Highlights

Pack and Extract: An Effective Graph-Based Approach for Feature Selection

Faisal N. Abu-Khzam, Joseph R. Barr, Cynthia Chaaya, Nadim Obeid, Peter Shaw

- Graph-based feature selection via weighted d-packing
- Compact feature sets with improved stability
- Strong predictive performance on biological datasets
- Significant dimensionality reduction with few features
- Biologically meaningful and interpretable selected genes

Pack and Extract: An Effective Graph-Based Approach for Feature Selection

Faisal N. Abu-Khzam[a,*], Joseph R. Barr[b], Cynthia Chaaya[a], Nadim Obeid[a], Peter Shaw[c]

[a] *Lebanese American University, Beirut, Lebanon*
[b] *Embry-Riddle Aeronautical University, Prescott, Arizona, USA*
[c] *Oujiang Laboratory, Wenzhou, Zhejiang, China*

Abstract

Many molecular biology data sets contain thousands of features, such as genes in single-cell experiments or other omics data. Standard predictive models often perform poorly in such settings, as they tend to retain a large number of features and fail to adequately handle correlations among them. Combinatorial approaches, such as feature selection based on the Independent Dominating Set (IDS) model, have shown promising results, but appear to have been largely overlooked, possibly due to their high computational complexity.

In this paper, we propose an improvement to IDS-based feature selection that (i) favors features exhibiting greater diversity and (ii) generalizes the notion of independent dominating set to a node packing model. The resulting formulation is used as an intermediate step within the feature selection process. We evaluate our approach on five genetic data sets and compare the results with those obtained using existing methods. Our experimental results demonstrate notable improvements in predictive accuracy over traditional models, as well as increased stability of the selected feature sets.

Keywords: Feature selection, Node Packing, Weighted Independent Dominating Set, Classification

*Corresponding author

Email addresses: faisal.abukhzam@lau.edu.lb (Faisal N. Abu-Khzam), barrj4@erau.edu (Joseph R. Barr), cynthia.chaaya@lau.edu (Cynthia Chaaya), nadim.obeid@lau.edu (Nadim Obeid), petershaw@ojlab.ac.cn (Peter Shaw)

1. Introduction

The principle of parsimony, or Occam's razor, is fundamental in philosophy and the natural sciences, and plays a central role in modeling physical phenomena and, more generally, in statistical modeling. In statistical learning, this principle manifests itself most clearly through *feature selection*, whose goal is to identify a small subset of informative variables while discarding redundancy and noise.

Feature selection seeks to preserve or improve predictive performance while reducing overfitting, enhancing interpretability, and lowering computational complexity. Existing approaches can be broadly classified as supervised or unsupervised. Supervised methods include techniques based on statistical tests such as the F-test and the chi-squared test, as well as information-theoretic and model selection criteria such as the Akaike and Bayesian Information Criteria. Unsupervised methods include dimensionality reduction and redundancy control techniques such as principal component analysis and variance inflation factor (VIF)-based approaches.

A wide range of heuristic strategies has also been proposed. Some rely on purity criteria, such as minimizing the Gini coefficient, and are commonly used in decision tree learning. Ensemble methods, including random forests and boosting, combine multiple decision trees to improve predictive performance. Other learning frameworks, such as *support vector machines (SVMs)*, rely on constrained optimization formulations that minimize convex loss functions subject to margin constraints. More recently, *regularization methods*, such as the *lasso* procedure, have become popular for shrinking the effective input dimension and controlling model complexity [1].

In the simplest setting involving two variables, pairwise inspection using Pearson's correlation coefficient may suffice to detect dependency, and removing one of two highly correlated variables is often adequate. However, this approach fails to capture correlations between a variable and a group of other variables. The variance inflation procedure partially addresses this limitation by using the *variance inflation factor (VIF)*, which relies on the coefficient of determination (R^2) from multiple linear regression to identify variables that are correlated with subsets of other variables. Nevertheless, such approaches remain limited in their ability to capture more global dependency structures.

An alternative perspective consists of modeling feature dependencies using graphs. In this setting, features correspond to vertices, and edges encode

pairwise correlations exceeding a prescribed threshold. Graph-based representations naturally capture structural relationships among features and enable the use of combinatorial optimization techniques for feature selection. In particular, selecting a minimum subset of features that dominates the correlation graph while remaining mutually independent leads to the *Independent Dominating Set* (IDS) model, which was proposed more than a decade ago [2]. Despite promising empirical results, this approach has received limited attention, possibly due to the reliance on integer linear programming formulations with exponential worst-case complexity.

Feature selection is especially critical in medical and biological applications, where data sets often exhibit extremely high dimensionality and strong correlations among variables. Reducing dimensionality in such settings facilitates both predictive modeling and the interpretation of underlying biological mechanisms [3, 4]. Empirical studies have also shown that no single feature selection method consistently outperforms others across all evaluation criteria, highlighting the importance of combining complementary perspectives and performance measures [5].

Motivated by these observations, we revisit graph-based feature selection and propose a more flexible combinatorial framework that extends the IDS model. Rather than restricting attention to immediate adjacency, we consider a generalized approach based on the notion of a *node packing*, or d-packing [6], which enforces separation between selected features at a prescribed graph distance. To differentiate between correlated features, we further incorporate a diversity criterion that favors informative variables, allowing the model to select representative features from correlated groups.

Contributions.. The contributions of this work are fourfold. First, we introduce a diversity-driven weighting scheme based on mutual information to guide the selection of informative features within correlated groups. Second, we formulate graph-based feature selection as a weighted d-packing problem on correlation graphs, generalizing independent dominating set–based models. Third, we develop an efficient greedy algorithm inspired by maximum-utility heuristics, thereby avoiding reliance on integer linear programming formulations. Finally, we demonstrate on multiple high-dimensional biological data sets that the proposed graph-based preselection, when combined with standard shrinkage methods, improves both predictive accuracy and the stability of selected feature sets.

2. Preliminaries

We assume familiarity with basic graph-theoretic terminology. A *dominating set* in a graph $G = (V, E)$ is a set $D \subseteq V$ such that every vertex in $V \setminus D$ has at least one neighbor in D. A *d-packing* in G is a subset $P \subseteq V$ satisfying the following conditions: (i) for any $v, w \in P$, the length of a shortest path between v and w is strictly greater than d, and (ii) for every vertex $v \in V \setminus P$, there exists a vertex in P at distance at most d from v. The special case $d = 1$ corresponds to the notion of a *maximal independent set*. In this context, a set of vertices is independent if no two of its elements are adjacent. The definition of a 1-packing ensures maximality in the following sense: if P is a 1-packing in G, then (i) no two vertices in P are adjacent, and (ii) no additional vertex can be added to P without violating independence. Hence, P is inclusion-wise maximal (i.e., it is not properly contained in a larger independent set).

