywords: Bayesian statistics; Relative belief ratio; Feature selection; Filter method; Kullback-Leibler divergence; Cancer classification; Microarray data; Gene expression analysis

1 Introduction

Classifying gene expression microarray datasets, often referred to as large-scale biological data analysis, has emerged as an important and intriguing research area for many scientists. The rapid advancement of microarray technology represents one of the most significant achievements in molecular biology, particularly for cancer detection (Alizadeh et al., 2000). Developing effective classification models is essential, as such models enable the accurate classification of previously unseen microarray samples once they have been trained on existing datasets. However, researchers encounter substantial challenges when detecting and classifying cancer using microarray gene expression data, primarily because these datasets typically contain a small number

of samples but a very large number of genes. To enhance classification performance, many genes must be removed through feature selection methods after being identified as redundant or irrelevant (noise features). Consequently, considerable effort has been devoted to developing improved feature selection techniques that enhance classification accuracy while reducing computational cost in cancer diagnosis and prognosis prediction (Alon et al., 1999a), (Armstrong et al., 2002), (Bhattacharjee et al., 2001).

Feature selection refers to the process of selecting a subset of relevant features (or variables) from a larger dataset to construct effective predictive models (Sobhanzadeh & Motieghader, 2019). In earlier studies, feature selection techniques primarily focused on identifying important genes prior to classifying microarray data for cancer diagnosis and prediction (Baldi & Long, 2001a), (Golub et al., 1999a), (Jafari & Azuaje, 2006a). Over time, numerous gene selection methods for microarray gene expression data have been developed to identify relevant genes associated with cancer diagnosis and prognosis (Alon et al., 1999a; Ben-Dor et al., 2000; Fox & Dimmic, 2006; Liu, Liu, & Zhang, 2010; Lu & Han, 2003; Ooi & Tan, 2003; Peng, Wu, & Jiang, 2010; Shen, Shi, & Kong, 2009).

A comprehensive review of feature selection techniques used in bioinformatics, along with their applications and potential in emerging areas, was provided in (Saeys, Inza, & Larranaga, 2007). Similarly, (Kamalov et al., 2025) reviewed several of the most recent feature selection approaches from a mathematical perspective. The threshold number of misclassification (TNoM) score was introduced in (Ben-Dor et al., 2000) as a method for gene selection. Subsequently, (Chandra & Gupta, 2011) proposed computing a p-value based on the TNoM score to evaluate the significance of each gene. Statistical significance tests, such as the ANOVA test and the t-test, have also been widely used as criteria for gene selection. In addition, several modified Bayesian approaches (Baldi & Long, 2001b; Fox & Dimmic, 2006; Zhou, Liu, & Wong, 2004) have been applied for gene selection, often relying on Gaussian distribution assumptions. Furthermore, clustering techniques have been used to reduce the dimensionality of gene expression data prior to classification (Chandra, Shanker, & Mishra, 2006), (Golub et al., 1999a).

Feature selection methods are generally categorized into three main types depending on whether and how they rely on a learning algorithm: filter, wrapper, and embedded methods. Filter methods evaluate the relevance of individual features to the target class labels independently of any learning algorithm. In contrast, wrapper and embedded methods incorporate the performance of the learning algorithm into the feature selection process. While embedded techniques guide the feature selection process based on the properties of the learning algorithm itself, wrapper methods typically evaluate multiple possible feature subsets by repeatedly training and testing the learning model (Pavya & Srinivasan, 2018).

In (Alzaatreh et al., 2025), an RBR-based filter method was introduced for binary classification problems. In this paper, the approach was extended to handle multiclass classification setting by utilizing the k-sample test proposed by (Al-Labadi & Alzaatreh, 2024). The proposed method assumes quantitative input features and a categorical target variable. One important application of this framework is cancer classification using gene expression data. To evaluate the effectiveness of the proposed method, several benchmark cancer datasets are used. The proposed

approach is compared with four well-established filter methods using both synthetic and real-world datasets. Experimental results indicate that the method improves classification performance while effectively identifying and ranking relevant features through the relative belief ratio (RBR) and its associated strength measure.

This work introduces a Bayesian feature selection approach based on the k-sample Relative Belief Ratio (RBR) for multiclass classification problems, with particular emphasis on high-dimensional gene expression data. The proposed method evaluates the statistical significance of features and ranks them according to their importance, providing a principled Bayesian measure of feature relevance. Such an approach is especially valuable for biological datasets, where identifying informative genes is essential for improving classification performance and supporting cancer diagnosis and prognosis studies.

Classifying high-dimensional datasets, particularly gene expression data, is challenging due to the small sample sizes and the presence of a large number of redundant or irrelevant features that can reduce classification accuracy. Standard feature selection methods such as Symmetrical Uncertainty (SU), Information Gain (IG), Minimum Redundancy Maximum Relevance (mRMR), and ReliefF have been widely used to address dimensionality issues. However, these approaches may struggle to balance dimensionality reduction with maintaining high classification accuracy and may not effectively capture feature importance in multiclass classification settings. To address these limitations, we extend the RBR-based filter method previously proposed for binary classification to the multiclass classification setting using the k-sample test framework. The proposed approach is evaluated on several benchmark cancer gene expression microarray datasets and compared with widely used feature selection methods, including IG, SU, mRMR, and ReliefF. Experimental results demonstrate that the proposed method achieves competitive or superior classification performance while effectively identifying and ranking relevant features.

The remainder of the paper is organized as follows. Section 2 provides an overview of filter-based feature selection techniques commonly used in gene expression analysis and data mining and discusses several widely used methods considered for comparison with the relative belief ratio (RBR) approach. Section 3 describes the proposed RBR filter-based feature selection method. Section 4 presents the gene expression datasets used in this study. Section 5 provides a brief overview of the classification algorithms employed in the analysis. Section 6 presents a comparative evaluation of the proposed method against widely used filter-based gene selection techniques using eight gene expression microarray cancer datasets, along with the corresponding results and discussion. Finally, Section 7 concludes the paper and outlines possible directions for future work.

2 Overview of Filter Based Feature Selection Methods

Filter methods perform feature selection based solely on the statistical properties of the training data, without relying on any learning algorithm. In these approaches, feature selection is conducted as a preprocessing step prior to model training. A typical filter algorithm involves two main stages. First, features are ranked according to specific criteria, which may vary across different methods. Then, the highest-ranked features are selected and used as inputs to the learning algorithm. Several

statistical techniques have been widely used in filter-based feature selection, including ANOVA, variance-based selection, correlation-based selection, and information gain. These methods aim to identify features that contribute significantly to explaining the variation in the target variable (Blessie & Karthikeyan, 2012). For instance, ANOVA-based feature selection evaluates differences in feature means across distinct classes or groups, using F-statistics and corresponding p-values to determine the significance of each feature (Muller & Fetterman, 2003). Variance-based selection, on the other hand, retains features that exhibit substantial variability across the dataset, under the assumption that features with higher variance are more informative for modeling and prediction. In this approach, the variance of each feature is computed and features are ranked accordingly, allowing those with greater variability to be prioritized and potentially improving predictive performance (Munson & Caruana, 2009).

To demonstrate the effectiveness of the proposed strategy, it is compared with several widely used gene selection techniques. Brief descriptions of the methods used for comparison are provided below.

Information Gain (IG)

Information Gain is a commonly used filter method for feature selection due to its simplicity and computational efficiency (Roobaert, Karakoulas, & Chawla, 2006). It measures how much knowing one variable, say A, reduces the uncertainty about another variable B, based on the concept of entropy.

The entropy of a discrete random variable B is calculated as

$$H(B) = -\sum_{b} P(b) \log_2 P(b), \tag{1}$$

where $P(b)$ is the probability that a sample belongs to class $b \in B$. The conditional entropy of B given A, denoted $H(B \mid A)$, measures the remaining uncertainty in B after observing A:

$$H(B \mid A = a) = -\sum_{b} \frac{P(a,b)}{P(a)} \log_2 \left(\frac{P(a,b)}{P(a)} \right), \tag{2}$$

where $P(a, b)$ is the joint probability of a sample being in class b and having feature value a. For nominal variables, the total conditional entropy $H(B \mid A)$ is obtained by averaging over all values of A. The Information Gain is then defined as

$$IG(B; A) = H(B) - H(B \mid A). \tag{3}$$

For continuous variables, differential entropy is used, replacing summations with integrals. IG is effective for identifying feature relevance and understanding variable relationships. However, it tends to favor categorical features with many distinct values, even if they offer limited predictive power.

Symmetric Uncertainty (SU)

Symmetric Uncertainty (SU) quantifies the degree of association between two variables and is often applied to evaluate the relationship between a feature and the target variable in a dataset. Like IG, SU utilizes entropy to measure the information contained within a variable. To overcome the inherent bias of Information Gain (IG) toward attributes with many distinct values, Symmetric Uncertainty (SU) introduces a normalization based on the combined entropy of A and B. The SU metric is formally given by

$$SU(A, B) = \frac{2 \times IG(A; B)}{H(A) + H(B)}, \tag{4}$$

where $H(A)$ and $H(B)$ denote the entropies of variables A and B, respectively, and $IG(A; B)$ refers to the Information Gain as previously defined. SU values are constrained to the interval $[0, 1]$, with a value of 1 indicating complete predictive power of one variable over the other, and a value of 0 suggesting no association. While SU helps address IG's bias, it tends to favor features with lower cardinality (Chai, Z. , Li, W., 2023).

Minimum Redundancy Maximum Relevance (mRMR)
mRMR is a multivariate filter that selects features maximizing relevance to the target via mutual information or correlation while minimizing redundancy among themselves. It greedily builds a subset that balances these two objectives, yielding compact, informative feature sets (Bugata, P., 2020).

ReliefF
ReliefF is an instance based weighting algorithm: for each randomly chosen sample, it finds the nearest "hits" (same class) and "misses" (different class) and updates feature weights based on how well features distinguish between hits and misses. It captures feature interactions and is robust to noise but does not explicitly remove redundancy (Urbanowicz, RJ. , Meeker, M. , La Cava, W. , Olson, RS., 2018).

