mathematics

MDPI

Article

Linear-Time Multivariate Unsupervised Feature Selection Using Nyström-Based Entropy Approximation

Gyu-Il Kim [1], Timur Khairulov [1], Wangduk Seo [2,*] and Jaesung Lee [1,*]

[1] Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea; gyu6491@cau.ac.kr (G.-I.K.); timurkhairulov@cau.ac.kr (T.K.)
[2] Division of AI Computer Science and Engineering, Kyonggi University, Suwon 16227, Republic of Korea
* Correspondence: wdseo@kyonggi.ac.kr (W.S.); curseor@cau.ac.kr (J.L.)

Abstract

Unsupervised feature selection seeks to identify informative, nonredundant features without using class labels. Motivated by Fano's inequality, we use the entropy of the selected feature subset as a task-agnostic heuristic for its information capacity. Because multivariate joint entropy is difficult to estimate and exact evaluation of all pairwise joint entropies is computationally costly, we propose a landmark-based block approximation inspired by the Nyström method. Features with high marginal entropy are selected as landmarks. Joint entropy is computed exactly for pairs involving at least one landmark, while non-landmark pairs are approximated in constant time per pair using the geometric mean of their marginal entropies. We establish deterministic bounds for the surrogate and incorporate the resulting hybrid estimator into an incremental feature selection strategy. Experiments on 25 public benchmark datasets show that the proposed method achieves the lowest average rank across eight evaluation metrics. It exhibits significant advantages over most baselines on information-preservation metrics while maintaining competitive classification and clustering performance.

Keywords: unsupervised feature selection; joint entropy; landmark-based approximation; Nyström method; pattern classification; computational efficiency

MSC: 68T05

1. Introduction

The rapid advancement of modern information technologies and data acquisition systems has led to the generation of massive high-dimensional datasets across a wide range of industrial and research domains [1–3]. Although these datasets contain rich information, their enormous number of features, often ranging from thousands to tens of thousands, can give rise to the curse of dimensionality [4,5]. Irrelevant, redundant, or noisy features can hide structure in the data. Such features also add storage and processing costs and may weaken the efficiency and generalization of downstream models. This problem has been approached from several perspectives. Supervised information-theoretic methods weigh relevance against redundancy, complementarity, and interaction with respect to a class variable [6–8]; fuzzy and granular-computing methods model the uncertainty that noisy or vague measurements introduce into feature significance [9]; and hybrid filter–wrapper frameworks have been developed for high-dimensional, low-sample-size biomedical data [10]. These approaches, however, require class labels during selection. The

Check for updates

Academic Editor: Renato Bruni

Received: 15 September 2026
Revised: 30 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

practical question is which features can be removed without losing information needed to represent the data when such labels are unavailable.

Feature selection has been widely studied as a representative dimensionality reduction technique for alleviating the challenges associated with high-dimensional data [11–15]. Among its various paradigms, unsupervised feature selection (UFS) addresses scenarios in which user-provided information, such as class labels, is unavailable during feature selection or the downstream task has not yet been determined at the time of analysis. UFS therefore has to judge feature quality from the data itself, before a particular predictive task is specified. Graph preservation, sparse representation, reconstruction, clustering, and pseudo-labeling offer different ways to select features without labels.

Existing UFS methods can preserve the local geometric or global representational structure of data, but many depend on a particular graph, clustering structure, or reconstruction model in their objective functions. They preserve the structures built into their objectives, but those structures do not directly measure information retained for future classification tasks. We therefore focus on information retained by the data itself: at selection time, neither the future task nor its class variable is available.

Information theory offers a basis for designing a UFS criterion independent of a specific class variable because it quantifies both the information contained in features and the dependencies among variables [16–18]. We use $H(S)$ as a task-agnostic heuristic for the information capacity of the selected subset. Fano's inequality [19] motivates this heuristic: a subset with low entropy cannot support accurate prediction of a class variable with high entropy. Direct optimization of $H(S)$ is difficult for two reasons. Searching all $\binom{d}{k}$ subsets quickly becomes impractical as d grows, while estimating the entropy of a larger selected subset requires increasingly sparse joint-state counts. We handle the search by adding one feature at a time. For each candidate, f, the score sums $H(s, f)$ over the features s already in S, so it does not require an estimate of the full distribution of $S \cup \{f\}$. The identity $H(s, f) = H(s) + H(f) - I(s; f)$ shows how this score combines the candidate's entropy with its dependence on selected features. Pairwise evaluation nevertheless remains costly: computing all pairs empirically takes $O(d^2 N)$ time. Restricting evaluation to pairs used by incremental search reduces that cost, but it can still be substantial when N, d, or k is large, particularly for datasets with thousands of features.

To reduce the remaining pairwise cost, we use a landmark-based approximation inspired by the Nyström principle. We rank features by marginal entropy and choose the highest-ranked ones as landmarks. A pair containing at least one landmark retains its empirically computed joint entropy. For two non-landmark features, we instead use the geometric mean of their marginal entropies. This division limits empirical pairwise calculations to landmark-related pairs while allowing the other entries to be obtained from quantities already computed for individual features. Compared with existing UFS methods, the proposed approach has four practical advantages. First, its criterion is task-agnostic: it depends only on the feature distribution and favors subsets with large information capacity, which, by Fano's inequality, is a prerequisite for accurate prediction of a class variable. Second, it requires no prior structure. No similarity graph, neighborhood size, reconstruction model, or number of clusters c is needed, whereas seven of the eight compared methods require c, which in practice is set to the number of classes. Third, selection is deterministic and has a single parameter, the landmark ratio α, avoiding iterative optimization and sensitivity to initialization. Fourth, the pairwise score is interpretable: by $H(s, f) = H(s) + H(f) - I(s; f)$, it rewards informative candidates and penalizes redundancy with already-selected features, and the landmark-based approximation reduces the number of empirically evaluated pairs to a fraction of approximately $2\alpha - \alpha^2$.

The main contributions of this study are summarized as follows.

- We use the entropy of the selected subset as a task-agnostic information-capacity heuristic for feature selection when the downstream class variable is unknown, motivated by the necessary condition for accurate prediction implied by Fano's inequality.
- We construct an incremental candidate score from pairwise joint entropies, avoiding direct estimation of the selected subset's high-dimensional joint entropy.
- We form a hybrid pairwise estimator by selecting high-marginal-entropy landmarks, evaluating pairs involving them empirically, and applying a geometric-mean surrogate to non-landmark pairs. We also derive bounds for the surrogate and establish its conservative underestimation property.
- We compare the method with existing UFS approaches on public datasets. The evaluation covers classification, clustering, information-preservation measures, and feature-selection time; class labels are not used during selection.

2. Related Work

UFS selects features without access to class labels, but existing methods differ in what they seek to retain and in what they must assume about the data. We review them along three axes: structure-preserving UFS, information-theoretic feature selection, and landmark- or sampling-based approximation of large matrices. Table 1 summarizes the methods used as baselines in this study together with the proposed method.

Table 1. Characteristics of the compared UFS methods. "Uses c": the method requires the number of clusters or projection dimension, which was set to the number of classes in our experiments. "Params": regularization or balancing parameters fixed in Section 4.1 (excluding iteration limits). ✓: required; –: not required.

Method	Criterion/Retained Property	Uses c	Iterative Opt.	Params
UDFS	Local discriminant structure, $\ell_{2,1}$ sparsity	✓	✓	2
CNAFS	Convex NMF + adaptive graph	✓	✓	4
EGCFS	Adaptive graph with constraints	✓	✓	2
RAFG	Robust regression + adaptive graph	✓	✓	3 (α, β, k)
BLSFS	Spectral pseudo-labels, bi-level	✓	✓	3 (β, γ, k)
BGLR	Anchor bipartite graph + low redundancy	✓	✓	3 (h, β, η)
GDFS	Local/global graph discrepancy	✓	✓	4
SDAE	Autoencoder reconstruction	–	✓	3
Proposed	Subset entropy $H(S)$ (information capacity)	–	–	1 (α)

2.1. Structure-Preserving Unsupervised Feature Selection

Graph-based UFS represents relationships among samples as a graph and selects features according to how well they retain those relationships. Early methods such as the Laplacian score and MCFS relied on similarity graphs fixed in advance [20,21]; the resulting graph, however, changes with the neighborhood size, similarity function, noise, and outliers. More recent methods learn the graph while selecting features. CNAFS couples convex nonnegative matrix factorization with adaptive graph learning [22]; EGCFS incorporates graph learning and graph constraints into the selection objective [23]; RAFG combines robust regression with adaptive ℓ_1-norm graph embedding to limit the influence of outliers [24]; BGLR represents samples through an anchor-based bipartite graph and adds low-redundancy regularization [25]; and GDFS learns discrepancies between local and global graph structures across views [26]. GroupFS extends this line by learning groups of related features under graph-smoothness and group-sparsity constraints [27].

A second line infers separation patterns from the data. UDFS combines discriminant analysis with $\ell_{2,1}$-norm regularization [28], and BLSFS uses pseudo-labels from spectral clustering in a bi-level framework [29]. Deep-learning approaches such as SDAE place

a selective layer in an autoencoder and rank features by their contribution to reconstruction [30].

These methods share three practical dependencies. First, the criterion is defined by a model of the data—a graph, a clustering, or a reconstruction—so features are rewarded for preserving that model, not for the information they carry. Second, most of them require the number of clusters or projection dimension c; in common practice, including the experiments reported here, c is set to the true number of classes, so a quantity derived from the labels enters the selection step. Third, they involve several regularization parameters and iterative optimization whose outcome may depend on initialization (Table 1). The proposed criterion avoids all three: it is defined directly on the empirical distribution of the features, requires no c, and has a single parameter, the landmark ratio α.