The problem of finding a minimum-size maximal independent set is equivalent to the Independent Dominating Set (IDS) problem, since every vertex outside the set must be adjacent to a vertex inside it. The IDS problem is a well-studied combinatorial optimization problem [7] and is known to be $\mathcal{NP}$-hard [8] and $\mathcal{PLS}$-hard [9]. The asymptotically fastest known exact algorithm for IDS runs in time $\mathcal{O}(1.3351^n)$ [10].

Feature Selection

The primary goal of feature selection is to identify a subset of independent variables that captures as much relevant information as possible using as few features as possible. More precisely, the aim is to select a subset of features for which a predictive model is optimal, where optimality is typically assessed using cross-validation.

The term *dimensionality reduction* is often used in this context, as training instances (x, y) with $x \in \mathbb{R}^p$ are replaced by (x', y) with $x' \in \mathbb{R}^q$, where $q < p$. An optimal feature set typically leads to a model that achieves a favorable bias–variance trade-off [11, 12]. This issue is particularly evident in standardized gene expression data, which often contain a large number of variables with substantial pairwise correlations. In such cases, two highly correlated features can often be replaced by a single representative feature without degrading model performance [11]. Variables that appear irrelevant are frequently dominated by noise, and their removal can improve generalization. Identifying an optimal feature subset becomes especially challenging in

high-dimensional settings where the number of features exceeds the number of samples [13].

Let $G = (V, E)$ be a graph constructed from tabular data, where vertices correspond to features and an edge connects two vertices if the corresponding features are deemed "close" according to a correlation measure, to be specified later. Computing a minimum d-packing in G yields a set of features that are sufficiently uncorrelated, while ensuring that every remaining feature lies within distance at most d from one of the selected features.

In this work, we further assume that features are associated with weights reflecting their statistical relevance. Our objective is therefore to identify a minimum-weight d-packing, where the weight of a packing is the sum of the weights of its vertices. We refer to the resulting optimization problem as the MINIMUM WEIGHTED d-PACKING problem (or equivalently, the (W,d)-PACKING problem). Observe that when $d = 1$, this formulation generalizes the classical INDEPENDENT DOMINATING SET problem. Consequently, the problem is $\mathcal{NP}$-hard via a straightforward reduction from the unweighted d-packing problem.

We assign feature weights based on a notion of *diversity*: features exhibiting greater variability with respect to the class label are considered more informative. This concept is discussed in [14]. Since our formulation is minimization-based, lower weights are assigned to more diverse features, or equivalently, the greedy heuristic is adapted to prioritize such features. Further details are provided in the next section.

The Role of VC Dimension

The Vapnik–Chervonenkis (VC) dimension measures the capacity of a hypothesis class, indicating the largest number of points that can be shattered by hypotheses from that class. A higher VC dimension corresponds to increased flexibility and lower bias, but also a greater risk of overfitting when the training sample size is limited.

Within the framework of Probably Approximately Correct (PAC) learning, one can derive bounds on the generalization error. For any hypothesis h, with probability at least $1 - \delta$, the following holds:

$$|R(h) - \hat{R}(h)| \leq \sqrt{\frac{2}{m}\left(d \ln\left(\frac{2m}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right)},$$

where $R(h)$ denotes the true risk, $\hat{R}(h)$ the empirical risk on a sample of size m, and d the effective VC dimension. Moreover, to achieve an error margin

ϵ with high probability, the sample size must satisfy

$$m = \Omega\left(\frac{d + \ln(1/\delta)}{\epsilon^2}\right).$$

These bounds highlight the trade-off between model complexity and the amount of data required for reliable learning.

Evaluation Metrics

If a model is too simple, for example by using too few features, it may suffer from high bias and fail to capture important structure in the data. Conversely, overly complex models may exhibit high variance and overfit noise, resulting in poor generalization.

To assess this trade-off, we use several evaluation metrics in addition to accuracy. Specifically, we report:

- **Recall:**

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

- **Precision:**

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

- **F1 score:**

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

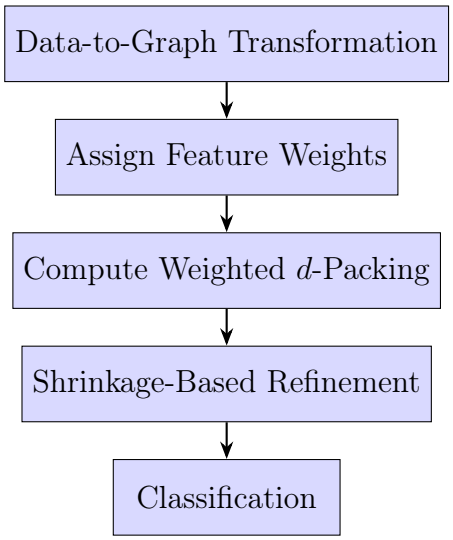
Barr et al. [15] proposed balancing bias and variance by gradually increasing model complexity until predictive performance peaks, an approach that they formally justified for generalized linear models and supported empirically.

In our setting, the d-packing parameter d plays a role analogous to a capacity parameter. Smaller values of d impose stricter separation between selected features, leading to higher bias, while larger values allow greater flexibility at the expense of increased variance. Selecting an appropriate value of d thus serves as a form of regularization and can improve overall model performance.

3. Methodology

We propose a graph-based feature selection framework that combines combinatorial optimization with statistical learning techniques. The overall approach follows a five-stage pipeline, illustrated in Fig. 1. First, the input data set is transformed into a correlation graph. Second, feature-level weights reflecting statistical diversity are assigned to graph vertices. Third, a weighted d-packing is computed to obtain a reduced and weakly correlated feature set. Fourth, additional shrinkage-based feature selection methods are applied as a refinement step. Finally, classification models are trained using the selected features.