3 Multiclass Feature Selection Using RBR

This section provides a detailed explanation of the proposed feature selection approach.

3.1 RBR Feature Selection Method

Let $\mathbb{S} \equiv (\boldsymbol{X}, Y)$ denote a dataset consisting of n samples and p features, where $n, p \in \mathbb{Z}_{>0}$. The matrix $\boldsymbol{X} = [x_{ij}]_{n \times p} \in \mathbb{R}^{n \times p}$ contains the feature values, and Y denotes the categorical target variable comprising m distinct classes, labeled from 1 to m. Additionally, x_{ij} denotes the i-th observation of the j-th feature. Denote the set of features (variable) in $\boldsymbol{X}$ by $\{V_1, \ldots, V_p\}$. Each feature $\{V_i\}$, assumed to be continuous, is divided into m subsets according to the number of classes $\{\mathbb{V}_{i,1}, \ldots, \mathbb{V}_{i,m}\}$ which includes the observations where the target variable is $1, 2, ..., m$ respectively.

We assume that $\mathbb{V}_{i,1}, \ldots, \mathbb{V}_{i,m}$ are m independent variables follow the normal distributions $N(\mu_1, \sigma_1^2)$, $\ldots$, $N(\mu_m, \sigma_m^2)$, respectively, where $(\mu_i, \sigma_i^2) \in \mathbb{R} \times \mathbb{R}^+$. Now,

consider the hypotheses:

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 = \cdots = \mu_m \\ H_a : &\quad \mu_i \neq \mu_j \text{ for some } i, j = 1, \ldots, m. \end{aligned} \tag{5}$$

(Al-Labadi & Alzaatreh, 2024) employ the relative belief ratio introduced by Evans (Evans, 2016) to infer inferences on an arbitrary parameter θ within the parameter space Θ, utilizing a statistical model characterized by the density function $f_\theta(x)$.

Let Π_θ denote the prior distribution of the parameter θ, with associated density function π_θ. Upon observing data x, the corresponding posterior distribution is denoted by $\Pi_\theta(\cdot \mid x)$, with density expressed as

$$\pi(\theta \mid x) = \frac{f_\theta(x)\pi(\theta)}{m(x)}, \tag{6}$$

where

$$m(x) = \int f_\theta(x)\pi(\theta)\, \mathrm{d}\theta$$

is the marginal likelihood, also known as the prior predictive density of the observed data. For a hypothesized value θ_0 of θ, and assuming that the densities π_θ and $\pi_\theta(\cdot \mid x)$ are continuous at θ, the relative belief ratio (RBR) given by

$$RB_\theta\left(\theta_0 \mid x\right) = \frac{\pi_\theta\left(\theta_0 \mid x\right)}{\pi_\theta\left(\theta_0\right)}. \tag{7}$$

The Relative Belief Ratio (RBR) is a robust tool for hypothesis testing and for making informed decisions based on available evidence (see, for example, the work of (Abdelrazeq, Al-Labadi, & Alzaatreh, 2020) and (Al-Labadi & Evans, 2017). The RBR compares the belief in a hypothesis from the prior state to the posterior state, providing a quantitative measure of how evidence has affected the level of belief in the hypothesis. The proposed method can be described as follows: for a given feature $\{V_j\}$ if H_a is accepted (i.e., $RB_\theta(\theta_0 \mid x) < 1$), it can be concluded that the feature used is significant for discriminating between the classes $(1, \ldots, m)$ of the target variable. Conversely, if H_0 is accepted (i.e., $RB_\theta(\theta_0 \mid x) > 1$), it can be concluded that the feature used is insignificant and can thus be eliminated from further analysis. The construction of RBR is described below.

In order to construct the RBR for the hypotheses in (5), we need to construct priors for μ_i and σ_i^2, for $i = 1, 2, \ldots, k$. We consider the following commonly used conjugate priors, (Al-Labadi & Alzaatreh, 2024).

$$\frac{1}{\sigma_i^2} \sim \text{gamma}_{\text{rate}}(\alpha_{0i}, \beta_{0i}) \tag{8}$$

$$\mu_i \mid \sigma_i^2 \sim \mathcal{N}(\mu_{0i}, \lambda_{0i}^2 \sigma_i^2), \tag{9}$$

where μ_{0i}, λ_{0i} and $(\alpha_{0i}, \beta_{0i})$ are hyper-parameters to be specified via elicitation. Full details of the hyperparameter choices can be found in (Al-Labadi & Alzaatreh, 2024).

The posterior distribution of (μ_i, σ_i^2) for $i = 1, 2, \ldots, k$ is given by

$$\frac{1}{\sigma_i^2} \mid x_{i1}, \ldots, x_{in_i} \sim \text{gamma}_{\text{rate}} \left(\alpha_{0i} + \frac{n_i}{2}, \beta_{x_i} \right) \tag{10}$$

$$\mu_i \mid \sigma_i^2, x_{i1}, \ldots, x_{in_i} \sim \mathcal{N} \left(\mu_{x_i}, \sigma_{x_i}^2 \right), \tag{11}$$

where

$$\mu_{x_i} = \left(n_i + \frac{1}{\lambda_{0i}^2} \right)^{-1} \left(\frac{\mu_{0i}}{\lambda_{0i}^2} + n_i \overline{x_i} \right), \quad \sigma_{x_i}^2 = \left(n_i + \frac{1}{\lambda_{0i}^2} \right)^{-1} \sigma_i^2$$

and

$$\beta_{x_i} = \beta_{0i} + \frac{(n_i - 1) S_i^2}{2} + \frac{n \left(\bar{x}_i - \mu_{0i} \right)^2}{2 \left(n_i \lambda_{0i}^2 + 1 \right)},$$

with $\overline{x_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$and $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i} \right)^2$.

The hyperparameters $\mu_{0i}, \lambda_{0i}, \alpha_{0i}, \beta_{0i}$ were selected using a fixed elicitation scheme for all features. Specifically, $\mu_{0i} = 1$ was used, and the standard deviation bounds were set to $s_{1i}^2 = 1$ and $s_{2i}^2 = 4$. For the inverse gamma, the shape and rate parameters were fixed at $\alpha_{0i} = 2$ and $\beta_{0i} = 2$.

In order to test the hypothesis, the Kullback-Leibler (KL) distance is used to compute the distance between the reference distribution and the prior/posterior distributions (Al-Labadi & Alzaatreh, 2024). It can be shown that when μ_i and σ_i have the prior distributions defined in (5) and (6), the prior distance is given by

$$D = d^{\text{prior}} (P, Q) = \sum_{i=1}^{k} \frac{\left(\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}} \sum_{i=1}^{k} \frac{\mu_i}{\sigma_i^2} - \mu_i \right)^2}{2\sigma_i^2}. \tag{12}$$

Also, if μ_i and σ_i have the posterior distributions defined in (7) and (8), the posterior distance is given by

$$D_x = d^{\text{post}} (P, Q) = \sum_{i=1}^{k} \frac{\left(\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}} \sum_{i=1}^{k} \frac{\mu_i}{\sigma_i^2} - \mu_i \right)^2}{2\sigma_i^2}. \tag{13}$$

The change in the distance between the prior and posterior distributions can be seen by the value of the relative belief ratio (Al-Labadi & Alzaatreh, 2024).

The strength of evidence provided by the relative belief ratio must be assessed to determine whether it there is a strong or weak evidence that supports or contradicts the proposed null hypothesis. To quantify this, the tail probability denoted as $\text{Str}(\theta_0 \mid x)$ is calculated, following the methodology of (M. Evans, 2015):

$$\text{Str}(\theta_0 \mid x) = \Pi \left(RB(\theta \mid x) \leq RB(\theta_0 \mid x) \mid x \right) = \int_{RB(\theta|x) \leq RB(\theta_0|x)} \pi(\theta \mid x) d\theta. \tag{14}$$

Here, A low value of $\text{Str}(\theta_0 \mid x)$, when $RB(\theta_0 \mid x) < 1$, shows strong evidence against θ_0, whereas a high value of $\text{Str}(\theta_0 \mid x)$, when $RB(\theta_0 \mid x) > 1$, indicates strong evidence in favor of θ_0. This calibration enhances the credibility of the conclusions by offering a more nuanced understanding of the evidence.

3.2 Computations of The RBR and Its Strength

As described in (Al-Labadi & Alzaatreh, 2024), let $V = (x_1, \ldots, x_k)$, M be a positive integer, and let $\hat{F}_D$ denote the empirical cdf of D for $i = 0, \ldots, M$. Let $\hat{d}_{i/M}$ be the estimate of $d_{i/M}$, the (i/M)-th prior quantile of D. Here $\hat{d}_0 = 0$ and $\hat{d}_1$ is the largest value of d. Let $\hat{F}_D(. \mid x)$ denote the empirical cdf of D. For $d \in \left[\hat{d}_{i/M}, \hat{d}_{(i+1)/M}\right]$, estimate $RB_D(d \mid x) = \pi_D(d \mid x)/\pi_D(d)$ by

$$R\hat{B}_D(d \mid x) = M\left\{\hat{F}_D\left(\hat{d}_{(i+1)/M)} \mid x\right) - \hat{F}_D\left(\hat{d}_{i/M} \mid x\right)\right\}. \tag{15}$$

We estimate $RB_D(0 \mid x) = \pi_D(0 \mid x)/\pi_D(0)$ by $R\hat{B}_D(0 \mid x) = M\hat{F}_D\left(\hat{d}_{p_0} \mid x\right)$ where $p_0 = i_0/M$ and i_0 are chosen so that i_0/M is not too small (roughly 0.05).

Lastly, we can estimate the Strength $DP_D\left(RB_D(d \mid x) \leq RB_D(0 \mid x) \mid x\right)$ by the finite sum:

$$\sum_{i \geq i_0 : \hat{RB}_D\left(\hat{d}_{i/M} \mid x\right) \leq \hat{RB}_D(0 \mid x)} \left(\hat{F}_D\left(\hat{d}_{(i+1)/M} \mid x\right) - \hat{F}_D\left(\hat{d}_{i/M} \mid x\right)\right). \tag{16}$$

The strength of the relative belief ratio (RBR) generates an importance score $\ell_j \in \mathbb{R}$ for each feature V_j were $j \in \{1, \ldots, p\}$. The numerical value of the strength of the RBR is used to rank the features, with a low value indicating stronger a higher importance score. Based on the RBR strength scores, the feature set will therefore be arranged from the most to the least important features. The RBR and feature selection algorithms are outlined in Algorithms I and II respectively.