2.2. Information-Theoretic Feature Selection

Information-theoretic criteria quantify the information in features and the dependencies among them. In the supervised setting, mRMR balances relevance to the class against pairwise redundancy [6], and Brown et al. showed that many such criteria are low-order approximations of conditional likelihood maximization [7]. R2CI extends this view by jointly modeling four kinds of correlation—class relevance, redundancy, complementarity, and interaction—and combining them in a single selection criterion [8]. Fuzzy and multi-granularity extensions replace Shannon entropy with fuzzy or granular uncertainty measures so that feature significance and interaction remain informative under noise and vagueness [9]. All of these criteria use the class variable C: relevance, complementarity, and interaction are defined through $I(f;C)$ or $I(f;C \mid s)$ and cannot be evaluated when labels are absent.

Unsupervised information-theoretic methods must therefore rely on feature–feature quantities alone. Existing approaches measure representativeness through entropy and relevance among features [16], maximize the information shared between selected and unselected features [31], rank features by pairwise dependence [32], or optimize a fractional objective built from pairwise fuzzy mutual information with a projection neural network [33]. Our work differs from these in two respects. Its objective is the subset entropy $H(S)$, motivated by Fano's inequality as a class-independent surrogate for information that would be available to any future class variable, rather than a heuristic combination of relevance and redundancy terms. Through $H(s,f) = H(s) + H(f) - I(s;f)$, the pairwise score rewards informative candidates and penalizes redundancy with already-selected features without referring to C. Its contribution on the computational side is not a new correlation term but a reduction in the $O(d^2N)$ cost of the pairwise joint-entropy matrix that such criteria require. Both direct multivariate entropy estimation, which becomes unreliable as the joint state space grows [34], and exhaustive pairwise evaluation are thereby avoided.

2.3. Landmark- and Sampling-Based Matrix Approximation

The Nyström method approximates a positive semidefinite kernel matrix from a subset of its columns: with landmark set L and remainder R, the unseen block is reconstructed as $K_{RR} \approx K_{RL}K_{LL}^{+}K_{LR}$ [35]. Its error is controlled when K is positive semidefinite and the landmarks are sampled uniformly, by column norms, or by leverage scores [36,37]. Related ideas appear in feature selection as column subset selection, which chooses columns that span the data matrix [38], and as representative-feature selection by feature similarity [39]. In UFS, landmarks are most often samples rather than features: anchor graphs replace the $N \times N$ similarity graph with an $N \times h$ bipartite graph between samples and h anchors [40], an idea used by BGLR [25].

The proposed approximation borrows the landmark idea but differs from these methods in three ways. First, landmarks are features chosen deterministically by marginal entropy rather than sampled, because high-entropy features are those most likely to enter the selected subset. Second, the pairwise joint-entropy matrix is not guaranteed to be positive semidefinite, and $\mathbf{Q}_{LL}$ can be ill-conditioned, so the Nyström reconstruction and its error guarantees do not carry over directly. Third, instead of reconstructing $\mathbf{Q}_{RR}$ from $\mathbf{Q}_{LL}$ and $\mathbf{Q}_{LR}$, each non-landmark entry is replaced by the rank-one surrogate $\sqrt{H(f_i)H(f_j)}$, which preserves the diagonal, never exceeds the true joint entropy (Lemma 2), and is separable, so that the non-landmark block need not be stored. For these reasons, we describe the method as a landmark-based block approximation inspired by the Nyström principle rather than as a Nyström reconstruction; an empirical comparison with the standard Nyström reconstruction is reported in Section 4.2.

3. Proposed Method

This section presents the proposed UFS method for preserving data information without using class labels. We first motivate the entropy-based selection criterion and describe the incremental search strategy. We then introduce the landmark-based approximation of pairwise joint entropy and analyze the overall algorithm and its computational complexity.

3.1. Motivation

In many practical settings, features must be selected before the downstream task is fixed. A dataset may be compressed or released once and later used for several prediction problems; labels may arrive only after the data have been collected; or the same reduced representation may serve multiple analyses. In these settings, the selected subset should retain information that any future task could exploit, rather than a structure tailored to one model. Existing UFS criteria are defined through such models—a similarity graph, a clustering into c groups, or a reconstruction network—and therefore reward features for preserving the model rather than for the information they carry. We instead seek a criterion that is defined only by the empirical distribution of the features and can be optimized without graph construction, cluster numbers, or iterative optimization. Subset entropy meets these requirements. Fano's inequality does not justify it as a criterion for arbitrary tasks, but it shows why a subset with low entropy is unsuitable for any task, as described next. Let $X \in \mathbb{R}^{N\times d}$ denote a data matrix consisting of N samples and d features, and let $F = \{f_1, f_2, \ldots, f_d\}$ denote the complete feature set. The objective of UFS is to select a subset $S \subset F$ containing k informative features without using class labels. We denote by C the class variable of a downstream classification task that is unknown at the time of feature selection. The variable C is not used by the proposed feature-selection procedure; it is introduced only to motivate the entropy criterion.

Let P_e denote the minimum probability of error when predicting C from the selected feature subset S. Fano's inequality establishes the following relationship between P_e and the conditional entropy $H(C \mid S)$. Throughout this section, logarithms are taken to base 2, and entropy is measured in bits.

$$H(C \mid S) \leq h(P_e) + P_e \log(|\mathcal{C}| - 1), \tag{1}$$

where $\mathcal{C}$ is the set of possible classes of C, and $h(P_e)$ denotes the binary entropy function. Using $h(P_e) \leq 1$, a simplified lower bound on the prediction error can be obtained as

$$P_e \geq \frac{H(C \mid S) - 1}{\log |\mathcal{C}|}. \tag{2}$$

Therefore, accurate prediction of C from S requires the conditional entropy $H(C \mid S)$ to be small. The conditional entropy can be expressed as

$$H(C \mid S) = H(C, S) - H(S). \tag{3}$$

In UFS, however, the class variable C is unavailable, and thus, $H(C \mid S)$, $H(C, S)$, and $I(C; S)$ cannot be computed or optimized directly. A class-independent bound follows from $I(C; S) \leq H(S)$:

$$H(C \mid S) = H(C) - I(C; S) \geq H(C) - H(S). \tag{4}$$

Combining Equation (4) with Equation (2) gives

$$P_e \geq \frac{H(C) - H(S) - 1}{\log |\mathcal{C}|}. \tag{5}$$

The right-hand side depends on the selected features only through $H(S)$: for every class variable C, a subset with small entropy forces a large lower bound on the error, and increasing $H(S)$ lowers this bound. In particular, error-free prediction ($P_e = 0$, hence $H(C \mid S) = 0$) requires $H(S) \geq H(C)$. This condition is necessary but not sufficient; a lower bound does not by itself determine the actual error, and high-entropy noise, for example, increases $H(S)$ without carrying information about C. We therefore adopt $H(S)$ as a task-agnostic heuristic for information capacity, motivated by this necessary condition, and evaluate its usefulness empirically in Section 4.

For a subset, S, its joint entropy is

$$H(S) = -\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \log p(\mathbf{s}). \tag{6}$$

Here, $\mathcal{S}$ is the joint state space of the selected features, and $p(\mathbf{s})$ is the probability of state $\mathbf{s}$. Given the current subset, S, we evaluate an unselected feature, f, through $J(S, f) = H(S \cup \{f\})$. The next feature is chosen by

$$f^+ = \arg\max_{f \in F \setminus S} J(S, f). \tag{7}$$

This rule compares the information in augmented subsets without using a class label or classifier. It is a heuristic for unknown downstream tasks, not a guarantee that the selected features will be relevant to any particular class variable. Directly evaluating Equation (7) remains difficult because estimating the high-dimensional joint entropy $H(S, f^+)$ becomes statistically unreliable as the size of S increases.

3.2. Incremental Search

Finding a globally optimal subset of size k under the joint-entropy objective would require comparing up to $\binom{d}{k}$ candidate subsets by exhaustive search. We therefore adopt a greedy incremental strategy that selects one feature at a time, although this strategy does not guarantee a globally optimal subset. Starting with $S = \emptyset$, the search adds one feature at each step. Given the current selected feature subset, S, the feature f^+ to be added at the next step is selected from the remaining candidate features by maximizing the joint entropy of the augmented subset as

$$f^+ = \arg\max_{f \in F \setminus S} H(S \cup \{f\}). \tag{8}$$

The selected feature is then added to the current subset, and this procedure is repeated until the predefined number of features has been selected as

$$S \leftarrow S \cup \{f^+\}. \tag{9}$$

Nevertheless, directly evaluating Equation (8) requires repeatedly estimating high-dimensional joint entropy for every candidate feature. As the number of selected features increases, the number of possible joint states grows exponentially, making it difficult to reliably estimate $H(S \cup \{f\})$ from a limited number of samples. To alleviate this difficulty, we approximate the high-dimensional joint entropy using low-order pairwise joint-entropy terms.