Figure 1: Overview of the proposed feature selection pipeline



3.1. Graph Construction

Following [2], features are represented as vertices in an undirected graph. Let V denote the set of vertices corresponding to the columns of the data matrix. Two vertices $u, v \in V$ are connected by an edge if the corresponding features exhibit sufficient correlation, defined as

$$\big|\mathrm{cor}(u, v)\big| \geq \tau,$$

where $\mathrm{cor}(u, v)$ denotes the Pearson correlation coefficient and $\tau \in (0, 1]$ is a user-defined threshold parameter. Pearson correlation is used throughout

this work due to its simplicity, interpretability, and widespread use in biological data analysis. The resulting graph captures pairwise dependencies between features and serves as the structural basis for subsequent combinatorial selection.

3.2. Feature Diversity and Weight Assignment

To differentiate between correlated features, we associate each vertex with a weight that reflects the feature's statistical relevance. As a measure of feature diversity, we use mutual information [16], which quantifies the dependency between a feature X and the target variable Y and is well suited for continuous-valued data:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Features exhibiting higher mutual information with respect to the target variable are considered more informative. Since the underlying combinatorial problem is formulated as a minimization problem, lower weights are assigned to more informative features, thereby biasing the selection process toward informative variables.

3.3. Weighted d-Packing Formulation

Given the weighted correlation graph $G = (V, E)$ and an integer $d \geq 1$, our objective is to compute a subset $P \subseteq V$ that forms a d-packing and optimizes the total weight of the selected vertices. Again, the case $d = 1$ corresponds to a weighted independent dominating set, while larger values of d enforce stronger separation between selected features. Obviously, this formulation generalizes the classical Independent Dominating Set problem, which is $\mathcal{NP}$-hard even in the unweighted case. Consequently, exact methods based on integer linear programming do not scale to high-dimensional feature spaces, motivating the use of efficient heuristic approaches.

3.4. Greedy Algorithm for Weighted d-Packing

Algorithm 1 presents a greedy heuristic for computing a weighted d-packing. The algorithm iteratively selects vertices based on a utility function defined as the sum of weights in the closed neighborhood of a vertex. This strategy is inspired by the classical Chvátal's maximum-utility heuristic for

Algorithm 1: Greedy Weighted d-Packing Heuristic

```
Data: col is an array of size n initialized to 0;
queue is an empty queue;
dpack is an empty set.
Result: dpack
x ← getMaxDegUncol()
while x ≠ −1 do
    add x to queue;
    add x to dpack;
    col[x] ← 1;
    while queue is not empty do
        v ← queue.poll();
        if col[v] ≤ d then
            for each neighbor w of v do
                if col[w] = 0 then
                    add w to queue;
                    col[w] ← col[v] + 1;
                end
            end
        end
    end
    x ← getMaxDegUncol();
end
Return dpack
```

the Set Cover problem, where elements with the largest marginal contribution are selected first.

While the pseudocode is self-explanatory, we should note that array `col` stores the distance of each vertex from the nearest selected vertex. The function `getMaxDegUncol` returns an uncolored vertex with maximum neighborhood utility. The algorithm ensures that selected vertices are separated by a distance greater than d, while all remaining vertices are dominated within distance d.

Computational Complexity.. Constructing the correlation graph requires $O(p^2)$ time in the number of features p, which is unavoidable for pairwise correlation-based approaches. The greedy d-packing algorithm runs in $O(k|V| + |E|)$

time, where k is the packing size. To see this, first note that each edge is explored at most once during the breadth-first search procedure (and each vertex joins the queue exactly once), while the linear-time `getMaxDegUncol` routine is called exactly k times. In practice, the packing size is typically much smaller than the number of features. Since subsequent shrinkage methods operate on the reduced feature set, the overall pipeline scales well to high-dimensional data while avoiding the super-polynomial complexity of exact combinatorial formulations.

3.5. Shrinkage-Based Refinement

After graph-based pre-selection, we apply shrinkage methods to further refine the feature set. Regularization techniques mitigate overfitting by penalizing model complexity, thereby improving the bias–variance trade-off.

We employ Elastic Net regularization, which combines ℓ_1 and ℓ_2 penalties and minimizes

$$\arg\min_{w} \left(\sum_{i=1}^{n} (y_i - f(x_i; w))^2 + \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} \|w\|_2^2 \right),$$

where $\lambda_1, \lambda_2 \geq 0$. Elastic Net is particularly effective in the presence of correlated features, as it can retain groups of correlated variables while eliminating redundant ones.

We also consider Support Vector Machine Recursive Feature Elimination (SVM-RFE), which iteratively ranks features according to their contribution to a trained SVM classifier and removes the least informative features.

3.6. Classification Models

In the final stage, classification models are trained using the selected features. We use Random Forests and Support Vector Machines with linear kernels, and perform hyperparameter tuning to optimize predictive performance.

4. Experimental Results

Our experiments were conducted on six benchmark data sets commonly used in feature selection studies: colon cancer (ColonCA [17] [18]), leukemia [19], lymphoma, and small blue round cell tumor (SRBCT), obtained from [2], the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [20], and a

heart disease dataset [21]. These data sets span both binary and multi-class classification settings and are characterized by high dimensionality and strong feature correlations. Table 1 summarizes the main characteristics of the gene-expression and WDBC data sets.

For each data set, we computed the coefficient of variation for each feature and used mutual information with the class label as the feature weight. Pairwise Pearson correlation coefficients were then computed between features to build a correlation graph: two vertices are connected whenever the absolute correlation exceeds a given threshold. The resulting weighted graph was processed using our proposed heuristic, henceforth dubbed d-WPE, for varying values of the distance parameter d and the correlation threshold. The selected feature sets were optionally refined using Elastic Net or SVM-RFE, followed by classification. Stability was computed using the same protocol as in [2] to ensure comparability.

Dataset	Features	Instances	Classes	Class Distribution
Colon Cancer	2001	62	2	Normal: 20 – Tumor: 40
Leukemia	7128	72	2	AML: 25 – ALL: 47
Lymphoma	4027	66	3	FL: 9 – CLL: 11 – DLBCL: 46
SRBCT	2309	83	4	1: 29 – 2: 11 – 3: 18 – 4: 25
Breast Cancer	30	569	2	Malignant: 212 – Benign: 357

Table 1: Summary of datasets

4.1. Colon Dataset Results

For the colon cancer dataset [2], Elastic Net and SVM-RFE without d-WPE reduced the original 2001 features to 11, achieving a classification accuracy of 0.77. Incorporating d-WPE with a correlation threshold of 0.6 and $d = 2$ further reduced the feature set to three features, with an accuracy of 0.69. Increasing d resulted in a single selected feature but lowered the accuracy to 0.62.