Algorithm 1: RBR Algorithm

1: For each $i = 1, \ldots, k$, specify the hyperparameters μ_{0i}, λ_{0i} and $(\alpha_{0i}, \beta_{0i})$ following the procedures outlined in (8) and (9).
2: For each $i = 1, \ldots, k$, draw samples of μ_i and σ_i^2 according to the distributions given in (10) and (11).
3: Evaluate the quantity D as defined in (12).
4: Repeat steps 1–3 to obtain r_1 replicated values of D.
5: Again, generate μ_i and σ_i^2 using (10) and (11).
6: Compute D_x using the formula presented in (13).
7: Repeat steps 5 and 6 to generate r_2 replicated samples of D_x.
8: Calculate the relative belief ratio as outlined in (15).
9: Estimate the corresponding strength measure using the expression in (16).

Algorithm 2: Multiclass Feature Algorithm

For $i = 1, \ldots, p$

1: Take Take $\{\mathbb{V}_{i,1}, \ldots, \mathbb{V}_{i,m}\}$ which includes the observations where the target variable is 1,2,...,m respectively of the training data set $\mathbb{X}$.

2: Apply RBR Algorithm for $\mathbb{V}_{i,1}, \ldots, \mathbb{V}_{i,m}$.

3: Obtain the importance score ℓ_i (the Strength of RBR) .

End

4: Sort $\ell_1, \ldots, \ell_p$ in ascending order.

5: Choose the top features to be retained from the sorted list of ℓ_i's.

4 Gene Datasets Description

This section provides a description of the utilized gene expression datasets. Each dataset is described in terms of its sample composition, features, and classification tasks, highlighting their relevance to cancer research and diagnosis.

Alon (Colon Cancer) The dataset used in this study comprises 62 samples and originates from the work of (Alon et al., 1999a). It contains gene expression measurements for 6,500 human genes, obtained using an Affymetrix oligonucleotide array. Among the samples, 40 are tumor tissues and 22 are normal colon tissues, making the dataset relevant for colon cancer classification. For the purpose of analysis, a subset of 2,000 genes was selected based on the highest minimum intensity observed across all samples. The dataset is labeled into two categories: tumor and normal.

Golub (Leukemia) The dataset introduced by Golub et al. (1999) comprises 72 samples, including 47 from individuals diagnosed with acute lymphoblastic leukemia (ALL) and 25 from those with acute myeloid leukemia (AML) (Golub et al., 1999a). Bone marrow specimens were collected at the point of diagnosis, and gene expression profiling was performed using Affymetrix Hgu6800 microarrays. The dataset includes 7,129 gene expression features corresponding to Affymetrix probe sets. Since ALL originates from two different lymphocyte lineages T-cell and B-cell the dataset has been analyzed using a three class classification scheme: AML, ALL-T, and ALL-B.

Gravier (Breast Cancer) Gravier conducted an analysis of small invasive ductal carcinomas classified as T1T2N0—indicating no axillary lymph node involvement with the aim of predicting metastatic potential in node negative breast cancer cases. A total of 168 patient samples were evaluated over a five year period. Gene expression data encompassing 2,905 features were utilized, categorized into two classes: A total of 111 patients who did not experience any adverse events following diagnosis were categorized as having a "good" prognosis, while 57 patients who developed early metastases were classified as "poor" prognosis cases (Gravier, Eleonore et al., 2010).

Khan (SRBCT): The dataset introduced by (Khan et al., 2001) addresses the classification of small round blue cell tumors (SRBCTs) in pediatric patients, encompassing neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing sarcoma family of tumors (EWS). It comprises 63 training samples generated using cDNA microarray technology, initially measuring the expression of 6,567 genes. For analysis, the data was filtered to retain 2,308 genes relevant to the classification of four class labels corresponding to the four SRBCT categories.

Pomeroy (CNS Tumor): The dataset centers on embryonal tumors of the central nervous system (CNS), with particular emphasis on medulloblastomas, the most prevalent malignant brain tumors in pediatric patients. It comprises 60 samples, each described by 7,128 features representing gene expression levels. The dataset is categorized into two classes and specifically targets Central Nervous System Embryonal Tumors (Pomeroy et al., 2002).

Shipp (Lymphoma): This dataset, originally presented by (Shipp et al., 2002), investigates diffuse large B-cell lymphomas (DLBCL), which constitute approximately 30–40% of adult non Hodgkin lymphoma cases. It includes 58 patient samples, each profiled with 6,817 gene expression measurements. The samples are divided into two groups: 32 correspond to patients who responded successfully to CHOP based chemotherapy, while the remaining 26 are associated with fatal or treatment resistant disease.

Sorlie (Breast Cancer): This dataset, based on the study by (Sørlie et al., 2001), focuses on breast carcinoma and utilizes gene expression profiles obtained through cDNA microarray technology. It consists of 85 samples, each characterized by 456 features corresponding to the expression levels of 427 distinct genes. The dataset encompasses five categories: 78 carcinoma samples, 3 benign tumors, and 4 normal tissue samples.

Yeoh (Leukemia): The Yeoh dataset based on (Yeoh et al., 2002) comprises diagnostic bone marrow gene expression profiles from 248 children, each assigned to one of six prognostic subtypes: T-cell lineage ALL (43 cases), E2A-PBX1 (27), TEL-AML1 (79), BCR-ABL (15), MLL rearrangements (20), and hyperdiploid karyotypes with more than 50 chromosomes (64). Gene expression was measured using Affymetrix oligonucleotide microarrays interrogating 12,600 probe sets. In the present work, a random subset of 1,000 features was drawn from the original 12,625 probe sets, while the full cohort of 248 pediatric ALL samples was retained unchanged.

5 Classification methods

This section presents an overview of the main classification algorithms utilized in the data analysis: Logistic Regression, k-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM). Each algorithm is introduced with its mathematical foundation, core concepts, and applicability to the classification tasks relevant to the datasets examined in this study.

Logistic Regression

Logistic regression is a popular technique used for binary classification, where the objective is to model the probability that a given observation belongs to a specific class. The method utilizes the logistic (sigmoid) function to convert a linear combination of input features into a probability value constrained between 0 and 1. The conditional probability $P(y = 1 \mid x)$ is estimated as:

$$P(y = 1 \mid x) = \sigma(\boldsymbol{\theta}^T x + \theta_0), \tag{17}$$

where $\sigma(z)$ denotes the sigmoid function, defined by:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{18}$$

In this formulation, $\boldsymbol{\theta}$ is the coefficient vector, x represents the input feature vector, and θ_0 is the intercept (bias) term. The resulting probability is typically thresholded at 0.5 to assign a binary class label to each observation.

The parameters $\boldsymbol{\theta}$ and θ_0 are estimated by maximizing the likelihood of the observed data, which is equivalent to minimizing the negative log-likelihood. The corresponding loss function, $\mathcal{L}(\boldsymbol{\theta}, \theta_0)$, is given by:

$$\mathcal{L}(\boldsymbol{\theta}, \theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{19}$$

where $\hat{y}_i = P(y = 1 \mid x_i)$ is the predicted probability for the i-th observation, y_i is the true label, and n is the total number of observations. Optimization algorithms such as stochastic gradient descent are typically employed to minimize the loss and estimate the model parameters (Boateng & Abaye, 2019).

k-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is a non-parametric, observation-based learning technique suitable for both classification and regression tasks. In classification, a sample is labeled according to the most common class among its k nearest neighbors. For regression, the predicted value is computed as the average of the target values of the k closest neighbors (Sun & Huang, 2010).

For a given point x_q, the distance to all training points $\{x_1, x_2, \ldots, x_n\}$ is computed. The k nearest neighbors are determined using a specified distance metric, commonly the Euclidean distance, which is defined as:

$$d(x_q, x_i) = \sqrt{\sum_{j=1}^{m} (x_{qj} - x_{ij})^2}, \tag{20}$$

where x_q and x_i are m-dimensional vectors representing the point and the i-th training sample, respectively.

For classification tasks, the predicted class label $\hat{y}_q$ for the point x_q is determined using majority voting among the class labels of the k nearest neighbors:

$$\hat{y}_q = \arg\max_{c} \sum_{i \in \mathcal{N}_k(x_q)} \mathbb{I}(y_i = c), \tag{21}$$

where $\mathcal{N}_k(x_q)$ represents the indices of the k nearest neighbors of x_q, $\mathbb{I}(\cdot)$ is the indicator function, and y_i is the class label of the i-th neighbor.

Random Forest

Random Forest is an ensemble learning algorithm applicable to both classification and regression tasks. It constructs multiple decision trees during training and

combines their individual predictions to improve accuracy and reduce overfitting. This method incorporates bagging (Bootstrap Aggregating) and random feature selection to generate a diverse set of decision trees (Rigatti & Steven, 2017).

For each tree in the ensemble, a bootstrap sample is generated by randomly drawing instances from the original training set $\mathcal{S} = \{(\boldsymbol{z}_1, t_1), (\boldsymbol{z}_2, t_2), \ldots, (\boldsymbol{z}_m, t_m)\}$ with replacement. A new training subset $\mathcal{S}_b$ of size m is created for each tree using this sampling procedure. Each decision tree is then trained on a different $\mathcal{S}_b$.

At each decision node, a random subset of input features is selected, and the best split is identified based on a chosen criterion. For classification problems, the Gini impurity is commonly used and is defined as:

$$\mathcal{G} = \sum_{j=1}^{K} q_j(1 - q_j), \tag{22}$$

where q_j is the proportion of data points at the node that belong to class j, and K is the total number of classes.