Let the currently selected subset be $S = \{f_1, f_2, \ldots, f_t\}$, where $t \geq 1$. Inspired by the order-two entropy-approximation approach [18], we define a pairwise surrogate for the joint entropy of the augmented subset as

$$\widehat{H}_2(S \cup \{f\}) = \frac{1}{t}\left(\sum_{\{u,v\} \subseteq S} H(u,v) + \sum_{s \in S} H(s,f)\right). \tag{10}$$

The first sum contains only features already selected in S and is therefore constant with respect to the candidate feature f. At a fixed iteration, $1/t$ does not change across candidates. We can therefore rank them using only the terms that contain f:

$$\widetilde{J}(S,f) = \sum_{s \in S} H(s,f). \tag{11}$$

Accordingly, the feature added to the current subset is selected as

$$f^+ = \arg\max_{f \in F \setminus S} \widetilde{J}(S,f) = \arg\max_{f \in F \setminus S} \sum_{s \in S} H(s,f). \tag{12}$$

The resulting pairwise score reflects two properties of a candidate: its entropy and its dependence on features already in S. Specifically, because $H(s,f) = H(s) + H(f) - I(s;f)$, maximizing the sum of pairwise joint entropies favors a candidate with high marginal entropy while penalizing high mutual information with the features already selected. Candidates with higher marginal entropy and lower mutual information with S consequently receive higher scores. This remains a ranking surrogate: interactions involving three or more features may make its ordering differ from that of exact joint entropy.

Because $S = \emptyset$ at initialization, the pairwise score in Equation (11) cannot be evaluated. The first feature is therefore selected as the one with the largest total pairwise joint entropy with all other features,

$$f_{\text{init}} = \arg\max_{f \in F} \sum_{g \in F \setminus \{f\}} H(f,g), \tag{13}$$

which applies the score of Equation (12) with all remaining features in place of S. In practice, the pairwise entropies in Equation (13) are taken from the approximation described in Section 3.3. Subsequent features are iteratively added according to Equation (12). The pairwise score removes the need to estimate the full joint distribution of the selected features. Computing $H(s,f)$ for many pairs still requires repeated probability estimates and can dominate the work on high-dimensional datasets.

3.3. Approximation of Pairwise Joint Entropy

The incremental search strategy introduced in the previous subsection replaces the direct estimation of high-dimensional joint entropy with pairwise joint entropy between

the currently selected features and each candidate feature. Nevertheless, constructing the complete pairwise joint-entropy matrix for a dataset containing N samples and d features requires $O(d^2N)$ time. With thousands of features, pairwise joint-entropy estimation can dominate the cost of feature selection.

To reduce this computational cost, we select landmark features according to their marginal entropies and use a rank-one surrogate for pairs of non-landmark features. The design borrows the landmark idea of the Nyström method but is not a Nyström reconstruction: the pairwise joint-entropy matrix is not guaranteed to be positive semidefinite, so we do not reconstruct the non-landmark block from the landmark columns and instead fill it with a rank-one surrogate built from marginal entropies. We therefore refer to the procedure as a landmark-based block approximation. The key idea is to directly compute joint entropy for pairs involving high-marginal-entropy features while applying a low-cost surrogate to pairs consisting only of non-landmark features. To this end, the marginal entropy of each feature in the complete feature set $F = \{f_1, f_2, \ldots, f_d\}$ is first computed as

$$H(f_i) = -\sum_{x \in \mathcal{X}_i} p_i(x) \log p_i(x). \tag{14}$$

Here, $\mathcal{X}_i$ denotes the state space of feature f_i, and $p_i(x)$ denotes the probability of observing state x.

Definition 1. *All features are sorted in descending order according to their marginal entropies as*

$$H(f_{(1)}) \geq H(f_{(2)}) \geq \cdots \geq H(f_{(d)}). \tag{15}$$

Let α denote the landmark ratio, and let $m = \max(1, \mathrm{round}(\alpha d))$ be the number of landmark features. The landmark set L and the non-landmark set R are, respectively, defined as

$$L = \left\{ f_{(1)}, f_{(2)}, \ldots, f_{(m)} \right\}, \qquad R = F \setminus L, \qquad m = \max\big(1, \mathrm{round}(\alpha d)\big). \tag{16}$$

We use $0 < \alpha < 1$ to set the number of landmarks, which also sets the size of the exact-evaluation region.

A larger α shifts more pairs to exact evaluation and increases computation; a smaller α shifts more pairs to the surrogate. The choice therefore trades computation against potential approximation loss.

Let $\mathbf{Q} \in \mathbb{R}^{d \times d}$ denote the pairwise joint-entropy matrix of the complete feature set. Each element of $\mathbf{Q}$ is defined as

$$Q_{ij} = \begin{cases} H(f_i), & \text{if } i = j, \\ H(f_i, f_j), & \text{if } i \neq j. \end{cases} \tag{17}$$

After simultaneously permuting the rows and columns of $\mathbf{Q}$, so that the landmark features precede the non-landmark features, the matrix can be partitioned according to Definition 1 as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{LL} & \mathbf{Q}_{LR} \\ \mathbf{Q}_{RL} & \mathbf{Q}_{RR} \end{bmatrix}. \tag{18}$$

In the partitioned matrix, $\mathbf{Q}_{LL}$ contains landmark–landmark pairs, while $\mathbf{Q}_{LR}$ and $\mathbf{Q}_{RL}$ contain pairs with one landmark. The block $\mathbf{Q}_{RR}$ contains pairs in which both features are non-landmarks. We directly compute empirical joint entropy for the blocks $\mathbf{Q}_{LL}$, $\mathbf{Q}_{LR}$, and $\mathbf{Q}_{RL}$, while applying a low-cost approximation only to $\mathbf{Q}_{RR}$.

The surrogate for the non-landmark block is constructed from a nonnegative rank-one matrix that preserves the marginal entropies on its diagonal.

Lemma 1. *Let* $\widetilde{\mathbf{Q}} = \mathbf{u}\mathbf{u}^\top$*, where* $u_i \geq 0$ *for every feature* f_i*, and suppose that its diagonal entries satisfy* $\widetilde{Q}_{ii} = H(f_i)$*. Then, its off-diagonal entries are*

$$\widetilde{Q}_{ij} = \sqrt{H(f_i)H(f_j)}. \tag{19}$$

Proof. Since $\widetilde{Q}_{ii} = u_i^2 = H(f_i)$ and $u_i \geq 0$, we have $u_i = \sqrt{H(f_i)}$. Therefore,

$$\widetilde{Q}_{ij} = u_i u_j = \sqrt{H(f_i)H(f_j)}, \tag{20}$$

as stated in Equation (19). □

We denote this off-diagonal surrogate by $\widetilde{H}(f_i, f_j) = \widetilde{Q}_{ij}$. Based on Lemma 1, we define the proposed hybrid pairwise joint-entropy estimator as follows.

Definition 2. *For any feature pair,* $f_i, f_j \in F$*, the hybrid estimator* $\widehat{H}(f_i, f_j)$ *is defined as*

$$\widehat{H}(f_i, f_j) = \begin{cases} H(f_i, f_j), & f_i \in L \text{ or } f_j \in L, \\ \widetilde{H}(f_i, f_j), & f_i \in R \text{ and } f_j \in R, \end{cases} \tag{21}$$

The estimator in Definition 2 preserves the exact joint entropy for feature pairs containing at least one landmark and applies the geometric-mean surrogate only when both features are non-landmarks. Because the surrogate depends only on the marginal entropies of the two features, it can be evaluated in constant time for each feature pair once all marginal entropies have been computed. Accordingly, the incremental score in Equation (11) is evaluated using the hybrid estimator as

$$\widehat{\widetilde{J}}(S, f) = \sum_{s \in S} \widehat{H}(s, f). \tag{22}$$

The following lemma establishes the ranges of the true pairwise joint entropy and the geometric-mean surrogate.

Lemma 2. *For any two features,* f_i *and* f_j*, the true joint entropy satisfies*

$$\max\{H(f_i), H(f_j)\} \leq H(f_i, f_j) \leq H(f_i) + H(f_j). \tag{23}$$

For positive marginal entropies, the geometric-mean surrogate satisfies the arithmetic–geometric–harmonic mean relation

$$\frac{2H(f_i)H(f_j)}{H(f_i) + H(f_j)} \leq \widetilde{H}(f_i, f_j) \leq \frac{H(f_i) + H(f_j)}{2}. \tag{24}$$

The corresponding inequalities for zero marginal entropy follow by continuity. In particular, the geometric mean does not exceed the larger of the two marginal entropies, and therefore,

$$\widetilde{H}(f_i, f_j) \leq \max\{H(f_i), H(f_j)\} \leq H(f_i, f_j). \tag{25}$$

Proof. The joint entropy can be expressed as

$$H(f_i, f_j) = H(f_i) + H(f_j \mid f_i). \tag{26}$$

Since $0 \le H(f_j \mid f_i) \le H(f_j)$, we have $H(f_i) \le H(f_i, f_j) \le H(f_i) + H(f_j)$. By symmetry, $H(f_j) \le H(f_i, f_j)$, which establishes Equation (23). Equation (24) follows from the arithmetic–geometric–harmonic mean inequality. Furthermore,

$$\sqrt{H(f_i)H(f_j)} \le \max\{H(f_i), H(f_j)\} \le H(f_i, f_j), \tag{27}$$

which establishes Equation (25). □

Lemma 2 shows that the geometric-mean surrogate conservatively underestimates rather than overestimates the true joint entropy. The following theorem provides an absolute approximation-error bound for the non-landmark region in which the surrogate is applied.