Reducing the correlation threshold to 0.55 significantly improved performance. At $d = 2$, our d-WPE heuristic selected a single feature while maintaining an accuracy of 0.77 and achieving perfect stability. Table 2 compares these results with competing methods, and Fig. 2 shows the evaluation metrics as a function of d.

Method	Corr. Thresh.	Accuracy	Stability	Precision	F1-score	Sensitivity	Specificity	Num. of Features
d-WPE	0.55	0.77	1.00	0.78	0.82	0.88	0.60	1
d-WPE	0.60	0.62	1.00	0.71	0.67	0.63	0.60	1
Stability Lasso [2]	0.60	0.83	0.61	–	–	–	–	5
Elastic Net	–	0.77	0.61	0.78	0.82	0.88	0.60	23
Elastic Net with SVM-RFE	–	0.77	0.72	0.78	0.82	0.88	0.60	11

Table 2: Comparison of d-WPE vs. other methods on the colon cancer dataset [2].

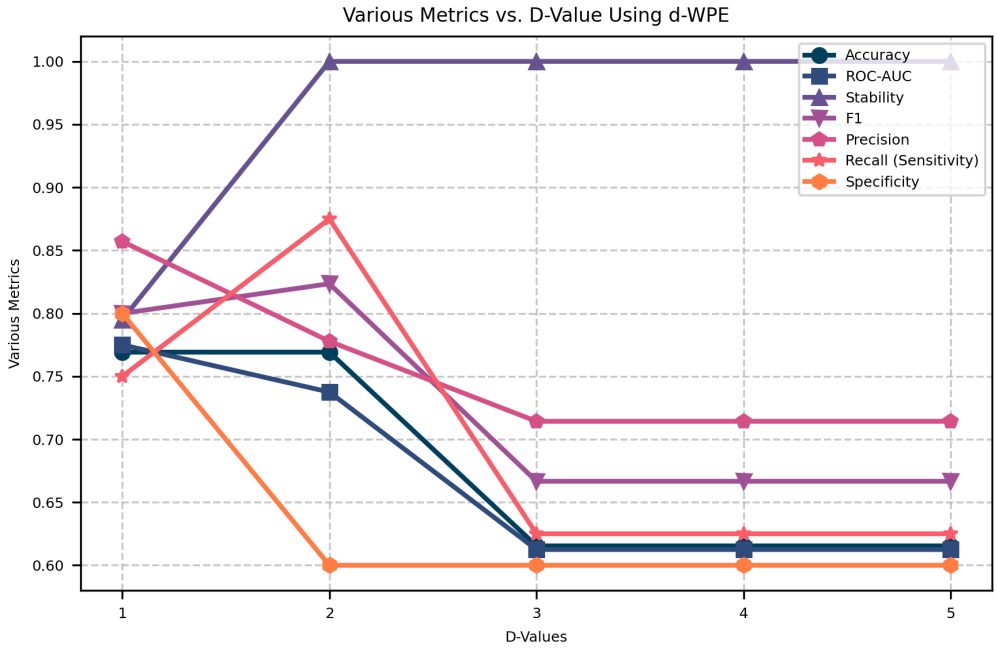


Figure 2: Metric results vs. d when applying d-WPE with correlation threshold 0.55 on the colon cancer dataset [2].

4.2. Leukemia Dataset Results

Applying d-WPE to the leukemia dataset [2] with a correlation threshold of 0.6 reduced the original 7128 features to a single feature at $d = 6$, achieving an accuracy of 0.94 with perfect stability. Lowering the threshold to 0.55 resulted in four selected features with the same accuracy and improved balance across evaluation metrics. Table 3 summarizes the comparison, and Fig. 3 shows the evolution of the metrics as a function of d.

Method	Corr. Thresh.	Accuracy	Stability	Precision	F1-score	Sensitivity	Specificity	Num. of Features
d-WPE	0.55	0.94	1.00	0.86	0.92	1.00	0.92	4
d-WPE	0.60	0.94	1.00	1.00	0.91	0.83	1.00	1
SVM-RFE with MIDS [2]	0.60	0.99	0.70	–	–	–	–	14
Elastic Net	–	1.00	0.70	1.00	1.00	1.00	1.00	65
Elastic Net with SVM-RFE	–	1.00	0.86	1.00	1.00	1.00	1.00	32

Table 3: Comparison of d-WPE vs. other methods on the leukemia dataset [2].

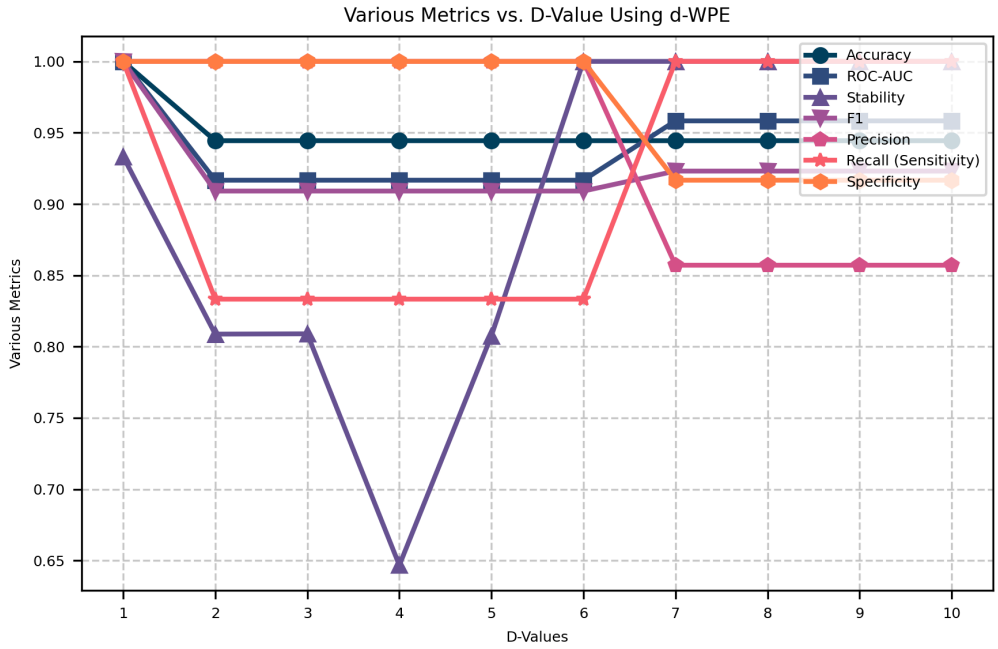


Figure 3: Metric results vs. d when applying d-WPE with correlation threshold 0.55 on the leukemia dataset [2].