For classification, each tree in the forest provides a predicted label $\hat{t}^{(r)}$ for an input $\boldsymbol{z}$, and the final prediction $\hat{t}$ is obtained through majority voting:

$$\hat{t} = \arg\max_j \sum_{r=1}^{R} \mathbb{I}(\hat{t}^{(r)} = j), \tag{23}$$

where R is the total number of trees in the forest, j represents a class label, and $\mathbb{I}(\cdot)$ is the indicator function.

Random Forest reduces overfitting by aggregating the predictions of multiple trees, making it more stable and less sensitive to noise. It is also capable of handling high dimensional data, supports both classification and regression tasks, and can provide feature importance scores based on the average reduction in impurity contributed by each feature.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method designed to perform classification by identifying the optimal hyperplane that best separates data points from different classes in the feature space. This hyperplane maximizes the margin between the classes and is determined by minimizing the following objective function $\Lambda(\boldsymbol{\alpha})$:

$$\Lambda(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + C\sum \delta_i \tag{24}$$

subject to the constraints:

$$t_i\left[\boldsymbol{\alpha}^T K(\boldsymbol{z}_i) + \beta\right] \geq 1 - \delta_i, \quad \delta_i \geq 0, \quad i = 1, \ldots, n. \tag{25}$$

In this formulation, $\boldsymbol{\alpha}$ represents the weight vector, β is the bias term, and δ_i are slack variables introduced to handle misclassified or ambiguous training samples. Each $\boldsymbol{z}_i$

denotes the feature vector for the i-th training observation, while t_i corresponds to its class label. The kernel function K is employed to project the input data into a higher-dimensional space, enabling the construction of a nonlinear decision boundary. This study applies a linear kernel. The regularization parameter C controls the trade-off between maximizing the margin and minimizing classification error; higher values of C place greater emphasis on minimizing misclassification.

SVMs are inherently binary classifiers. To address multiclass problems, one can decompose the task into multiple binary classification problems using strategies such as one-vs-one or one-vs-all, or alternatively, adopt multiclass extensions of the SVM framework (Campbell & Sturim, 2006).

6 Experiments, Analysis and Results

The RBR feature selection approach requires numeric features and a qualitative target. This section illustrates the performance of RBR feature selection applied to two types of datasets: synthetic and gene datasets.

6.1 Synthetic Datasets

The RBR feature selection method is applied on synthetic datasets characterized by relevant and irrelevant features. This approach enables the assessment of the method ability to identify relevant and irrelevant features. Synthetic datasets are generated using the Friedman function, a nonlinear function utilized for testing and validating statistical models and machine learning algorithms due to its inherent complexity and controllable noise introduction. The function is defined as

$$b = 10\sin(\pi a_1 a_2) + 20(a_3 - 0.5)^2 + 10a_4 + 5a_5 + \epsilon, \tag{26}$$

where a_1, a_2, a_3, a_4, a_5 represent independent variables uniformly distributed on [0, 1] and $\epsilon \sim N(0, 1)$ (Friedman, 1991).

1000 synthetic datasets were generated using the Friedman function. Each dataset was designed to contain 10 features, with 5 designated as important features and the remaining 5 as noise, including target variables with 3 classes. The RBR feature selection method was then applied, and the RBR and strength values of each feature were averaged across all datasets.

The results, as shown in Table 1, indicate that RBR feature selection method was able to correctly identify 4 out of 5 important features correctly and failed to identify important_3 feature as a relevant one with weak evidence (smaller strength means stronger evidence in the significance case). Furthermore, all noisy variables were correctly identified.

Table 1: Multiclass synthetic Results

Feature	RBR	Strength
important_1	0.00000	0.000000
important_2	0.00000	0.000000
important_3	11.76704	0.785142
important_4	0.00000	0.000000
important_5	0.04752	0.003360
noise_1	8.73552	0.766002
noise_2	9.44144	0.749796
noise_3	9.53912	0.789618
noise_4	9.11628	0.736568
noise_5	9.48844	0.786382

6.2 Gene datasets experimental design

The experimental analysis in this section centers on multiclass classification using the datasets outlined in Table 2. Each class represents a category assigned to observations based on the nominal target variable Y. The datasets consist of numerical features with varying dimensionalities and sample sizes. The feature count ranges from 456 to 7,129, whereas the number of samples varies from 58 to 248. The target variable includes between 2 and 5 distinct classes. Additionally, the dimensionality ratio varies from 5.365 to 118.8, while the class imbalance ranges from 0.19 to 0.55 across the datasets.

All datasets used in the evaluation are gene microarray data, capturing the expression levels of thousands of genes at once. All these datasets are used in machine learning to identify and classify genes associated with cancer diseases. In high dimensional datasets, such as those in the table, many features may be redundant or irrelevant. The most informative genes were identified, facilitating more accurate and robust disease classification and contributing to a better understanding and diagnosis of various cancers and other genetic conditions.

Table 2: Multiclass Gene Datasets Descriptions.

Dataset Name	**No. Observations**	**No. Features**	**No. Classes**	**Dimensionality**	**Imbalance**
Alon (Colon Cancer)	62	2000	2	32.258	0.55
Golub (Leukemia)	72	7129	3	99.014	0.531
Gravier (Breast Cancer)	168	2905	2	17.292	0.513513
Khan (SRBCT)	63	2308	4	36.635	0.347
Pomeroy (CNS Tumor)	60	7128	2	118.800	0.538
Shipp (Lymphoma)	58	6817	2	117.534	0.327
Sorlie (Breast Cancer)	85	456	5	5.365	0.343
Yeoh (Leukemia)	248	1000	6	4.032	0.190

*The data sources: R package datamicroarray. Dimensionality is defined as the ratio of the number of features to the number of observations. Class imbalance is measured as the ratio of observations in the minority class to those in the majority class.

To assess the effectiveness of the proposed approach, the following was applied:
Let $\{V_i \mid i = 1, 2, \ldots, p\}$ represent the set of features in the feature space $\boldsymbol{X}$. Each dataset $\mathbb{S}$ was split into training and testing subsets, with 50% of the data used for

feature ranking and the remaining 50% for evaluation. The performance of the RBR method was compared against four baselines Information Gain (IG), Symmetrical Uncertainty (SU), Minimum Redundancy Maximum Relevance (mRMR), and ReliefF. Each feature selection method assigns an importance score $\ell_j \in \mathbb{R}$ to each feature V_i. Features were then ranked in descending order based on these scores from the most to the least important, and the top 100 features were retained.

Next, a threshold parameter k was used to generate ten different feature subset sizes, corresponding to the top $\{10, 20, \ldots, 100\}$ features from the top 100 selected previously. When $k = 100$, the full set of top 100 features was retained, it served as the baseline model.

A five fold cross-validation approach was used in the testing phase of the experimental evaluation. In each fold, the test set was used to assess the performance of the selected feature subsets across four classification algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), (introduced in Section 5). To mitigate overfitting in this high dimensional, low sample regime ($n \in \{58\text{–}248\}$, $p \gg n$), all hyperparameters were optimized via nested cross-validation within each training fold: the elastic net mixing parameter $\alpha \in \{0, 0.5, 1\}$ and penalty strength λ (50 logarithmically spaced values) for logistic regression;the number of neighbors $k \in \{1, 3, 5, 7, 9\}$ with per-fold standardization for KNN; the number of variables per split and node size for random forest (500 trees, mtry $\in \{\sqrt{p}, p/10, p/3\}$, node size = 1); and the cost $C \in \{0.1, 1, 10, 100\}$ and kernel width $\gamma \in \{1/p, 0.1, 1, 10\}$ for SVM with a radial basis kernel (with predictors standardized per fold). Following hyperparameter selection, each model was refit on the full training partition and evaluated on its held-out test split, and the resulting F_1 scores were averaged across the five outer folds to yield the final performance matrices.

For each dataset classifier pair, the proposed RBR filter was compared to the four baselines features selections methods at all ten feature subset thresholds. Statistical significance was assessed via two-sided Wilcoxon signed rank tests on the paired micro-F_1scores, with $p < 0.05$ indicating significant differences. To complement the p-values with a practical effect measure, we calculated the paired gap

$$d_i = \text{Rb_St}_i - \text{Baseline}_i, \qquad i = 1, \ldots, 10,$$

and reported the median of the ten d_i values. This median gives a robust, distribution free estimate of the typical micro-F_1 gain (positive value) or loss (negative value) achieved by RBR whenever the Wilcoxon test indicates significance.

6.2.1 Gene Datasets Results

In this section, we start with an overview of all dataset results, followed by a detailed discussion of three datasets where RBR shows good, moderate, and poor performance compared with the benchmark feature selection methods. The complete set of result figures and tables for all datasets is provided in the Appendix.

The micro-F_1 score results for each dataset, obtained after applying the Relative Belief Ratio (RBR) and four filter methods, Information Gain (IG), Symmetrical Uncertainty (SU), Minimum Redundancy Maximum Relevance (mRMR), and ReliefF, are shown in the figures. Next, the micro-F_1 scores, averaged over five fold cross

validation, are plotted against ten different k thresholds to illustrate classification performance across varying feature subset sizes during testing.

The results obtained from RBR and the other baseline methods are then analyzed and compared in terms of overall micro F_1 and the ability to identify optimal subsets of informative features. For the majority of datasets, applying feature selection techniques led to improved classification performance compared to the model that utilized the full set of top $k = 100$ features.

The RBR method demonstrated consistent effectiveness across all classifiers, with accuracy showing a more pronounced improvement in some datasets compared to alternative baseline methods than others. For instance, the SVM accuracy rises from an initial value of 0.675 to approximately 0.74 on the Gravier dataset (Figure 8). Starting from the baseline with the top 100 features, RBR achieved better logistic regression starting accuracy in the Alon, Golub, and Shipp datasets (Figures 3, 7, 11), while it has the lowest in Khan and Yeoh (Figures 9, 5). The starting performance was random with no noticeable trend for the other dataset classifiers.

Interestingly, RBR reached the peak score in logistic regression on Alon, Khan, Shipp, and Sorlie (Figures 3, 9, 11, 12), suggesting that the method was able to eliminate noisy features, while it decreased to the lowest in Khan, Pomeroy, and Yeoh (Figures 9, 10, 5), which could be the result of eliminating some correlated features.