Theorem 1. *Let $m = \max(1, \mathrm{round}(\alpha d))$ be the number of landmarks selected according to Definition 1, and assume that the non-landmark set R is nonempty. Define the maximum marginal entropy among the non-landmark features as*

$$\tau = \max_{f \in R} H(f) = H\left(f_{(m+1)}\right). \tag{28}$$

Then, for every non-landmark feature pair $f_i, f_j \in R$, the absolute approximation error of the geometric-mean surrogate satisfies

$$0 \le H(f_i, f_j) - \widetilde{H}(f_i, f_j) \le \tau. \tag{29}$$

Proof. The conservative underestimation property in Lemma 2 gives

$$H(f_i, f_j) - \widetilde{H}(f_i, f_j) \ge 0. \tag{30}$$

Using the upper bound of the true joint entropy, we further obtain

$$\begin{aligned} H(f_i, f_j) - \widetilde{H}(f_i, f_j) &\le H(f_i) + H(f_j) - \sqrt{H(f_i)H(f_j)} \\ &\le \max\{H(f_i), H(f_j)\}. \end{aligned} \tag{31}$$

The second inequality follows because the minimum of two nonnegative values does not exceed their geometric mean. By Definition 1, $f_i, f_j \in R$ implies $H(f_i) \le \tau$ and $H(f_j) \le \tau$. Hence,

$$\max\{H(f_i), H(f_j)\} \le \tau. \tag{32}$$

Combining these inequalities proves Equation (29). □

For each non-landmark pair, Theorem 1 bounds the surrogate error when both entropies are evaluated under the same probability distribution. Finite-sample entropy estimation error lies outside this approximation bound. The possible absolute gap is larger for pairs whose two marginal entropies are high. Such features are preferentially included in the landmark set, and their pairwise joint entropies are therefore evaluated exactly. We apply the surrogate only to non-landmark pairs; there, both marginal entropies are at most τ, which gives the absolute-error threshold in Theorem 1. A bounded entropy error does not imply that the exact and approximate scores order candidate features identically. If two candidates have close scores, a small pairwise error can still reverse their order. The proposed hybrid estimator preserves exact joint entropy in the regions associated with landmark features while reducing repeated probability-distribution estimation in the non-landmark region.

3.4. Overall Algorithm and Computational Complexity

We use the hybrid estimator from Section 3.3 to supply the pairwise scores for incremental selection. First, the marginal entropy of each feature is computed, and the landmark set L and non-landmark set R are constructed according to Definition 1. The hybrid estimator in Definition 2 is then applied to all feature pairs to construct the symmetric matrix $\widehat{\mathbf{Q}}$. Directly computed empirical joint entropy is used for pairs containing at least one landmark, whereas the geometric-mean surrogate is used when both features are non-landmarks.

The first selected feature is f_{init}, the feature with the largest off-diagonal row sum of $\widehat{\mathbf{Q}}$, as specified in Equation (13). For each unselected feature $f_i \notin S$, its cumulative score is then defined as

$$r_i = \sum_{f_s \in S} \widehat{H}(f_i, f_s). \tag{33}$$

The feature having the largest cumulative pairwise joint entropy with the current selected subset is selected as

$$f^+ = \arg\max_{f_i \in F \setminus S} r_i. \tag{34}$$

After selecting f^+, the selected subset is updated as $S \leftarrow S \cup \{f^+\}$. Only $\widehat{H}(f_i, f^+)$ is added to the cumulative score of each remaining candidate feature. Therefore, pairwise contributions accumulated in previous iterations do not need to be recomputed. This procedure is repeated until $|S| = k$, and the complete procedure is summarized in Algorithm 1.

Algorithm 1 Entropy -based feature selection with landmark-based block approximation

Input: Data matrix $X \in \mathbb{R}^{N \times d}$, target number of features k, and landmark ratio α
Output: Selected feature subset S

Compute $H(f_i)$ for every $f_i \in F$
Sort the features in descending order of $H(f_i)$
Set $m = \max(1, \text{round}(\alpha d))$
Define $L = \{f_{(1)}, \ldots, f_{(m)}\}$ and $R = F \setminus L$
for $i = 1$ to d **do**
    for $j = i$ to d **do**
        if $i = j$ **then**
            $\widehat{Q}_{ij} \leftarrow H(f_i)$
        else if $f_i \in L$ or $f_j \in L$ **then**
            $\widehat{Q}_{ij} \leftarrow H(f_i, f_j)$
        else
            $\widehat{Q}_{ij} \leftarrow \sqrt{H(f_i)H(f_j)}$
        end if
        $\widehat{Q}_{ji} \leftarrow \widehat{Q}_{ij}$
    end for
end for
$f_{\text{init}} \leftarrow \arg\max_{f_i \in F} \sum_{j \neq i} \widehat{Q}_{ij}$
$S \leftarrow \{f_{\text{init}}\}$
$r_i \leftarrow \widehat{H}(f_i, f_{\text{init}})$ for every $f_i \in F \setminus S$
while $|S| < k$ **do**
    $f^+ \leftarrow \arg\max_{f_i \in F \setminus S} r_i$
    $S \leftarrow S \cup \{f^+\}$
    $r_i \leftarrow r_i + \widehat{H}(f_i, f^+)$ for every $f_i \in F \setminus S$
end while
    return S

We next analyze the time complexity of the proposed method. Computing the marginal entropy of a single feature requires processing N samples. Therefore, computing the marginal entropies of all d features requires $O(dN)$ time. Sorting the d features according to their marginal entropies requires an additional $O(d \log d)$ time. Let $m = \max(1, \mathrm{round}(\alpha d))$ denote the number of landmark features. The total number of unordered feature pairs is $\binom{d}{2}$, whereas the number of pairs in which both features are non-landmarks is $\binom{d-m}{2}$. Therefore, the number of feature pairs M_{exact} for which the exact joint entropy is computed is

$$\begin{aligned} M_{\mathrm{exact}} &= \binom{d}{2} - \binom{d-m}{2} \\ &= \frac{m(2d-m-1)}{2}. \end{aligned} \tag{35}$$

Because computing the exact joint entropy of a feature pair requires $O(N)$ time, the total computational cost of the exact region is

$$O(M_{\mathrm{exact}} N) = O\Big(\alpha(2-\alpha)d^2 N\Big). \tag{36}$$

With $m \approx \alpha d$, the pairs requiring approximation are precisely those whose two features lie outside the landmark set. Their number is

$$M_{\mathrm{approx}} = \binom{d-m}{2}. \tag{37}$$

The marginal entropies have already been computed, so evaluating the geometric mean for one such pair takes $O(1)$ time. The approximate block therefore costs

$$O\big(M_{\mathrm{approx}}\big) = O\Big((1-\alpha)^2 d^2\Big). \tag{38}$$

Selection from the completed matrix uses running scores. At each of at most k steps, finding the largest score and updating the remaining candidates each cost $O(d)$, giving $O(kd)$ for the search. Adding this to matrix construction yields

$$O\Big(dN + d\log d + \alpha(2-\alpha)d^2 N + (1-\alpha)^2 d^2 + kd\Big). \tag{39}$$

The full $d \times d$ matrix requires $O(d^2)$ storage. When N is sufficiently large, processing samples for the exactly evaluated pairs is generally the largest time component. A conventional approach that computes the exact joint entropy for every feature pair requires $O(d^2 N)$ time. In contrast, the proposed method performs the $O(N)$ exact computation only for feature pairs containing at least one landmark. For a sufficiently large d, the proportion of feature pairs subject to exact computation can be approximated as

$$\frac{M_{\mathrm{exact}}}{\binom{d}{2}} \approx 1 - (1-\alpha)^2 = 2\alpha - \alpha^2. \tag{40}$$

For example, when $\alpha = 0.1$, exact computation is applied to approximately 0.19 of all feature pairs. Consequently, compared with computing the exact joint entropy for every pair, the proposed method replaces repeated probability-distribution estimation with geometric-mean computation for approximately 81% of the feature pairs. For a fixed landmark ratio, the method reduces sample-based pairwise entropy computations by a constant proportion; its worst-case asymptotic time complexity nevertheless remains $O(d^2 N)$. The cost is linear in the number of samples N, and the constant multiplying $d^2 N$ is reduced from one to $2\alpha - \alpha^2$; in this sense, the method is linear-time in N with a reduced quadratic dependence

on d. Direct empirical joint-entropy computation is retained for pairs containing high-marginal-entropy landmarks.

4. Experimental Results

We evaluate the selected features from two perspectives. Classification and clustering measure their usefulness in downstream analyses, while class-related and label-independent information measures examine what the subsets retain. Results for the proposed method are compared with those of existing UFS methods on public benchmarks.

4.1. Experimental Settings

The experiments cover 25 public datasets obtained from the UCI Machine Learning Repository and Kaggle. They span bioinformatics, images, audio, and game-playing data and include both numerical and nominal features; dataset size ranges from 50 to 8124 samples and from 22 to 11,340 features.

Entropy calculations require discrete states, so we binarized each numerical feature at its own mean: values above the mean received 1 and all others received 0. Nominal features with more than two categories were one-hot encoded, whereas the remaining nominal features were retained. Although the maximum selected-feature count was set to 300, the main comparison uses the five highest-ranked features from each method. The landmark ratio was fixed at $\alpha = 0.1$ for all datasets and was not tuned; the number of landmarks is $m = \max(1, \text{round}(\alpha d))$. The sensitivity of the results to α and to the surrogate is examined in Section 4.2. The effect of feature count is examined separately up to 50 features. For each dataset and method, feature selection was repeated on 10 random 80% subsamples, and the reported means and standard deviations were calculated across these 10 runs. Table 2 summarizes the name, number of samples, number of features, data type, and application domain of each dataset.

Table 2. Summary of the datasets used in the experiments.