4.3. Lymphoma Dataset Results

The lymphoma dataset [2] involves multi-class classification. To accommodate this, Elastic Net and SVM-RFE were applied using a one-vs.-rest strategy. Using d-WPE with a correlation threshold of 0.6 reduced the number of features from 4027 to seven at $d = 1$, achieving an accuracy of 0.90. With a lower threshold of 0.55, d-WPE reduced the feature set to three features at $d = 5$ while achieving an accuracy of 0.85 and a stability score of 0.80. Table 4 reports the detailed comparison, and Fig. 4 shows the metric behavior with respect to d.

Method	Corr. Thresh.	Accuracy	Stability	Precision	F1-score	Sensitivity	Specificity	Num. of Features
d-WPE	0.55	0.85	0.80	0.83	0.80	0.84	0.94	3
d-WPE	0.60	0.90	0.91	0.87	0.89	0.95	0.96	7
SVM-RFE with MIDS [2]	0.60	0.99	0.60	–	–	–	–	9
Elastic Net	–	1.00	0.47	1.00	1.00	1.00	1.00	145
Elastic Net with SVM-RFE	–	0.95	1.00	0.92	0.94	0.98	0.98	5

Table 4: Comparison of d-WPE vs. other methods on the lymphoma dataset [2].

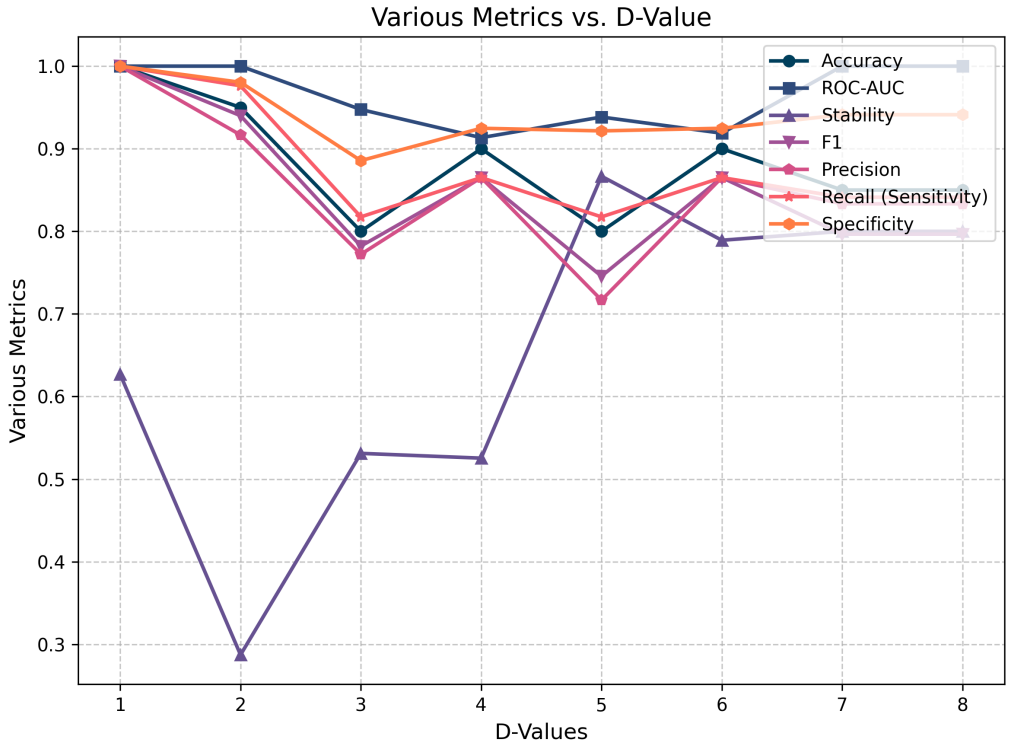


Figure 4: Metric results vs. d when applying d-WPE with correlation threshold 0.55 on the lymphoma dataset [2].

4.4. SRBCT Dataset Results

We applied the same multi-class procedure to the SRBCT dataset [2]. Using our d-WPE heuristic with a correlation threshold of 0.6, the number of features was reduced from 2308 to 19 at $d = 8$, achieving an accuracy of 0.76. With a threshold of 0.55, the method reduced the features to four at $d = 5$ with an accuracy of 0.81. Increasing d further reduced the number of features (e.g., to three at $d = 7$), but with a substantial drop in accuracy. Table 5 and Fig. 5 report the detailed results.

Method	Corr. Thresh.	Accuracy	Stability	Precision	F1-score	Sensitivity	Specificity	Num. of Features
d-WPE	0.55	0.81	1.00	0.83	0.82	0.83	0.93	4
d-WPE	0.60	0.76	0.78	0.80	0.77	0.78	0.91	19
SVM-RFE with MIDS [2]	0.60	0.96	0.58	–	–	–	–	24
Elastic Net	–	1.00	0.47	1.00	1.00	1.00	1.00	215
Elastic Net with SVM-RFE	–	1.00	0.86	1.00	1.00	1.00	1.00	14

Table 5: Comparison of d-WPE vs. other methods on the SRBCT dataset [2].