Under Random Forest, the peak score is recorded in Gravier and Pomeroy, while it drops to the lowest point in Golub, Khan, Sorlie, and Yeoh (Figures 7, 9, 12, 5). On the other hand, looking at KNN, RBR reached the peak performance only in Alon and Golub (Figures 3, 7), while it shows moderate to low performance in other datasets.

Finally, for SVM, RBR shows strong performance in terms of reaching the peak and maintaining consistently good scores in Colon, Gravier, and Sorlie (Figures 3, 8, 12), while it demonstrates moderate performance in the remaining datasets.

Further investigation assessed how often the highest performance score of the comparison feature selection methods beat, tied, or lost to the peak performance of RBR across the 8 datasets and 4 classifier pairs, using a practical relevance threshold of 0.1 pp (differences below this value were counted as ties).

For example, in Table 3, the first row indicates that Information Gain won 10 times against RBR, meaning its peak performance score was higher than the top performance of RBR in 10 dataset–classifier cases. For instance, in Figure 10, under KNN, IG wins against RBR while, at the same time, it loses against it under Random Forest, so it counts as one win and one loss (RBR is omitted because it serves as the reference).

Table 3: Win/tie/loss record for each feature selection versus RBR

FS	**Wins**	**Ties**	**Losses**
IG	10	9	13
SU	11	8	13
MRMR	14	6	12
ReliefF	12	6	14

RBR secures more wins than losses against IG, SU, and ReliefF and is outperformed more frequently by mRMR. To provide an absolute benchmark, Table 4 reports the grand mean F_1 score achieved by each filter across all datasets and classifiers.

Table 4: Dataset mean F_1 for every filter

Dataset	RBR	IG	SU	MRMR	ReliefF
Alon (Colon Cancer)	0.730	0.674	0.666	0.663	0.694
Golub (Leukemia)	0.955	0.979	0.974	0.948	0.973
Gravier (Breast Cancer)	0.689	0.659	0.660	0.700	0.650
Khan (SRBCT)	0.939	0.973	0.991	0.972	0.980
Pomeroy (CNS Tumor)	0.610	0.630	0.615	0.634	0.614
Shipp (Lymphoma)	0.876	0.870	0.858	0.901	0.895
Sorlie (Breast Cancer)	0.725	0.714	0.711	0.683	0.703
Yeoh (Leukemia)	0.923	0.955	0.955	0.960	0.954

RBR delivered the highest mean F_1 on the Alon and Sorlie datasets. It remained within the top three performers on Gravier and Shipp, confirming competitive robustness in those scenarios. However, the method exhibited the lowest average score on Khan, Pomeroy, and Yeoh.

Aggregate performance was assessed by computing the mean F_1 for each filter under every classifier, with averages taken across the ten subset thresholds and the eight datasets. The resulting classifier-specific scores are presented below.

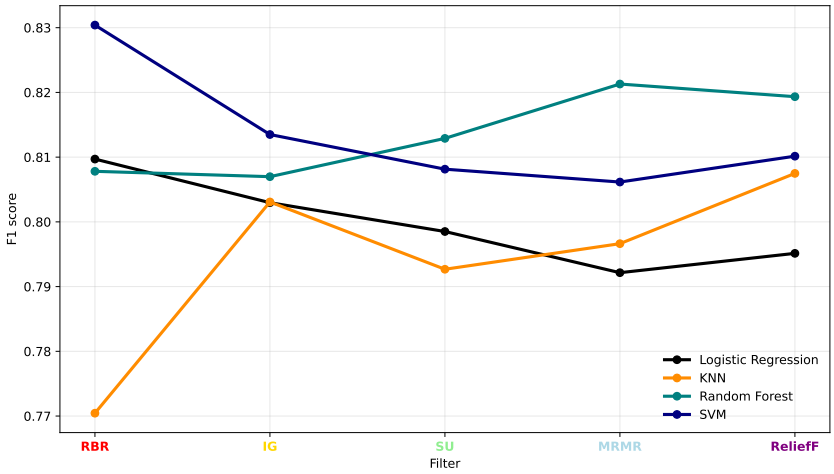


Figure 1: mean F_1 feature selection/classifier comparison

For better insight, we rank each filter method with respect to its performance under each classifier, with one indicating the best and five the lowest.

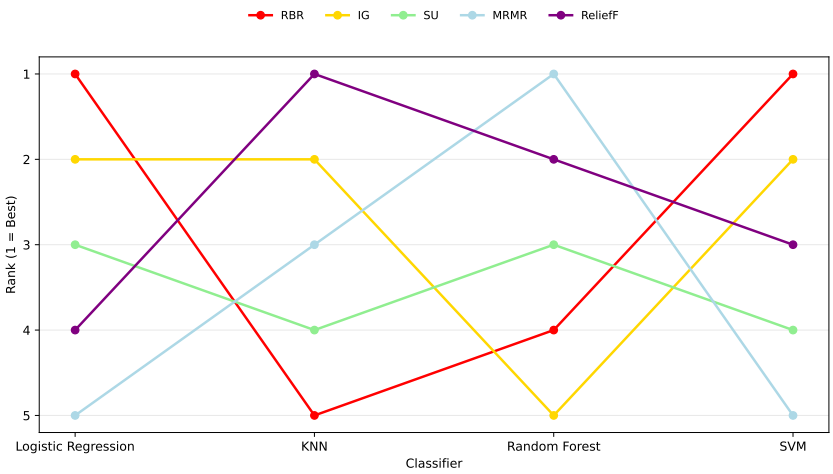


Figure 2: mean F_1 Feature Selection/Classifier Rank

In the following paragraph, we will dive deep into three cases where RBR performed well, moderately, and poorly compared to other baseline feature selection methods.

Alon (Colon Cancer) In the Alon dataset, RBR demonstrates consistently positive and statistically significant gains under all classifiers, particularly logistic regression, KNN, and SVM. The gains shown in Table 6 ranged from +5.00 pp to +11.43 pp, with corresponding p-values below 0.02 in most comparisons in Table 5. These advantages were most pronounced at intermediate feature retention levels, where peak scores were observed between 50% and 60% under logistic regression and SVM, and around 40–50% under KNN (Figure 3).

For example, the +10.4762 pp improvement over ReliefF under logistic regression coincided with this mid-range peak region. Performance declined gradually when feature retention fell below 20%, particularly for logistic regression. These results indicate that RBR is most effective on this dataset.

Table 5: Wilcoxon p-values for the *Alon* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0059**	**0.0098**	**0.0107**	**0.0152**
KNN	**0.0059**	**0.0098**	**0.0107**	0.1689
Random Forest	0.1602	0.1723	0.3750	**0.0431**
SVM	**0.0098**	**0.0039**	**0.0039**	0.0645

Table 6: Median F_1 difference score for the *Alon* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+5.0000	+9.2857	+10.0000	+10.4762
KNN	+10.9524	+11.1905	+4.7619	+3.3333
Random Forest	-1.9048	-0.4762	+2.8571	-5.2381
SVM	+9.5238	+10.9524	+11.4286	+6.1905



Figure 3: Alon dataset performance for each FS

Sorlie (Breast Cancer) On the Sorlie dataset, RBR demonstrated strong performance compared to the benchmark methods, as reflected in the statistical analysis and tuning behavior. Significant improvements were primarily observed under logistic regression and SVM (Figure 4), with positive gains reaching +7.69 pp and +11.33 pp (Table 8), respectively, and corresponding p-values below 0.02 (Table 7).

Under logistic regression, pronounced performance peaks were observed between 90–80% and around 30–20% feature retention, where RBR consistently outperformed competing methods (Figure 4). For instance, the +8.08 pp advantage over mRMR ($p = 0.0020$) was attained near these optimal regions, followed by a gradual decline below 20% retention.

Under SVM, a clear peak was recorded between 70–60% features, accompanied by sustained superiority over mRMR and ReliefF, while a noticeable drop was observed around 20% retention (Figure 4). In contrast, under KNN and Random Forest, most comparisons were non-significant, with small negative median differences and relatively flat performance curves.

Table 7: Wilcoxon p-values for the *Breast Cancer (Sorlie)* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0098**	**0.0059**	**0.0020**	**0.0195**
KNN	0.0858	0.1602	0.5566	0.2324
Random Forest	0.3223	0.3223	0.5566	0.0840
SVM	0.0663	0.0840	**0.0020**	**0.0020**

Table 8: Median micro-F_1 difference scores for the *Breast Cancer (Sorlie)* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+7.69	+7.00	+8.08	+5.58
KNN	-7.00	-6.66	-3.41	-2.48
Random Forest	-0.36	-0.57	-0.18	-2.91
SVM	+6.83	+6.31	+11.33	+7.92

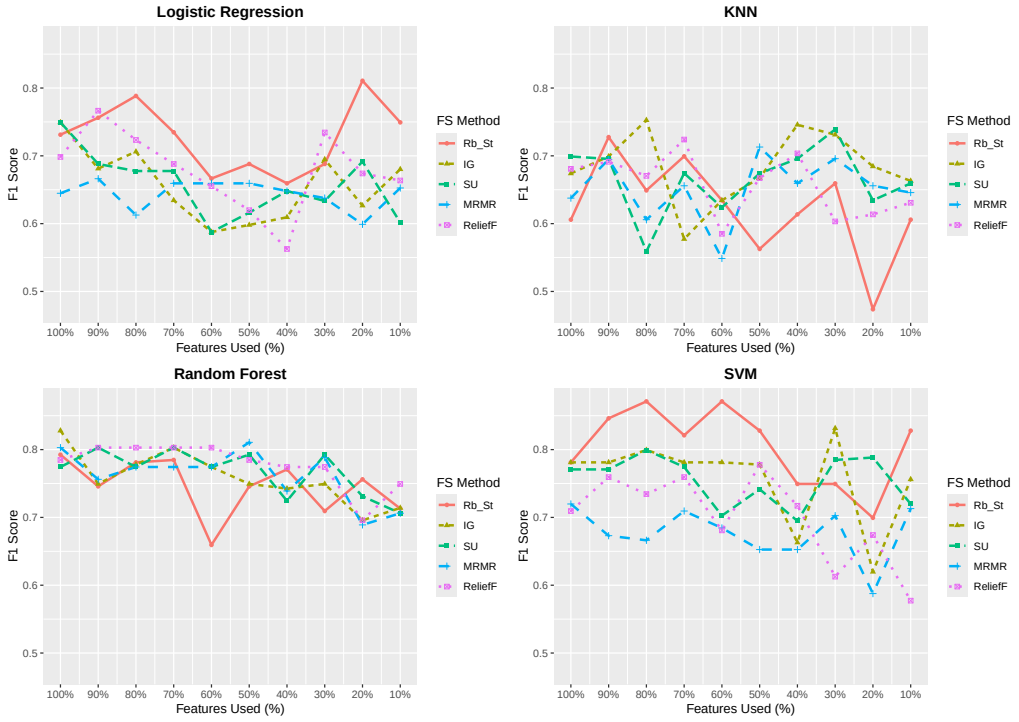


Figure 4: Sorlie dataset performance for each FS

Yeoh (Leukemia) In the Yeoh dataset, RBR had the lowest performance, as is clearly observed in Figure 5, supported by negative gains (Table 10) and statistically significant performance outcomes across classifiers (Table 9). Median losses ranged from −1.74 pp to −4.17 pp, with all associated p-values below 0.03, indicating systematic underperformance relative to all baselines.