Dataset	$\lvert W \rvert$	$\lvert F \rvert$	Type	Domain
ALLAML	72	7129	Numerical	Biology
Arcene	100	9920	Numerical	Biology
Audiology	226	93	Nominal	Biology
Ba	1404	320	Nominal	Image
CLL.SUB.111	111	11,340	Numerical	Biology
Coil20	1440	1024	Numerical	Image
Colon	62	5994	Nominal	Biology
GLIOMA	50	4434	Numerical	Biology
Isolet	1560	617	Numerical	Audio
Leukemia	72	7070	Nominal	Biology
LSVT	126	310	Numerical	Biology
Lymphoma	96	4026	Nominal	Biology
Mushrooms	8124	98	Nominal	Biology
Nci9	60	9712	Nominal	Biology
ORL	400	1024	Numerical	Image
Promoters	105	228	Nominal	Biology
Prostate.GE	102	5966	Numerical	Biology
SCADI	70	141	Nominal	Biology
Semeion	1593	256	Nominal	Image
SPECT	265	22	Nominal	Biology
Splice	3190	287	Nominal	Biology
Tic-Tac-Toe	958	27	Nominal	Game
Umist	575	644	Numerical	Image
WarpAR10P	130	2400	Numerical	Image
WarpPIE10P	210	2420	Numerical	Image

To evaluate the performance of the proposed method, we selected eight representative UFS methods for comparison: UDFS [28], CNAFS [22], EGCFS [23], GDFS [26], BGLR [25], BLSFS [29], RAFG [24], and SDAE [30]. No additional dataset-specific hyperparameter search was conducted for the compared methods, and the same fixed parameter settings were applied across all datasets. The detailed settings of the compared methods are as follows.

- UDFS (2011) : UDFS combines discriminant analysis with $\ell_{2,1}$-norm sparsity regularization to select discriminative features while accounting for correlations among features. The two balancing parameters were both set to 1.
- CNAFS (2020): CNAFS combines convex nonnegative matrix factorization with adaptive graph learning to select features that simultaneously preserve the low-dimensional representation and local geometric structure of the data. The parameters were set to $\alpha = 0.01$, $\beta = 100$, $\lambda = 100$, and $\gamma = 100$. The additional implementation parameters were fixed at $\epsilon = 1$ and NITER $= 3$, and the number of basis vectors for nonnegative matrix factorization was set equal to the total number of input features. The target number of clusters c was set to the number of classes in each dataset.
- EGCFS (2020): EGCFS jointly learns an adaptive graph structure during feature selection and incorporates graph constraints into its optimization problem to preserve the structural relationships in the data. The two balancing parameters were set to $\alpha = 0.001$ and $\lambda = 0.1$, respectively.
- GDFS (2025): GDFS is a graph-based UFS method that selects informative features by jointly considering the geometric structure and discriminative information of the data. The balancing parameters were set to $\alpha = 10$, $\beta = 1$, $\gamma = 0.01$, and $\rho = 0.1$, and the maximum number of iterations was fixed at 20.
- BGLR (2024): BGLR employs an anchor-based bipartite graph to efficiently represent the relationships among samples and select representative features while suppressing feature redundancy. The number of anchors was set to $h = 200$, while the low-redundancy regularization parameter and graph-entropy parameter were set to $\beta = 0.01$ and $\eta = 100$, respectively. The maximum number of iterations was fixed at 100.
- BLSFS (2024): This method uses pseudo-labels from spectral clustering within a bi-level framework to connect classification-level and feature-level information. We set its balance parameter to $\beta = 1$ and its $\ell_{2,1}$-norm regularization parameter to $\gamma = 100$. The KNN graph used $k = 10$, and optimization was limited to 20 iterations.
- RAFG (2024): RAFG combines robust regression and adaptive graph learning to reduce the influence of noisy features and outliers. Its $\ell_{2,p}$-norm and adaptive graph regularization parameters were $\alpha = 1$ and $\beta = 0.1$, respectively. We used $p = 1$, five nearest neighbors ($k = 5$), and a limit of 20 iterations.
- SDAE (2024): A selective layer in SDAE's deep autoencoder identifies features through reconstruction of the full feature space. The lasso regularization rate was 10^{-5}, the learning rate was 10^{-3}, the hidden-layer dimension was 300, and training lasted 100 epochs.

For methods that require the number of clusters or a projection dimension (all compared methods except SDAE), this quantity was set to the number of classes in each dataset, following common practice. The proposed method does not require it. Apart from this quantity, class labels are reserved for evaluation, and no feature-selection method receives them during selection. We evaluate the selected subsets using Decision Tree, Naive Bayes, and K-means accuracy, together with Mutual Information (MI), Normalized Mutual Information (NMI), Pattern Discrimination Power (PDP), Entropy Ratio, and Gini Impurity. For all evaluation measures, a larger value indicates a higher-quality feature subset.

To evaluate the predictive performance of the selected features in downstream classification tasks, we employ Decision Tree and Naive Bayes classifiers, evaluated by 10-fold cross-validation. Given N_{test} samples in a test fold, classification accuracy is defined as

$$\text{Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{I}(\widehat{y}_i = y_i). \tag{41}$$

Here, y_i and $\widehat{y}_i$ denote the true and predicted class labels of the i-th sample, respectively. The indicator function $\mathbb{I}(\cdot)$ returns one if the given condition is true and zero otherwise. We calculate Decision Tree and Naive Bayes accuracy by inserting each classifier's predictions into Equation (41).

We also run K-means to examine clustering structure retained by the selected features. Let z_i denote the cluster assignment produced by K-means and let y_i denote the true class label of the i-th sample. Clustering accuracy, which accounts for the optimal correspondence between cluster labels and class labels, is defined as

$$\text{KMeans_ACC} = \frac{1}{N} \max_{\pi \in \Pi} \sum_{i=1}^{N} \mathbb{I}(y_i = \pi(z_i)). \tag{42}$$

Here, Π denotes the set of possible one-to-one mappings between cluster labels and class labels, and π represents the optimal label permutation. We find the optimal cluster-to-class correspondence with the Hungarian algorithm. The number of clusters was set to the number of classes in each dataset. For each subset, we ran K-means with 10 different initializations and evaluated its mean clustering accuracy.

MI is used to evaluate the amount of information shared between the selected feature subset and the class variable. Let S denote the joint state induced by the selected features and let Y denote the class variable. MI is defined as

$$I(S;Y) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{y \in \mathcal{Y}} p(\mathbf{s}, y) \log\left(\frac{p(\mathbf{s}, y)}{p(\mathbf{s})p(y)}\right). \tag{43}$$

Here, $\mathcal{S}$ and $\mathcal{Y}$ denote the joint state space of the selected features and the set of class labels, respectively. A larger MI value indicates that the selected feature subset shares more information with the class variable. We also report NMI because raw MI depends in part on the entropy magnitudes of the variables being compared. NMI is defined as

$$\text{NMI}(S, Y) = \frac{2\, I(S;Y)}{H(S) + H(Y)}. \tag{44}$$

NMI generally ranges from zero to one, with a larger value indicating a stronger statistical dependency between the selected feature subset and the class labels.

PDP examines how many different sample patterns the selected features produce, without using class labels. Let $\mathcal{U}(S)$ denote the set of distinct joint patterns generated by the selected feature subset S over all samples. PDP is defined as

$$\text{PDP}(S) = \frac{|\mathcal{U}(S)|}{N}. \tag{45}$$

If every sample has a distinct feature pattern, then $\text{PDP}(S) = 1$. If all samples share the same pattern, then $\text{PDP}(S) = 1/N$. A larger PDP value means that more distinct patterns appear among the samples.

The Entropy Ratio is used to measure how much of the information contained in the original feature set is preserved by the selected subset. Let F denote the complete feature set and S denote the selected feature subset. The Entropy Ratio is defined as

$$\text{Entropy Ratio}(S, F) = \frac{H(S)}{H(F)}. \tag{46}$$

Because entropy is monotonic under the inclusion of variables, $S \subseteq F$ implies $H(S) \leq H(F)$. Thus, when $H(F) > 0$, the Entropy Ratio ranges from zero to one. A larger Entropy Ratio corresponds to a larger share of the original feature set's entropy retained by S.

We use Gini Impurity to characterize how evenly the selected-feature patterns are distributed. Let $\mathcal{U}(S)$ denote the set of distinct patterns generated by the selected feature subset and let $p(\mathbf{u})$ denote the probability of observing pattern $\mathbf{u}$. Gini Impurity is defined as

$$\text{Gini Impurity}(S) = 1 - \sum_{\mathbf{u} \in \mathcal{U}(S)} p(\mathbf{u})^2. \tag{47}$$

A higher Gini Impurity indicates a less concentrated distribution across the observed patterns. Thus, a high value supports greater diversity of observed patterns; by itself, it does not establish class distinguishability. Because every feature takes at most two states after preprocessing, a subset of k features can produce at most 2^k distinct patterns, and the number of observed patterns cannot exceed the number of samples N. Hence,

$$\text{PDP}(S) \leq \frac{\min(2^k, N)}{N}, \qquad \text{Gini Impurity}(S) \leq 1 - \frac{1}{\min(2^k, N)}, \tag{48}$$

where the second bound is attained when the observed patterns are equally frequent. For the $k = 5$ features used in the main comparison, Gini Impurity therefore cannot exceed $1 - 1/32 \approx 0.969$, and PDP cannot exceed $32/N$ on datasets with more than 32 samples. Means and standard deviations are rounded to two decimal places; a standard deviation shown as 0.00 is smaller than 0.005, not exactly zero. All experiments are conducted on an Intel i5–12400F CPU (Intel Corporation, Santa Clara, CA, USA) with 82 GB RAM using MATLAB R2025b (The MathWorks, Inc., Natick, MA, USA).

Each method is summarized on each dataset by its mean over the 10 random subsamples, and the methods are compared across the 25 datasets. The Friedman test, with the Iman–Davenport correction, is first applied to the average ranks to test whether all methods perform equally. When this hypothesis is rejected, two post-hoc analyses follow at $\alpha = 0.05$: the Nemenyi test for all pairwise comparisons, with critical difference $\text{CD} = q_{0.05}\sqrt{k(k+1)/(6N)} = 3.102\sqrt{9 \cdot 10/(6 \cdot 25)} \approx 2.40$, and the Wilcoxon signed-rank test between the proposed method and each baseline, with Holm correction for the eight comparisons. Because the Nemenyi test controls the error rate over all $k(k-1)/2 = 36$ pairs, it is conservative for comparisons against a single method, so we report both. Tests are performed across datasets rather than across splits, since splits drawn from the same dataset are not independent.