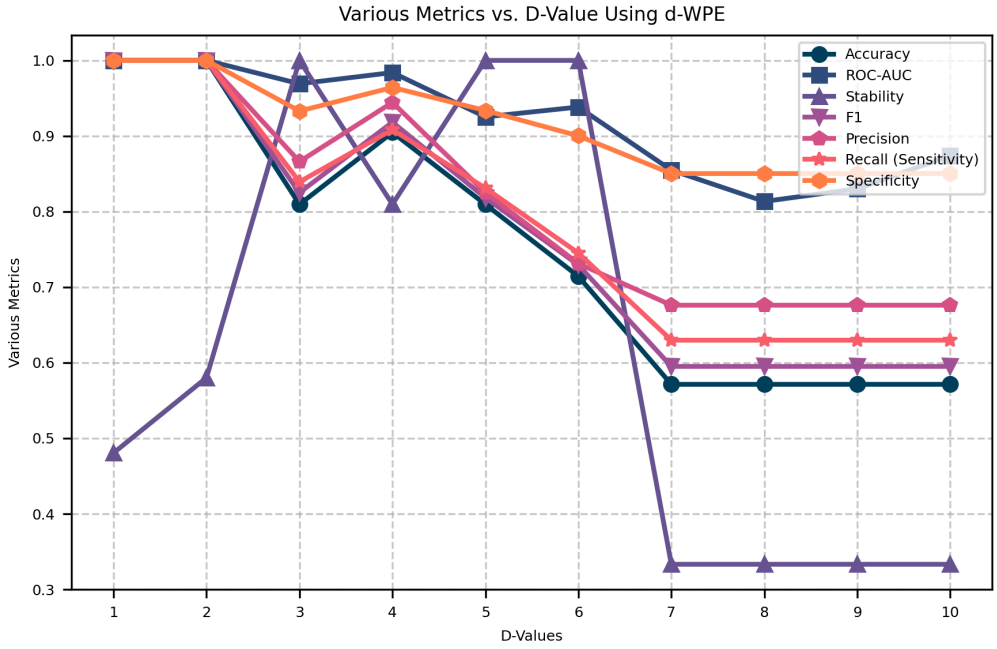


Figure 5: Metric results vs. d when applying d-WPE with correlation threshold 0.55 on the SRBCT dataset [2].

4.5. Breast Cancer Dataset Results

We tested the proposed method on the WDBC data set [20] using correlation thresholds of 0.55 and 0.70 for graph construction. The results are compared with recent feature selection methods from [22], [23], [24], and [25] in Table 6. Using $\tau = 0.70$, d-WPE selected three features and achieved an accuracy of 0.9825 with ROC AUC 0.9984. Using $\tau = 0.55$, the method selected two features and achieved a perfect ROC AUC, while maintaining competitive performance in the other metrics. We further note that comparison with the previously used methods (in Tables 2-5) was not possible,

simply because these methods were never used on the WDBC dataset.

Method	Accuracy	F1-score	Sensitivity	Specificity	ROC AUC	Num. of Features
d-WPE (0.55)	0.9386	0.9176	0.9070	0.9577	1.0000	2
d-WPE (0.70)	0.9825	0.9756	0.9524	0.9825	0.9984	3
WFSIB [22]	0.9912	0.9880	0.9858	0.9943	–	10
ABCoDT [23]	0.9718	–	–	–	–	2
ELM-RBF [24]	0.9569	0.9830	0.9756	0.9439	–	9
PCA-CA [25]	0.9900	0.9761	0.9600	–	–	10
Elastic Net	0.9561	0.942529	0.953488	0.957746	0.9971	24

Table 6: Comparison of feature selection methods on the breast cancer dataset

4.6. Heart Disease Dataset Results

We also tested the proposed method on a heart disease dataset [21]. In this case, the method reduced the number of selected features from 13 to 1, and the evaluation metrics remained constant across d values because the computed d-packing set did not change. For reference, Noroozi et al. [26] evaluated several feature selection methods on the same dataset and reported that the smallest feature subset (four features) was obtained via backward selection. In contrast, our method selected a single feature while maintaining comparable performance.

4.7. Classification Results

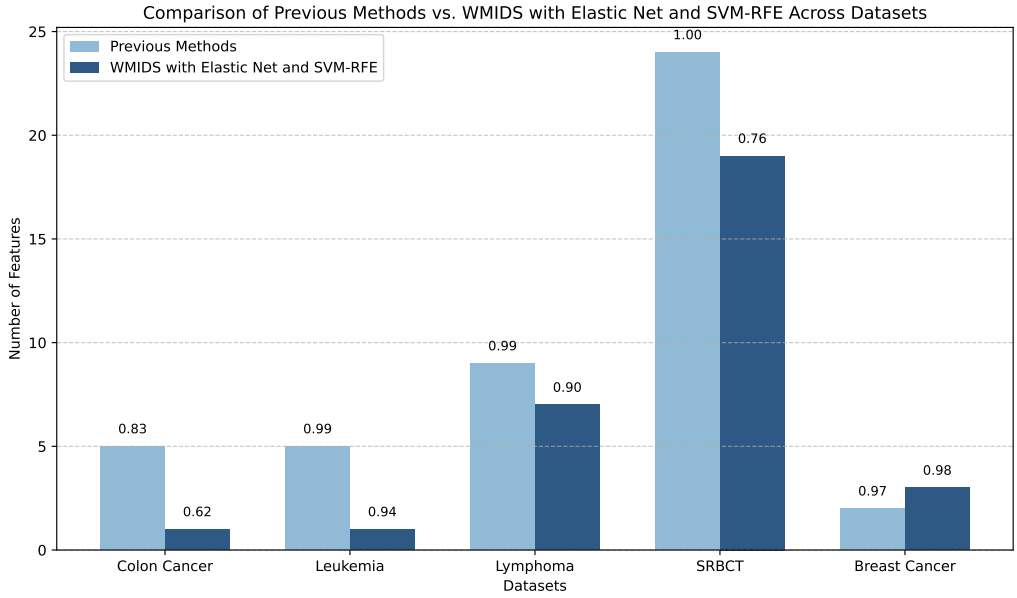


Figure 6: Feature selection comparison with methods from [2] and [27] versus d-WPE combined with Elastic Net and SVM-RFE. Classification accuracy is shown above the bars.

Overall, d-WPE consistently achieved substantial feature reduction. Figure 6 summarizes the minimum reduced feature counts in comparison with [2] and [28, 27]. Our method typically achieves stronger feature reduction with slightly lower accuracy in some cases, but with noticeably higher stability compared to the graph-based approaches in [2]. The difference is particularly pronounced on the leukemia and SRBCT datasets.

The correlation threshold plays a central role in controlling the sparsity (or density) of the correlation graph, and therefore the achievable feature reduction. For example, on the colon dataset, thresholds 0.55 and 0.6 both yield a single selected feature but with different accuracies (Table 2). On SRBCT, $\tau = 0.60$ yields 19 selected features with accuracy 0.76, whereas $\tau = 0.55$ yields four selected features with improved accuracy (Table 5). These observations indicate that τ should be selected carefully for each data set to balance reduction and predictive performance.

Finally, the choice of vertex weights can also influence results. For instance, on the WDBC dataset, replacing Gini-based weights by mutual in-

formation led to a noticeable increase in accuracy, supporting the role of diversity-aware weighting in correlated feature settings.

4.8. Biological Validation of Selected Features

To evaluate the biological relevance of features selected by d-WPE, we mapped the selected feature indices back to their corresponding gene identifiers using the original annotation data for each dataset. The colon cancer dataset [18] uses UniGene identifiers linked to GenBank accessions, while the leukemia dataset [17] uses Affymetrix HuGeneFL (Hu6800) probe identifiers. Gene annotations were resolved via the Bioconductor hu6800.db annotation package and NCBI Gene linkage.