For example, under KNN, RBR performed 4.17 pp worse than both SU and mRMR, and these deficits were highly significant. The figures further demonstrate that no performance peaks were achieved across any classifier. A clear, consistent decline was recorded after removing 30% of features. Even within the initial range of 100–70%, RBR remained the lowest compared to the baseline methods.

Table 9: Wilcoxon p-values for the *Yeoh* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0076**	**0.0109**	**0.0077**	**0.0059**
KNN	**0.0039**	**0.0039**	**0.0020**	**0.0039**
Random Forest	**0.00195**	**0.0284**	**0.0273**	**0.0059**
SVM	**0.00195**	**0.00195**	**0.00195**	**0.00195**

Table 10: Median micro-F_1 difference scores for the *Yeoh* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	-2.50	-1.74	-2.54	-2.10
KNN	-3.77	-4.17	-4.17	-2.90
Random Forest	-2.82	-2.52	-2.00	-2.43
SVM	-2.03	-2.43	-2.40	-2.43

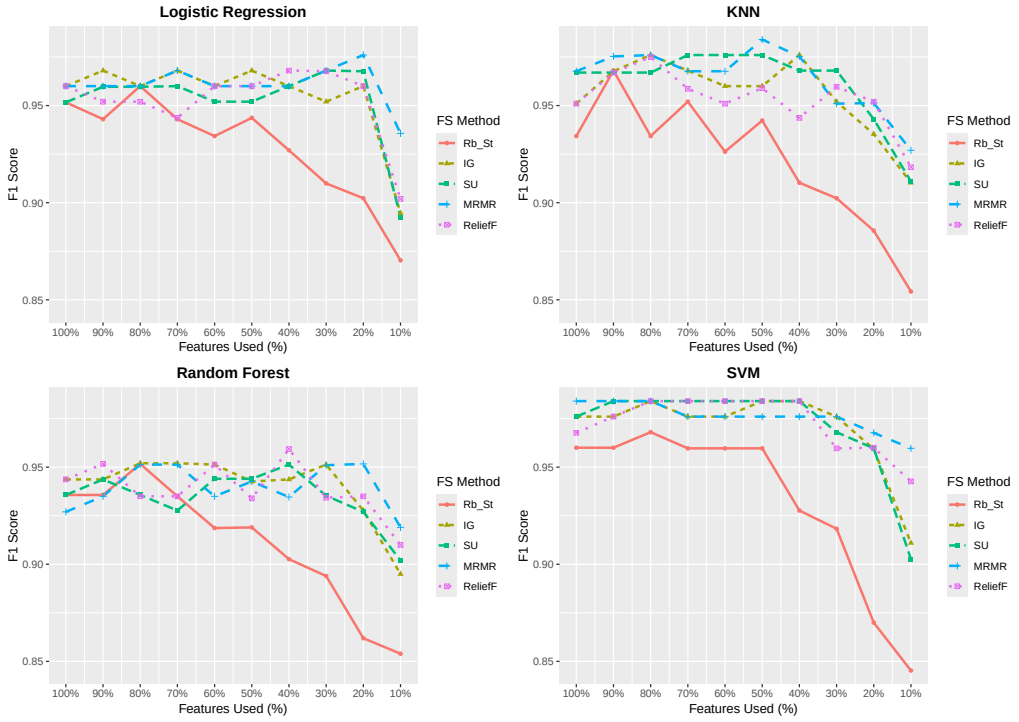


Figure 5: Yeoh dataset performance for each FS

The complete set of experimental figures and results tables for all datasets are provided in the Appendix.

7 Conclusions

In conclusion, classifying gene expression data continues to be a crucial aspect of cancer prediction and diagnosis, especially given the challenges associated with high dimensional data and small sample sizes. In this study, a feature selection method was developed based on the k-sample Bayesian Relative Belief Ratio (RBR), which demonstrated decent performance in multiclass classification problems across synthetic and gene expression datasets. By leveraging the RBR score and strength, the proposed method effectively identified and ranked significant genes that contribute to the classification in the majority of datasets with cancer types, including leukemia, colon tumor, lung cancer, breast cancer, lymphoma, and CNS tumor.

The application of the RBR method to synthetic datasets further validated its capacity to identify relevant features even in the presence of a substantial number of irrelevant ones. The results showed that the RBR method achieved 90% correct

feature identification. The averaged RBR and strength values for the features provided insight into their relative importance, as shown in Table 1.

The experimental analysis of eight multiclass cancer gene datasets highlighted the efficacy of the RBR method in guiding feature selection and achieving the best classification performance under two classifier. The results showed that the RBR method outperformed or were competetoive to baseline feature selection technique (Information Gain (IG), Symmetrical Uncertainty (SU), Minimum Redundancy Maximum Relevance (mRMR), and ReliefF) in dimensionality reduction while maintaining optimal predictive accuracy. The ability of the RBR method to rank features based on their relative importance, even in high dimensional gene expression cancer datasets, highlights its utility in real world applications.

Despite its advantages, the RBR method incurs higher computational costs due to the hyperparameter tuning required for prior and posterior distributions. Furthermore, variations in the chosen interval $[a_i, b_i]$ for the prior μ_i can introduce slight randomness, highlighting the need for careful parameter selection.

The proposed method potential for extension to categorical features and its stability across diverse datasets represent promising directions for future research. Overall, the RBR-based feature selection algorithm offers a powerful tool for addressing the challenges of multiclass gene expression data classification, providing a pathway for improved cancer diagnostics and a deeper understanding of the molecular mechanisms underlying these diseases.

Declarations

Author Contribution:
All authors contributed equally to the manuscript, including the methodology, data analysis, investigation, and preparation of the final version. All authors reviewed and approved the final draft prior to submission.

Ethics, Consent to Participate, and Consent to Publish declarations:
Not applicable.

Data Availability:
The datasets analyzed in this study are publicly available gene expression microarray datasets. A detailed description of all datasets, including their characteristics and sources, is provided in Section 4. All datasets used in this study are accessible through the datamicroarray R package (Ramey, 2016) and the original studies cited in Section 4. No new datasets were generated during the current study.

Clinical Trial Number:
Not applicable.

Competing Interests:
The authors declare that they have no competing interests.

Funding declaration:
No financial fund has been received for this research.

Appendix

Table 11: Wilcoxon p-values for the *Alon* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0059**	**0.0098**	**0.0107**	**0.0152**
KNN	**0.0059**	**0.0098**	**0.0107**	0.1689
Random Forest	0.1602	0.1723	0.3750	**0.0431**
SVM	**0.0098**	**0.0039**	**0.0039**	0.0645

Table 12: Median micro-F_1 difference score for the *Alon* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+5.0000	+9.2857	+10.0000	+10.4762
KNN	+10.9524	+11.1905	+4.7619	+3.3333
Random Forest	-1.9048	-0.4762	+2.8571	-5.2381
SVM	+9.5238	+10.9524	+11.4286	+6.1905



Figure 6: Alon dataset performance for each FS

Table 13: Wilcoxon p-values for the *Golub* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	0.6250	0.4922	**0.0098**	**0.0273**
KNN	0.1091	**0.0204**	0.5533	**0.0378**
Random Forest	**0.0066**	**0.0066**	**0.0066**	**0.0066**
SVM	0.1085	0.3722	**0.0273**	0.2119

Table 14: Median micro-F_1 difference score for the *Golub* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	-2.8571	-2.8571	-2.8571	-2.8571
KNN	-1.6071	-2.8571	+2.0833	-2.8571
Random Forest	-2.8571	-2.8571	-2.8571	-2.8571
SVM	-1.8452	-0.5952	+4.1667	-0.5952