4.2. Comparison Result

Tables 3–5 present the performance of the proposed and baseline methods in terms of Naive Bayes classification accuracy, Decision Tree classification accuracy, and K-means clustering accuracy, respectively. Each entry reports the mean and standard deviation over repeated runs, and the best result for each dataset is highlighted in bold. The last row reports the average rank of each method across the 25 datasets, where a lower rank indicates better overall performance. The two classifiers and K-means provide three different downstream views of the selected subsets. The proposed method achieves the best average rank for

all three evaluation measures, obtaining ranks of 2.96, 3.22, and 4.36 for Naive Bayes, Decision Tree, and K-means, respectively. For Naive Bayes classification, the proposed method achieves the best or tied-best result on 11 datasets: GLIOMA, SCADI, SPECT, Tic-Tac-Toe, Coil20, LSVT, ORL, Promoters, Splice, Umist, and WarpAR10P. Particularly clear improvements are observed on Coil20 and Promoters. Similar behavior is observed with the Decision Tree classifier, for which the proposed method achieves the best result on nine datasets and records an average rank of 3.22, compared with 4.28 for RAFG, the second-best method. Its Decision Tree results are best on GLIOMA, SCADI, SPECT, Tic-Tac-Toe, Coil20, Promoters, Splice, Umist, and WarpAR10P, with the largest margins on Coil20, Promoters, and Splice. For K-means clustering, the proposed method attains the best average rank of 4.36 and achieves the best accuracy on six datasets: SPECT, Coil20, ORL, Promoters, Splice, and WarpAR10P. A baseline may lead on a particular dataset; across the reported downstream measures, however, each baseline has a higher average rank. We did not optimize a clustering objective, yet our method had the lowest average rank for K-means accuracy. It held the same position for Naive Bayes and Decision Tree accuracy.

To examine whether the selected subsets retain class-related information, we measure MI and NMI against the class variable in Tables 6 and 7. Class labels are used for this evaluation, but not for feature selection. Each table reports means and standard deviations across repeated runs, marks the best value for each dataset in bold, and gives average method ranks in its final row. For MI, our method has an average rank of 1.56; CNAFS is next at 3.88. It also records the best or tied-best MI on 18 of the 25 datasets: GLIOMA, Prostate.GE, SCADI, SPECT, Tic-Tac-Toe, Arcene, Coil20, Colon, Leukemia, LSVT, Lymphoma, Nci9, ORL, Promoters, Splice, Umist, WarpAR10P, and WarpPIE10P. Thus, on many of the evaluated datasets, the subsets retain information about the class variable despite being selected without labels. Particularly clear improvements are observed on SCADI, Coil20, Colon, Promoters, and WarpPIE10P. Although BLSFS, EGCFS, GDFS, RAFG, and SDAE achieve the highest MI on a small number of individual datasets, their performance is less consistent across the entire benchmark collection. The NMI results show a similar overall tendency. The proposed method achieves the best average rank of 2.36, followed by GDFS with an average rank of 4.68. It obtains the best or tied-best NMI on 14 datasets: GLIOMA, SPECT, Tic-Tac-Toe, Arcene, Coil20, Colon, Leukemia, Lymphoma, Nci9, ORL, Promoters, Splice, WarpAR10P, and WarpPIE10P. The improvements on Coil20, Colon, Nci9, and Promoters are particularly notable. Some baseline methods perform better on specific datasets, such as UDFS on SCADI, EGCFS on Isolet, GDFS on Mushrooms, RAFG on Semeion, and SDAE on Audiology. These dataset-specific wins do not extend to the baselines' average ranks across all 25 datasets. Because MI depends on the entropy scales of the variables, the NMI results provide an additional normalized view of class-related information preservation. The proposed method also obtains the lowest average rank under this measure. Across these datasets, the MI and NMI results are consistent with using subset entropy as a class-independent information-preservation surrogate. In particular, the proposed method retains class-relevant information consistently across datasets despite not using class labels during feature selection.

Table 3. Comparison of Naive Bayes classification accuracy obtained by the proposed and baseline methods. Results are reported as mean ± standard deviation. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	0.59 ± 0.06	0.62 ± 0.10	0.60 ± 0.08	0.59 ± 0.07	0.63 ± 0.10	**0.78 ± 0.06**	0.60 ± 0.05	0.62 ± 0.09	0.72 ± 0.10
CLL.SUB.111	0.47 ± 0.07	0.45 ± 0.10	0.49 ± 0.09	**0.51 ± 0.07**	0.47 ± 0.07	0.48 ± 0.04	0.46 ± 0.05	**0.51 ± 0.05**	0.46 ± 0.09
GLIOMA	**0.52 ± 0.10**	0.40 ± 0.12	0.41 ± 0.08	0.30 ± 0.10	0.35 ± 0.09	0.38 ± 0.13	0.37 ± 0.07	**0.52 ± 0.10**	0.30 ± 0.09
Prostate.GE	0.62 ± 0.05	0.56 ± 0.07	0.57 ± 0.10	0.66 ± 0.09	0.59 ± 0.09	0.47 ± 0.06	0.60 ± 0.06	**0.67 ± 0.17**	0.47 ± 0.10
SCADI	**0.73 ± 0.07**	0.66 ± 0.17	0.43 ± 0.05	0.40 ± 0.10	0.58 ± 0.08	0.62 ± 0.09	0.49 ± 0.08	0.69 ± 0.12	0.41 ± 0.07
SPECT	**0.82 ± 0.04**	0.74 ± 0.05	0.78 ± 0.03	0.76 ± 0.02	0.80 ± 0.04	0.73 ± 0.05	0.75 ± 0.05	0.76 ± 0.04	0.80 ± 0.02
Tic-Tac-Toe	**0.77 ± 0.02**	0.65 ± 0.02	0.66 ± 0.02	0.63 ± 0.01	0.65 ± 0.01	0.66 ± 0.03	0.66 ± 0.03	0.66 ± 0.03	0.75 ± 0.03
Arcene	0.56 ± 0.08	0.53 ± 0.12	0.54 ± 0.06	0.54 ± 0.10	0.53 ± 0.07	0.51 ± 0.11	0.56 ± 0.04	**0.57 ± 0.08**	0.54 ± 0.07
Audiology	0.41 ± 0.06	0.25 ± 0.02	0.54 ± 0.12	0.29 ± 0.12	0.26 ± 0.02	0.23 ± 0.04	0.27 ± 0.02	0.25 ± 0.05	**0.60 ± 0.05**
Ba	0.12 ± 0.02	0.14 ± 0.05	0.12 ± 0.03	0.15 ± 0.04	**0.16 ± 0.04**	0.09 ± 0.02	0.10 ± 0.03	0.14 ± 0.04	0.09 ± 0.01
Coil20	**0.45 ± 0.03**	0.11 ± 0.03	0.33 ± 0.03	0.09 ± 0.01	0.09 ± 0.01	0.13 ± 0.03	0.25 ± 0.05	0.08 ± 0.01	0.26 ± 0.03
Colon	0.58 ± 0.11	0.59 ± 0.07	0.58 ± 0.08	0.60 ± 0.06	0.59 ± 0.07	0.53 ± 0.09	**0.61 ± 0.07**	0.58 ± 0.05	**0.61 ± 0.07**
Isolet	0.15 ± 0.02	0.11 ± 0.02	**0.18 ± 0.04**	0.14 ± 0.01	0.11 ± 0.02	0.17 ± 0.04	0.13 ± 0.02	0.12 ± 0.02	0.13 ± 0.02
Leukemia	0.60 ± 0.05	0.60 ± 0.06	0.61 ± 0.06	0.60 ± 0.05	0.62 ± 0.06	0.64 ± 0.06	0.58 ± 0.04	0.62 ± 0.07	**0.69 ± 0.11**
LSVT	**0.69 ± 0.05**	0.61 ± 0.08	0.68 ± 0.05	0.65 ± 0.03	0.68 ± 0.05	0.62 ± 0.05	**0.69 ± 0.05**	0.59 ± 0.05	0.64 ± 0.03
Lymphoma	0.44 ± 0.05	0.48 ± 0.08	0.46 ± 0.04	0.47 ± 0.06	0.46 ± 0.05	**0.52 ± 0.05**	0.43 ± 0.05	0.45 ± 0.04	0.50 ± 0.08
Mushrooms	0.86 ± 0.00	0.71 ± 0.12	0.68 ± 0.11	**0.93 ± 0.00**	0.88 ± 0.07	0.57 ± 0.06	0.63 ± 0.05	0.70 ± 0.11	0.52 ± 0.00
Nci9	0.12 ± 0.06	**0.14 ± 0.08**	0.11 ± 0.05	**0.14 ± 0.06**	0.07 ± 0.06	0.08 ± 0.07	0.09 ± 0.04	0.10 ± 0.04	0.10 ± 0.06
ORL	**0.14 ± 0.03**	**0.14 ± 0.03**	0.12 ± 0.01	**0.14 ± 0.03**	**0.14 ± 0.03**	0.11 ± 0.03	0.12 ± 0.03	0.09 ± 0.04	0.12 ± 0.01
Promoters	**0.81 ± 0.03**	0.57 ± 0.09	0.59 ± 0.08	0.60 ± 0.12	0.56 ± 0.13	0.64 ± 0.12	0.60 ± 0.12	0.64 ± 0.09	0.62 ± 0.07
Semeion	0.33 ± 0.02	0.25 ± 0.04	0.30 ± 0.03	0.27 ± 0.03	0.31 ± 0.06	0.29 ± 0.04	0.26 ± 0.03	**0.38 ± 0.05**	0.30 ± 0.05
Splice	**0.88 ± 0.00**	0.52 ± 0.02	0.53 ± 0.02	0.52 ± 0.00	0.52 ± 0.01	0.52 ± 0.01	0.53 ± 0.02	0.53 ± 0.03	0.52 ± 0.01
Umist	**0.31 ± 0.04**	0.30 ± 0.08	0.21 ± 0.05	0.24 ± 0.06	0.27 ± 0.04	0.21 ± 0.05	0.22 ± 0.03	0.19 ± 0.07	0.21 ± 0.03
WarpAR10P	**0.27 ± 0.08**	0.24 ± 0.08	0.21 ± 0.04	0.24 ± 0.10	0.16 ± 0.05	0.24 ± 0.07	0.23 ± 0.05	0.16 ± 0.04	0.23 ± 0.06
WarpPIE10P	0.33 ± 0.07	0.26 ± 0.04	0.28 ± 0.05	0.24 ± 0.03	0.31 ± 0.08	0.17 ± 0.04	0.26 ± 0.05	**0.41 ± 0.05**	0.23 ± 0.07
Avg. Rank	**2.96**	5.74	4.92	4.92	5.26	5.92	5.44	4.72	5.12