Colon cancer.. The two features selected by d-WPE at $\tau = 0.55$ (achieving accuracy 0.70 with a single feature) correspond to:

- Feature 266: *CSRP1* (cysteine and glycine rich protein 1, Entrez Gene 1465), a LIM domain protein expressed in smooth muscle cells.
- Feature 575: *MORF4L2* (mortality factor 4 like 2, Entrez Gene 9643), a chromatin remodeling factor involved in transcriptional regulation.

Notably, *CSRP1* is a smooth muscle marker that lies directly within the primary biological signal identified by Alon et al. [18]: their analysis found that colon tumor samples were distinguished from normal tissue largely by the reduced expression of smooth muscle genes, reflecting the higher proportion of smooth muscle cells in normal colon tissue. The fact that d-WPE independently selects a gene from this core signal (with no prior biological knowledge encoded in the method) provides strong evidence that the graph-theoretic approach captures genuinely informative features rather than statistical artifacts. A recent study has also reported *CSRP1* as differentially expressed in colon adenocarcinoma [29].

Leukemia.. The features selected by d-WPE at $\tau = 0.60$ (single feature, accuracy 0.94) and $\tau = 0.55$ (four features, accuracy 0.94) map to the following genes:

The single feature at $\tau = 0.60$, *CST3*, ranks 49th out of 7129 features in the Golub et al. [17] distinction metric, placing it well within the top 1% of discriminative genes for the ALL/AML classification task. *TRDV2*, a T cell receptor delta chain gene, ranks 82nd. These rankings confirm that d-WPE selects features that are individually highly discriminative.

Feature	Gene	Threshold	PubMed hits
1881	*CST3* (cystatin C)	$\tau = 0.60$	6
3344	*CCR3* (C-C chemokine receptor 3)	$\tau = 0.55$	8
6166	*TRDV2* (T cell receptor δ variable 2)	$\tau = 0.55$	5
6853	*SFTPA1* (surfactant protein A1)	$\tau = 0.55$	0
3548	*PRAME* (PRAME nuclear receptor)	$\tau = 0.55$	106
575	*NUP188* (nucleoporin 188)	plateau	0

Table 7: Leukemia features selected by d-WPE, their corresponding genes, and PubMed citation counts for the gene in the context of leukemia.

5. Discussion of Results

The experimental results can be interpreted in terms of the bias–variance trade-off and, more broadly, statistical learning theory. In our graph-based feature selection model based on d-packing, the distance parameter d implicitly controls model capacity by limiting how many features can be selected simultaneously and how correlated they may be. Smaller values of d allow larger feature sets, which may reduce bias but increase variance due to redundancy and correlation. Larger values of d enforce stronger separation constraints, leading to more compact feature sets with reduced variance, potentially at the cost of increased bias.

Although deriving explicit VC-dimension bounds for classifiers built on d-packing–selected features is nontrivial, the empirical behavior observed across all datasets is consistent with this interpretation. In particular, tuning d proved to be an effective mechanism for controlling model complexity in practice. While accuracy sometimes decreased slightly as d increased, stability and specificity consistently improved. These gains are especially important in high-dimensional biological settings, where robustness and reproducibility of selected features are often more valuable than marginal improvements in predictive accuracy.

From the perspective of graph-based modeling, the proposed d-WPE FRAMEWORK generalizes earlier IDS-based approaches by relaxing strict independence constraints. This added flexibility allows the method to better adapt to the structure of correlation graphs derived from real data. Across all evaluated datasets, d-WPE achieved substantial improvements in stability. For example, stability increased by approximately 0.2 on the colon cancer dataset, by about 0.05 on the leukemia dataset, and by roughly 0.6 on the

lymphoma dataset for suitable values of d. These results suggest that enforcing distance-based separation between selected features is an effective way to mitigate instability caused by highly correlated variables.

The leukemia dataset illustrates particularly well the practical implications of choosing the node packing parameter d. In this case, different values of d led to models with distinct trade-offs between accuracy and precision. While one choice of d yielded slightly higher accuracy, another offered better precision and stability. Such situations highlight the importance of combining quantitative evaluation with domain knowledge. Interestingly, the emergence of a single dominant feature at higher values of d may warrant further biological investigation. From a graph-theoretic viewpoint, such a feature could correspond to a structurally central vertex, analogous to those observed in cluster editing models where a single vertex can belong to multiple clusters via the notion of vertex splitting [30], and may represent a biologically meaningful candidate gene.

From a molecular biology perspective, one of the main advantages of the proposed approach is its ability to drastically reduce the number of selected features. When feature selection models retain hundreds of variables, biological interpretation becomes impractical. By contrast, the compact feature sets produced by d-WPE make downstream analysis feasible. Standard tools can then be applied, including functional annotation and pathway analysis using resources such as NCBI [31], STRING [32], and DAVID [33]. When appropriate, these analyses can be complemented by experimental validation, such as survival analysis or Western blot experiments.

6. Conclusion

We introduced a graph-based framework for feature selection based on a weighted d-packing model. The key idea is to enforce a controlled separation between selected features while incorporating a notion of diversity through feature-dependent weights. This combination allows the method to select representative features from correlated groups while favoring those that are more informative with respect to the target variable. The parameter d provides a natural way to control the trade-off between redundancy and coverage, and leads to a more stable selection process in practice.

From an algorithmic perspective, the proposed greedy procedure avoids the need for expensive optimization formulations and scales well to high-dimensional datasets. Experimentally, the method achieves strong feature

reduction, improved stability, and competitive predictive performance across several biological benchmarks.

Overall, the results suggest that combining node packing with diversity-aware weighting provides an effective and interpretable approach to feature selection in correlated settings. In contrast to many black-box learning pipelines, the proposed framework offers a transparent selection mechanism in which the role of each feature can be traced back to explicit structural and statistical criteria. This is particularly important in biological applications, where understanding why certain variables are selected is often as valuable as predictive performance itself. More broadly, this work illustrates how simple combinatorial structures can be adapted to capture statistical properties of data in a meaningful and explainable way.

Acknowledgments

This research project was supported by the Lebanese American University under the President's Intramural Research Fund PIRF0056.

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) Vol. 58, No. 1 (1996).