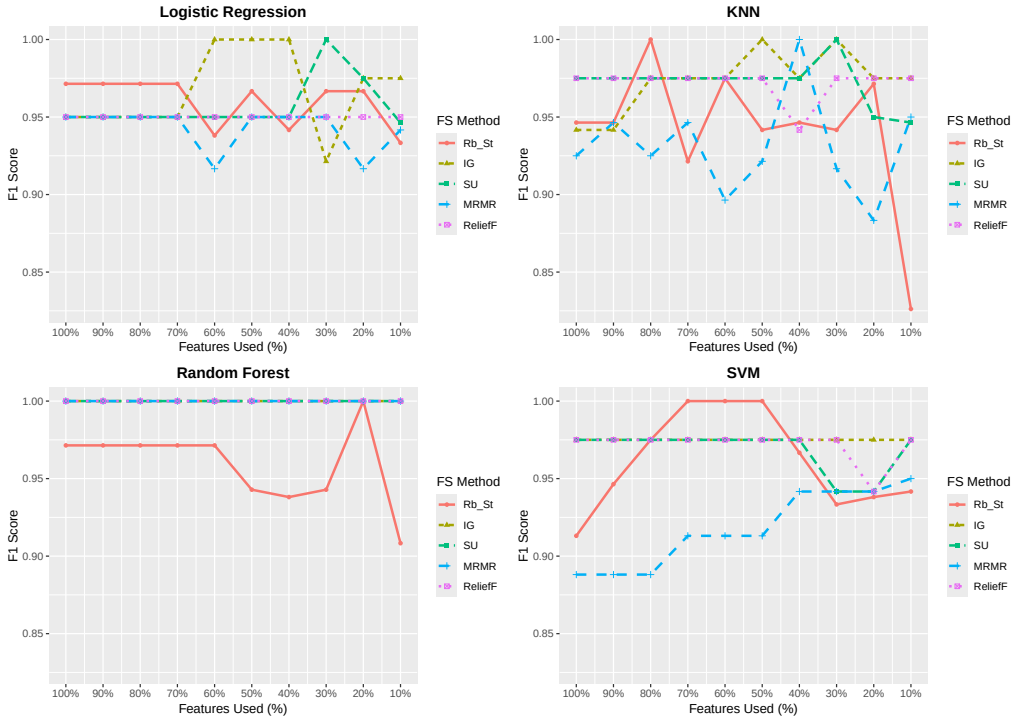


Figure 7: Golub dataset performance for each FS

Table 15: Wilcoxon p-values for the *Gravier* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0059**	0.4922	1.0000	0.1055
KNN	0.2754	0.3750	**0.0039**	0.0645
Random Forest	**0.0077**	**0.0273**	0.3750	**0.0209**
SVM	**0.0098**	**0.0059**	1.0000	**0.0208**

Table 16: Median micro-F_1 difference score for the *Gravier* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+2.4632	+1.1397	+0.0735	+4.7426
KNN	-2.3162	-1.1765	-3.5294	+1.7647
Random Forest	+6.6544	+4.9632	+1.6912	+6.7279
SVM	+4.2647	+6.8750	-0.6250	+4.3015

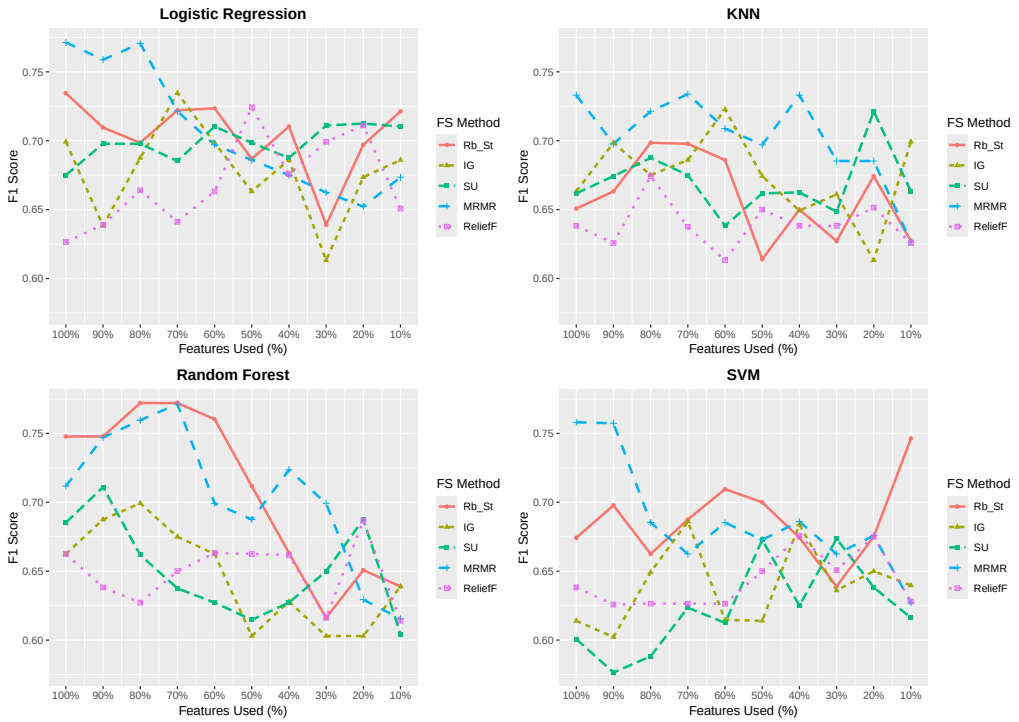


Figure 8: Gravier dataset performance for each FS

Table 17: Wilcoxon p-values for the *Khan* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	0.1142	**0.0160**	0.1189	0.4331
KNN	**0.0071**	**0.0071**	0.0569	**0.0112**
Random Forest	**0.0106**	**0.0074**	**0.0112**	**0.0097**
SVM	0.2356	**0.0339**	1.0000	0.0588

Table 18: Median micro-F_1 difference score for the *Khan* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	-2.5000	-2.5000	-2.5000	-2.5000
KNN	-3.3333	-4.5833	-0.8333	-4.5833
Random Forest	-2.5000	-5.0000	-2.9167	-2.5000
SVM	-1.2500	-1.2500	0.0000	0.0000

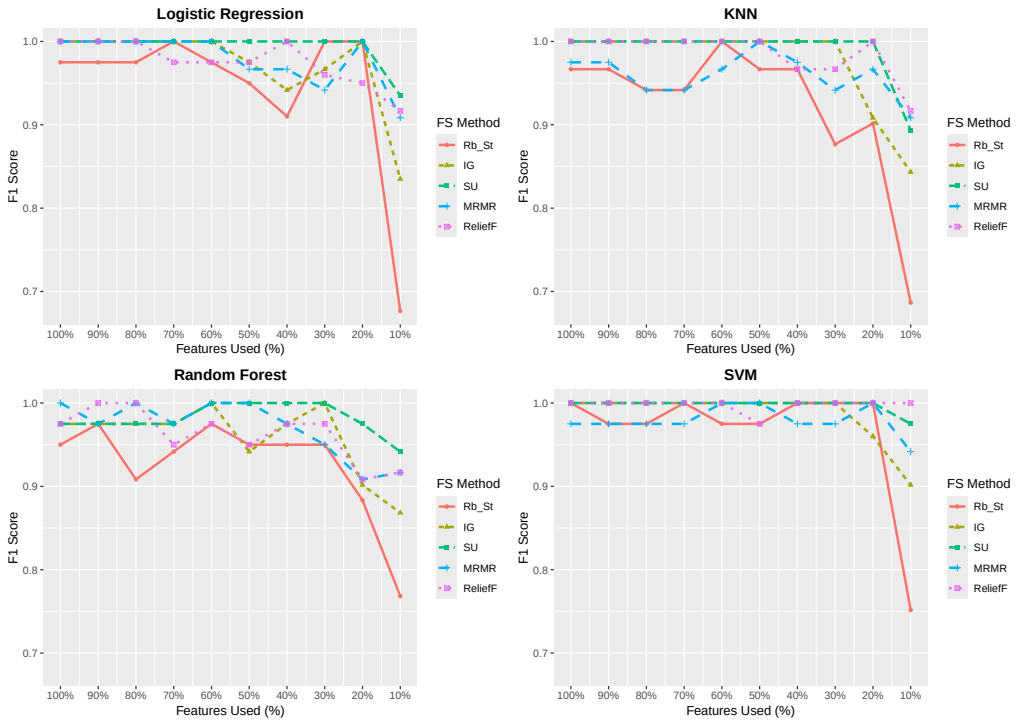


Figure 9: Khan dataset performance for each FS

Table 19: Wilcoxon p-values for the *Pomeroy* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0164**	0.4002	**0.0273**	0.3750
KNN	**0.0137**	0.2353	0.0887	**0.0108**
Random Forest	0.2603	0.4002	0.9156	**0.0104**
SVM	**0.0114**	**0.0339**	0.3173	**0.0256**

Table 20: Median micro-F_1 difference score for the *Pomeroy* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	-4.00	-3.33	-8.67	-3.00
KNN	-7.33	-3.33	-2.00	-6.67
Random Forest	+0.67	+0.33	+0.00	+4.00
SVM	+3.33	+1.67	+0.00	+3.33



Figure 10: Pomeroy dataset performance for each FS

Table 21: Wilcoxon p-values for the *Shipp* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	0.0656	0.0925	0.8885	0.1602
KNN	**0.0059**	**0.0371**	**0.0020**	**0.0020**
Random Forest	**0.0175**	**0.0371**	0.6250	0.1369
SVM	0.7659	0.1501	0.2123	0.8457

Table 22: Median micro-F_1 difference score for the *Shipp* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+3.75	+6.07	+1.25	+7.50
KNN	-7.14	-6.79	-8.39	-13.04
Random Forest	+7.50	+6.25	+0.54	-1.43
SVM	+1.25	+1.25	-1.25	+0.18