https://doi.org/10.3390/math14214057

Table 4. Comparison of Decision Tree classification accuracy obtained by the proposed and baseline methods. Results are reported as mean ± standard deviation. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	0.58 ± 0.07	0.57 ± 0.13	0.65 ± 0.08	0.58 ± 0.06	0.60 ± 0.10	**0.78 ± 0.08**	0.57 ± 0.05	0.61 ± 0.10	0.72 ± 0.07
CLL.SUB.111	0.44 ± 0.05	0.43 ± 0.08	**0.48 ± 0.08**	0.45 ± 0.07	0.47 ± 0.07	0.47 ± 0.06	0.46 ± 0.05	0.47 ± 0.06	0.45 ± 0.09
GLIOMA	**0.52 ± 0.14**	0.38 ± 0.09	0.36 ± 0.10	0.25 ± 0.09	0.38 ± 0.10	0.39 ± 0.14	0.34 ± 0.12	0.51 ± 0.11	0.30 ± 0.09
Prostate.GE	0.66 ± 0.07	0.59 ± 0.08	0.54 ± 0.09	**0.67 ± 0.10**	0.55 ± 0.11	0.49 ± 0.07	0.60 ± 0.06	0.66 ± 0.19	0.46 ± 0.10
SCADI	**0.73 ± 0.08**	0.65 ± 0.18	0.46 ± 0.08	0.40 ± 0.10	0.57 ± 0.09	0.62 ± 0.11	0.47 ± 0.08	0.70 ± 0.12	0.41 ± 0.08
SPECT	**0.82 ± 0.03**	0.79 ± 0.03	0.79 ± 0.02	0.79 ± 0.02	0.79 ± 0.03	0.80 ± 0.01	0.79 ± 0.02	0.79 ± 0.02	0.78 ± 0.03
Tic-Tac-Toe	**0.82 ± 0.02**	0.64 ± 0.02	0.67 ± 0.03	0.64 ± 0.01	0.65 ± 0.01	0.67 ± 0.02	0.67 ± 0.02	0.67 ± 0.04	0.80 ± 0.03
Arcene	0.53 ± 0.09	0.54 ± 0.08	0.50 ± 0.07	0.53 ± 0.07	0.56 ± 0.05	0.52 ± 0.11	**0.57 ± 0.04**	**0.57 ± 0.12**	0.55 ± 0.06
Audiology	0.39 ± 0.08	0.25 ± 0.02	0.53 ± 0.10	0.29 ± 0.12	0.27 ± 0.02	0.23 ± 0.04	0.27 ± 0.02	0.25 ± 0.05	**0.60 ± 0.06**
Ba	0.13 ± 0.02	0.14 ± 0.04	0.13 ± 0.03	0.15 ± 0.04	**0.16 ± 0.03**	0.08 ± 0.02	0.11 ± 0.04	0.15 ± 0.04	0.09 ± 0.01
Coil20	**0.48 ± 0.01**	0.11 ± 0.03	0.35 ± 0.04	0.09 ± 0.01	0.09 ± 0.01	0.13 ± 0.03	0.26 ± 0.05	0.09 ± 0.01	0.28 ± 0.03
Colon	0.58 ± 0.13	0.59 ± 0.08	0.51 ± 0.05	0.58 ± 0.07	0.56 ± 0.05	**0.62 ± 0.11**	0.60 ± 0.07	0.61 ± 0.09	0.61 ± 0.08
Isolet	0.15 ± 0.03	0.10 ± 0.02	**0.17 ± 0.04**	0.13 ± 0.01	0.10 ± 0.02	**0.17 ± 0.04**	0.13 ± 0.02	0.12 ± 0.03	0.12 ± 0.02
Leukemia	0.59 ± 0.09	0.57 ± 0.10	0.59 ± 0.08	0.57 ± 0.07	0.61 ± 0.08	0.63 ± 0.08	0.60 ± 0.09	0.62 ± 0.06	**0.70 ± 0.11**
LSVT	0.64 ± 0.04	0.61 ± 0.05	0.63 ± 0.06	0.66 ± 0.02	**0.67 ± 0.07**	0.60 ± 0.06	**0.67 ± 0.05**	0.64 ± 0.04	0.61 ± 0.06
Lymphoma	0.37 ± 0.05	0.41 ± 0.09	0.42 ± 0.05	0.44 ± 0.05	0.39 ± 0.07	**0.48 ± 0.05**	0.42 ± 0.06	0.43 ± 0.03	**0.48 ± 0.09**
Mushrooms	0.91 ± 0.00	0.72 ± 0.13	0.70 ± 0.12	**0.93 ± 0.00**	0.90 ± 0.07	0.57 ± 0.05	0.64 ± 0.05	0.71 ± 0.12	0.52 ± 0.00
Nci9	0.12 ± 0.06	0.12 ± 0.08	0.11 ± 0.08	**0.14 ± 0.06**	0.08 ± 0.05	0.07 ± 0.06	0.10 ± 0.05	0.12 ± 0.05	0.12 ± 0.06
ORL	0.15 ± 0.05	**0.16 ± 0.02**	0.11 ± 0.02	0.15 ± 0.04	0.13 ± 0.02	0.11 ± 0.03	0.12 ± 0.03	0.10 ± 0.06	0.12 ± 0.02
Promoters	**0.80 ± 0.06**	0.56 ± 0.10	0.54 ± 0.10	0.56 ± 0.14	0.54 ± 0.13	0.63 ± 0.11	0.58 ± 0.11	0.64 ± 0.09	0.61 ± 0.08
Semeion	0.34 ± 0.02	0.25 ± 0.04	0.30 ± 0.03	0.27 ± 0.03	0.33 ± 0.06	0.30 ± 0.03	0.27 ± 0.03	**0.39 ± 0.05**	0.31 ± 0.05
Splice	**0.88 ± 0.00**	0.53 ± 0.03	0.53 ± 0.02	0.52 ± 0.00	0.52 ± 0.01	0.52 ± 0.01	0.53 ± 0.01	0.53 ± 0.03	0.52 ± 0.01
Umist	**0.34 ± 0.04**	0.33 ± 0.08	0.23 ± 0.05	0.27 ± 0.07	0.28 ± 0.05	0.24 ± 0.05	0.25 ± 0.05	0.22 ± 0.09	0.23 ± 0.04
WarpAR10P	**0.27 ± 0.08**	0.24 ± 0.08	0.19 ± 0.05	0.22 ± 0.08	0.16 ± 0.07	0.23 ± 0.09	0.25 ± 0.05	0.16 ± 0.04	0.23 ± 0.07
WarpPIE10P	0.40 ± 0.07	0.28 ± 0.07	0.28 ± 0.05	0.24 ± 0.03	0.33 ± 0.07	0.19 ± 0.06	0.26 ± 0.07	**0.42 ± 0.07**	0.24 ± 0.08
Avg. Rank	**3.22**	5.70	5.48	5.32	5.28	5.36	5.12	4.28	5.24