[2] L. Shu, T. Ma, L. J. Latecki, Stable feature selection with minimal independent dominating sets, in: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB'13, Association for Computing Machinery, New York, NY, USA, 2013, p. 450–457. doi:10.1145/2506583.2506600.
URL https://doi.org/10.1145/2506583.2506600

[3] K. Kawamoto, C. A. Houlihan, E. A. Balas, D. F. Lobach, Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, Bmj 330 (7494) (2005) 765.

[4] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, B. J. Chow, G. Dwivedi, Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death, PloS one 14 (6) (2019) e0218760.

[5] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, F. E. Alsaadi, Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, Applied Soft Computing 86 (2020) 105836. doi:https://doi.org/10.1016/j.asoc.2019.105836.
URL https://www.sciencedirect.com/science/article/pii/S1568494619306179

[6] F. N. Abu-Khzam, G. B. Matar, S. Thoumi, Pack and measure: An effective approach for influence propagation in social networks, CoRR abs/2401.00525 (2024). arXiv:2401.00525, doi:10.48550/ARXIV.2401.00525.
URL https://doi.org/10.48550/arXiv.2401.00525

[7] Y. Wang, C. Li, M. Yin, A two phase removing algorithm for minimum independent dominating set problem, Applied Soft Computing 88 (2020) 105949.

[8] R. M. Karp, Reducibility among combinatorial problems, Springer, 2010.

[9] M. Borzechowski, The complexity class Polynomial Local Search (PLS) and PLS-complete problems, Institut für Informatik, Freien Universität Berlin, 2016.

[10] N. Bourgeois, F. Della Croce, B. Escoffier, V. Paschos, Fast algorithms for min independent dominating set, Discrete Applied Mathematics 161 (4) (2013) 558–572, seventh International Conference on Graphs and Optimization 2010. doi:https://doi.org/10.1016/j.dam.2012.01.003.
URL https://www.sciencedirect.com/science/article/pii/S0166218X1200011X

[11] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering 40 (1) (2014) 16–28, 40th-year commemorative issue. doi:https://doi.org/10.1016/j.compeleceng.2013.11.024.
URL https://www.sciencedirect.com/science/article/pii/S0045790613003066

[12] J. R. Barr, S. Zacks, Goodness-of-fit of statistical distributions, Encyclopedia with Semantic Computing and Robotic Intelligence 2 (02), 1850014 (2018).

[13] U. M. Khaire, R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University - Computer and Information Sciences 34 (4) (2022) 1060–1073. doi:https://doi.org/10.1016/j.jksuci.2019.06.012.
URL https://www.sciencedirect.com/science/article/pii/S1319157819304379

[14] M. Shaheen, N. Naheed, A. Ahsan, Relevance-diversity algorithm for feature selection and modified bayes for prediction, Alexandria Engineering Journal 66 (2023) 329–342. doi:https://doi.org/10.1016/j.aej.2022.11.002.
URL https://www.sciencedirect.com/science/article/pii/S1110016822007359

[15] J. R. Barr, M. Sobel, Y. Lu, B. A. Nguchu, P. Shaw, On the variability of statistical models, in: 2023 Fifth International Conference on Transdisciplinary AI (TransAI), IEEE, 2023, pp. 187–196.

[16] T. M. Cover, J. A. Thomas., Elements of. Information Theory. Second Edition, Wiley, 2006.

[17] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.

[18] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, 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) (1999) 6745–6750.

[19] T. Golub, golubEsets: exprSets for golub leukemia data, r package version 1.52.0 (2025). doi:10.18129/B9.bioc.golubEsets.
URL https://bioconductor.org/packages/golubEsets

[20] W. Wolberg, O. Mangasarian, N. Street, W. Street, Breast Cancer Wisconsin (Diagnostic), UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5DW2B (1993).

[21] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Heart Disease, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C52P4X (1989).

[22] K. Karimi, A. Ghodratnama, R. Tavakkoli-Moghaddam, Two new feature selection methods based on learn-heuristic techniques for breast cancer prediction: a comprehensive analysis, Annals of Operations Research 328 (09 2022). doi:10.1007/s10479-022-04933-8.

[23] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, L. Gu, Feature selection based on artificial bee colony and gradient boosting decision tree, Applied Soft Computing 74 (2019) 634–642. doi:https://doi.org/10.1016/j.asoc.2018.10.036.
URL https://www.sciencedirect.com/science/article/pii/S1568494618305933

[24] S. Mojrian, J. Hassannataj Joloudari, I. Felde, K. Kando, A. Szabo-Gali, N. László, A. Mosavi, Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system, 2020. doi:10.1109/RIVF48685.2020.9140744.

[25] M. Darzi, A. Liaei, M. Hosseini, H. Asghari, Feature selection for breast cancer diagnosis: A case-based wrapper approach, World Academy of Science, Engineering and Technology 53 (2011) 1142–1145.

[26] Z. Noroozi, A. Orooji, L. Erfannia, Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction, Scientific Reports 13 (1) (Dec 2023). doi:10.1038/s41598-023-49962-w.

[27] J. R. Barr, F. N. Abu-Khzam, P. Shaw, Feature selection via independent domination, in: 2023 Fifth International Conference on Transdisciplinary AI (TransAI), 2023, pp. 197–200. doi:10.1109/TransAI60598.2023.00048.

[28] F. N. Abu-Khzam, J. R. Barr, M. R. Benabid, P. Shaw, Feature selection via weighted independent domination, in: 2024 Conference on AI, Science, Engineering, and Technology (AIxSET), 2024, pp. 179–184. doi:10.1109/AIxSET62544.2024.00033.

[29] W. Zhang, X. Li, J. Chen, Identification of differentially expressed genes in colon adenocarcinoma and their prognostic significance, Frontiers in Genetics 14 (2023) 1120395.

[30] F. N. Abu-Khzam, E. Arrighi, M. Bentert, P. G. Drange, J. Egan, S. Gaspers, A. Shaw, P. Shaw, B. D. Sullivan, P. Wolf, Cluster editing with vertex splitting, Discrete Applied Mathematics 371 (2025) 185–195. doi:https://doi.org/10.1016/j.dam.2025.04.013.
URL https://www.sciencedirect.com/science/article/pii/S0166218X25001714

[31] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al., Ncbi taxonomy: a comprehensive update on curation, resources and tools, Database 2020 (2020) baaa062.

[32] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, et al., The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic acids research 51 (D1) (2023) D638–D646.

[33] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, David: database for annotation, visualization, and integrated discovery, Genome biology 4 (2003) 1–11.