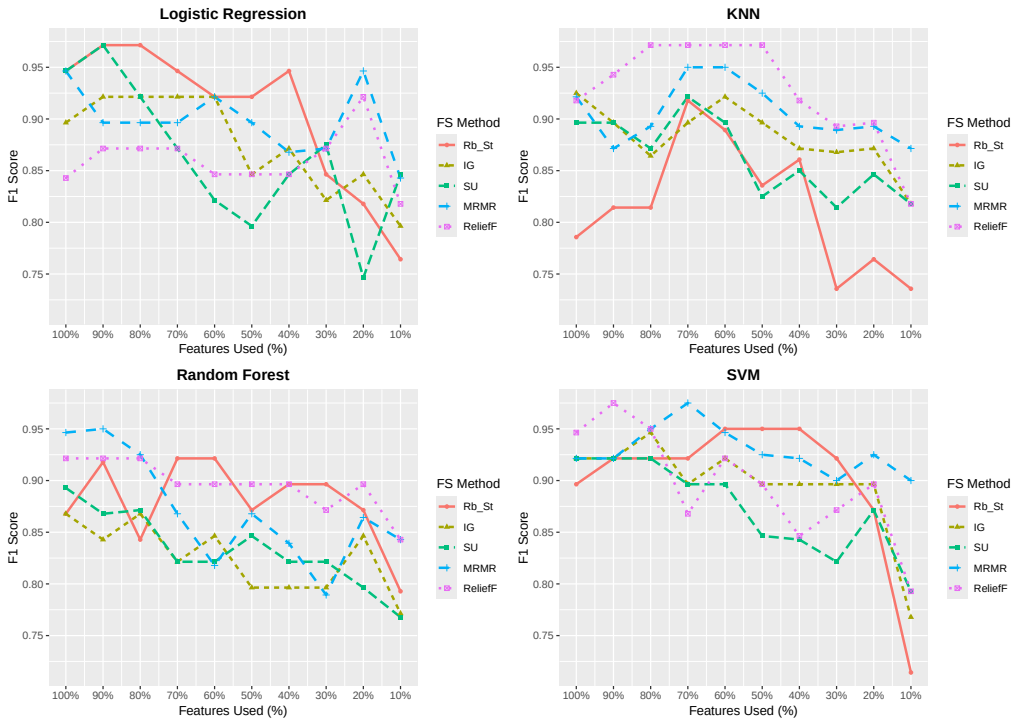


Figure 11: Shipp dataset performance for each FS

Table 23: Wilcoxon p-values for the *Sorlie* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0098**	**0.0059**	**0.0020**	**0.0195**
KNN	0.0858	0.1602	0.5566	0.2324
Random Forest	0.3223	0.3223	0.5566	0.0840
SVM	0.0663	0.0840	**0.0020**	**0.0020**

Table 24: Median micro-F_1 difference scores for the *Sorlie* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	+7.69	+7.00	+8.08	+5.58
KNN	-7.00	-6.66	-3.41	-2.48
Random Forest	-0.36	-0.57	-0.18	-2.91
SVM	+6.83	+6.31	+11.33	+7.92



Figure 12: Sorlie dataset performance for each FS

Table 25: Wilcoxon p-values for the *Yeoh* dataset

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	**0.0076**	**0.0109**	**0.0077**	**0.0059**
KNN	**0.0039**	**0.0039**	**0.0020**	**0.0039**
Random Forest	**0.00195**	**0.0284**	**0.0273**	**0.0059**
SVM	**0.00195**	**0.00195**	**0.00195**	**0.00195**

Table 26: Median micro-F_1 difference scores for the *Yeoh* dataset (pp)

Classifier	IG	SU	MRMR	ReliefF
Logistic Regression	-2.50	-1.74	-2.54	-2.10
KNN	-3.77	-4.17	-4.17	-2.90
Random Forest	-2.82	-2.52	-2.00	-2.43
SVM	-2.03	-2.43	-2.40	-2.43

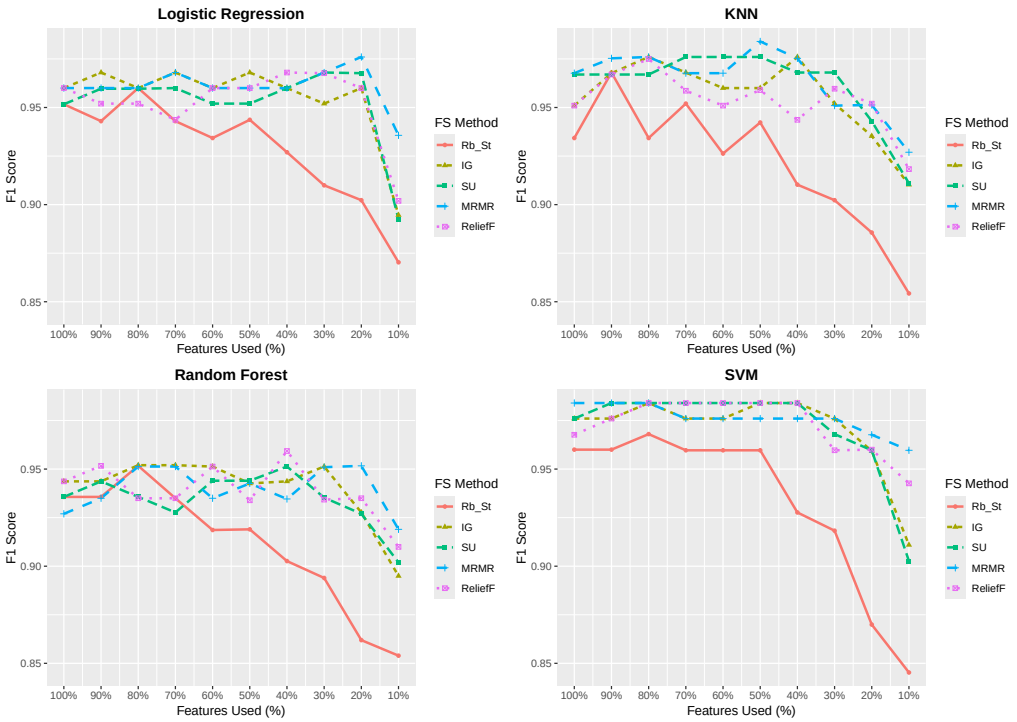


Figure 13: Yeoh dataset performance for each FS

References

Alzaatreh, A., Emarly, M., Mitra, R., Al-Labadi, L. (2025). Feature selection in binary classification problems using relative belief ratio. *Physica Scripta*, *100*(8), 086010.

Abdelrazeq, I., Al-Labadi, L., & Alzaatreh, A. (2020). On one-sample bayesian tests for the mean. *Statistics*, *54*(2), 424–440.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... others (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511.

Al-Labadi, & Alzaatreh. (2024). *K-sample bayesian test for the mean.* (Submitted)

Al-Labadi, & Evans. (2017). Optimal robustness results for relative belief inferences and the relationship to prior-data conflict.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999a). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745–6750.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., ... Korsmeyer, S. J. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, *30*(1), 41–47.

Baldi, P., & Long, A. D. (2001a). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509–519.

Baldi, P., & Long, A. D. (2001b). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509–519.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000). Tissue classification with gene expression profiles. In *Proceedings of the fourth annual international conference on computational molecular biology* (pp. 54–64).

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., ... others (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, *98*(24), 13790–13795.

Blessie, C., & Karthikeyan. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, *6*(3), 385-394. doi: 10.1260/1748-3018.6.3.385

Boateng, & Abaye, e. a. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, *7*(04), 190.

Campbell, & Sturim, e. a. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE signal processing letters*, *13*(5), 308–311.

Chandra, B., Shanker, S., & Mishra, S. (2006). A new approach: Interrelated two-way clustering of gene expression data. *Statistical Methodology*, *3*(1), 93–102.

Evans. (2016). Measuring statistical evidence using relative belief. *Computational and structural biotechnology journal*, *14*, 91–96.

Evans, M. (2015). *Measuring statistical evidence using relative belief ratio.* CRC press.

Fox, R. J., & Dimmic, M. W. (2006). A two-sample bayesian t-test for microarray data. *BMC bioinformatics*, *7*, 1–11.

Urbanowicz, RJ., & Moore, JH. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, 189–203.

Bugata, P., & Drotar, P. (2020). On some aspects of minimum redundancy maximum relevance feature selection. *Science China Information Sciences*, *63*(1), 112103.

Chai, Z., & Li, Y. (2023). Symmetric uncertainty based decomposition multi-objective immune algorithm for feature selection. *Swarm and Evolutionary Computation*, *78*, 101286.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, *19*(1), 1–67.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... others (1999a). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., ... Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, *62*(17), 4963–4967.

Gravier, Eleonore, Pierron, G., Vincent-Salomon, A., gruel, N., Raynal, V., Savignoni, A., ... Delattre, O. (2010, September). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*, *49*(12), 1125–1125.

Jafari, P., & Azuaje, F. (2006a). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, *6*, 1–8.

Chandra, B., & Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, *44*(4), 529–535. (ISSN: 1532-0464)

Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., & Safaraliev, M. (2025). Mathematical methods in feature selection: A review. *Mathematics*, *13*(6).

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., ... Meltzer, P. S. (2001, June). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679.

Liu, H., Liu, L., & Zhang, H. (2010). Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, *43*(1), 81–87.

Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. *Information Systems*, *28*(4), 243–268.

Muller, & Fetterman. (2003). *Regression and anova: an integrated approach using sas software*. John Wiley & Sons, Inc.

Munson, & Caruana. (2009). On feature selection, bias-variance, and bagging. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine learning and knowledge discovery in databases* (pp. 144–159). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ooi, C. H., & Tan, P. (2003). Genetic algorithms applied to multiclass prediction for the analysis of gene expression data. *Bioinformatics*, *19*(1), 37–44.

Pavya, & Srinivasan. (2018). Review of literature on filter and wrapper methods for feature selection. *International journal of engineering sciences & research*

technology, *7*(1), 137–143.

Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of biomedical informatics*, *43*(1), 15–23.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., ... Golub, T. R. (2002, January). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*(6870), 436–442.

Ramey, J. A. (2016). Datamicroarray: Collection of data sets for classification. *R Package, available at https://rdrr.io/github/ramhiser/datamicroarray*, *10*, .

Rigatti, & Steven. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31–39.

Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. In *Feature extraction: Foundations and applications* (pp. 463–470). Springer.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, *23*(19), 2507–2517.

Shen, Q., Shi, W.-m., & Kong, W. (2009). New gene selection method for multiclass tumor classification by class centroid. *Journal of Biomedical Informatics*, *42*(1), 59–65.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., ... Golub, T. R. (2002, January). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*(1), 68–74.

Sobhanzadeh, & Motieghader, e. a. (2019). Featureselect: a software for feature selection based on machine learning approaches. *BMC bioinformatics*, *20*(1), 1–17.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., ... Børresen-Dale, A.-L. (2001, September). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, *98*, 10869–10874.

Sun, & Huang. (2010). An adaptive k-nearest neighbor algorithm. In *2010 seventh international conference on fuzzy systems and knowledge discovery* (Vol. 1, pp. 91–94).

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., & Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*(2), 133–143.

Zhou, X., Liu, K.-Y., & Wong, S. T. (2004). Cancer classification and prediction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics*, *37*(4), 249–259.