Table 5. Comparison of K-means clustering accuracy obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	0.57 ± 0.02	0.56 ± 0.02	0.57 ± 0.02	0.55 ± 0.02	0.58 ± 0.04	**0.65 ± 0.04**	0.56 ± 0.02	0.61 ± 0.05	**0.65 ± 0.08**
CLL.SUB.111	0.43 ± 0.02	0.44 ± 0.02	0.43 ± 0.02	0.45 ± 0.02	0.43 ± 0.02	**0.46 ± 0.02**	0.45 ± 0.01	0.45 ± 0.02	**0.46 ± 0.02**
GLIOMA	0.45 ± 0.02	0.47 ± 0.03	0.42 ± 0.05	0.40 ± 0.04	0.42 ± 0.03	0.49 ± 0.06	0.50 ± 0.02	**0.51 ± 0.03**	0.38 ± 0.03
Prostate.GE	0.57 ± 0.02	0.57 ± 0.03	0.56 ± 0.03	**0.61 ± 0.06**	0.58 ± 0.04	0.54 ± 0.03	0.55 ± 0.03	**0.61 ± 0.07**	0.52 ± 0.01
SCADI	0.46 ± 0.04	0.61 ± 0.13	0.36 ± 0.04	0.36 ± 0.06	0.48 ± 0.07	0.56 ± 0.10	0.43 ± 0.07	**0.63 ± 0.11**	0.40 ± 0.06
SPECT	**0.58 ± 0.04**	0.56 ± 0.03	0.56 ± 0.02	0.54 ± 0.01	0.57 ± 0.02	0.56 ± 0.01	0.55 ± 0.02	0.55 ± 0.01	0.57 ± 0.02
Tic-Tac-Toe	0.56 ± 0.01	0.56 ± 0.01	0.56 ± 0.02	0.57 ± 0.01	**0.58 ± 0.01**	**0.58 ± 0.03**	0.56 ± 0.01	0.56 ± 0.01	0.56 ± 0.01
Arcene	0.55 ± 0.02	0.55 ± 0.02	0.55 ± 0.02	0.55 ± 0.03	0.55 ± 0.02	0.56 ± 0.05	**0.58 ± 0.01**	0.57 ± 0.05	0.56 ± 0.04
Audiology	0.22 ± 0.11	0.19 ± 0.09	0.28 ± 0.13	0.23 ± 0.12	0.22 ± 0.09	0.20 ± 0.08	0.23 ± 0.07	0.19 ± 0.10	**0.37 ± 0.17**
Ba	0.17 ± 0.01	0.18 ± 0.04	0.17 ± 0.02	0.19 ± 0.04	**0.20 ± 0.03**	0.13 ± 0.01	0.15 ± 0.03	0.19 ± 0.04	0.14 ± 0.01
Coil20	**0.41 ± 0.02**	0.13 ± 0.03	0.32 ± 0.02	0.12 ± 0.00	0.12 ± 0.00	0.14 ± 0.02	0.26 ± 0.05	0.11 ± 0.00	0.28 ± 0.02
Colon	0.56 ± 0.01	0.57 ± 0.02	0.55 ± 0.02	0.57 ± 0.02	0.58 ± 0.03	0.56 ± 0.03	0.58 ± 0.02	0.56 ± 0.03	**0.61 ± 0.05**
Isolet	0.17 ± 0.02	0.14 ± 0.02	**0.19 ± 0.03**	0.17 ± 0.01	0.14 ± 0.02	**0.19 ± 0.04**	0.15 ± 0.02	0.15 ± 0.02	0.15 ± 0.01
Leukemia	0.56 ± 0.02	0.56 ± 0.02	0.57 ± 0.05	0.58 ± 0.03	0.57 ± 0.02	**0.65 ± 0.06**	0.56 ± 0.02	0.57 ± 0.03	0.64 ± 0.05
LSVT	0.58 ± 0.01	0.57 ± 0.02	0.58 ± 0.03	**0.62 ± 0.02**	0.57 ± 0.03	0.57 ± 0.04	0.59 ± 0.03	0.55 ± 0.03	0.56 ± 0.02
Lymphoma	0.28 ± 0.01	0.35 ± 0.03	0.28 ± 0.01	0.34 ± 0.04	0.27 ± 0.01	**0.39 ± 0.05**	0.28 ± 0.01	0.32 ± 0.05	0.37 ± 0.08
Mushrooms	0.63 ± 0.02	0.63 ± 0.10	0.59 ± 0.08	**0.68 ± 0.02**	0.66 ± 0.03	0.55 ± 0.05	0.53 ± 0.01	0.64 ± 0.08	0.51 ± 0.00
Nci9	0.32 ± 0.02	**0.33 ± 0.02**	0.31 ± 0.03	0.32 ± 0.03	0.30 ± 0.02	0.25 ± 0.06	0.30 ± 0.02	**0.33 ± 0.03**	0.26 ± 0.04
ORL	**0.30 ± 0.02**	0.27 ± 0.02	0.26 ± 0.02	0.28 ± 0.02	0.27 ± 0.01	0.25 ± 0.02	0.27 ± 0.02	0.21 ± 0.06	0.26 ± 0.01
Promoters	**0.69 ± 0.04**	0.57 ± 0.02	0.57 ± 0.04	0.56 ± 0.04	0.56 ± 0.04	0.59 ± 0.06	0.56 ± 0.04	0.58 ± 0.06	0.55 ± 0.02
Semeion	0.25 ± 0.01	0.23 ± 0.02	0.25 ± 0.02	0.26 ± 0.02	0.27 ± 0.04	0.27 ± 0.02	0.24 ± 0.02	**0.34 ± 0.03**	0.28 ± 0.03
Splice	**0.66 ± 0.02**	0.41 ± 0.02	0.42 ± 0.02	0.51 ± 0.04	0.52 ± 0.01	0.49 ± 0.05	0.42 ± 0.01	0.49 ± 0.04	0.41 ± 0.04
Umist	0.31 ± 0.03	**0.33 ± 0.07**	0.25 ± 0.03	0.28 ± 0.04	0.30 ± 0.04	0.27 ± 0.03	0.26 ± 0.02	0.24 ± 0.06	0.24 ± 0.01
WarpAR10P	**0.33 ± 0.03**	0.32 ± 0.04	0.29 ± 0.02	0.31 ± 0.04	0.28 ± 0.03	0.32 ± 0.03	0.31 ± 0.01	0.24 ± 0.02	0.31 ± 0.02
WarpPIE10P	0.33 ± 0.02	0.30 ± 0.02	0.30 ± 0.02	0.29 ± 0.02	0.32 ± 0.05	0.26 ± 0.03	0.29 ± 0.02	**0.42 ± 0.05**	0.27 ± 0.03
Avg. Rank	**4.36**	5.40	5.40	4.80	4.80	4.60	5.60	4.68	5.36

https://doi.org/10.3390/math14214057

Table 6. Comparison of MI values obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	0.46 ± 0.07	0.41 ± 0.08	0.42 ± 0.08	0.38 ± 0.07	0.41 ± 0.12	0.47 ± 0.11	**0.48 ± 0.05**	0.33 ± 0.11	0.34 ± 0.12
CLL.SUB.111	0.55 ± 0.09	0.51 ± 0.05	0.52 ± 0.04	**0.56 ± 0.10**	0.52 ± 0.08	0.50 ± 0.06	0.51 ± 0.06	0.54 ± 0.07	0.42 ± 0.05
GLIOMA	**1.59 ± 0.13**	1.30 ± 0.13	1.17 ± 0.33	0.99 ± 0.29	1.33 ± 0.10	0.99 ± 0.25	1.35 ± 0.06	1.17 ± 0.23	0.43 ± 0.13
Prostate.GE	**0.46 ± 0.07**	0.31 ± 0.05	0.36 ± 0.09	0.44 ± 0.12	0.34 ± 0.06	0.05 ± 0.03	0.30 ± 0.03	0.32 ± 0.28	0.09 ± 0.04
SCADI	**1.69 ± 0.12**	1.15 ± 0.45	1.05 ± 0.19	0.15 ± 0.06	0.94 ± 0.39	0.91 ± 0.28	0.78 ± 0.30	1.11 ± 0.48	0.45 ± 0.19
SPECT	**0.24 ± 0.02**	0.15 ± 0.05	0.21 ± 0.05	0.18 ± 0.02	0.21 ± 0.02	0.12 ± 0.02	0.18 ± 0.04	0.17 ± 0.05	0.22 ± 0.02
Tic-Tac-Toe	**0.43 ± 0.03**	0.03 ± 0.01	0.09 ± 0.05	0.03 ± 0.00	0.02 ± 0.01	0.05 ± 0.04	0.05 ± 0.04	0.07 ± 0.04	0.40 ± 0.04
Arcene	**0.39 ± 0.07**	0.18 ± 0.10	0.31 ± 0.08	0.26 ± 0.07	0.32 ± 0.06	0.05 ± 0.03	0.29 ± 0.04	0.13 ± 0.07	0.05 ± 0.03
Audiology	1.85 ± 0.16	0.10 ± 0.09	1.90 ± 0.27	0.33 ± 0.66	0.23 ± 0.04	0.10 ± 0.03	0.15 ± 0.07	0.14 ± 0.14	**2.17 ± 0.17**
Ba	1.48 ± 0.10	1.59 ± 0.29	1.43 ± 0.15	1.63 ± 0.26	**1.73 ± 0.24**	1.00 ± 0.14	1.15 ± 0.29	1.63 ± 0.25	1.15 ± 0.12
Coil20	**2.43 ± 0.09**	0.65 ± 0.22	1.96 ± 0.11	0.53 ± 0.03	0.53 ± 0.02	0.72 ± 0.21	1.43 ± 0.29	0.51 ± 0.02	1.76 ± 0.07
Colon	**0.58 ± 0.09**	0.30 ± 0.11	0.40 ± 0.09	0.37 ± 0.10	0.44 ± 0.10	0.25 ± 0.06	0.17 ± 0.03	0.24 ± 0.06	0.16 ± 0.06
Isolet	1.28 ± 0.14	1.28 ± 0.22	1.35 ± 0.18	**1.47 ± 0.18**	1.20 ± 0.13	1.46 ± 0.34	1.00 ± 0.12	1.07 ± 0.13	0.99 ± 0.12
Leukemia	**0.47 ± 0.09**	0.25 ± 0.07	0.37 ± 0.09	0.31 ± 0.11	0.38 ± 0.12	0.15 ± 0.08	0.15 ± 0.03	0.33 ± 0.09	0.24 ± 0.12
LSVT	**0.33 ± 0.07**	0.15 ± 0.08	0.31 ± 0.08	0.05 ± 0.02	0.26 ± 0.08	0.07 ± 0.03	**0.33 ± 0.05**	0.09 ± 0.05	0.23 ± 0.10
Lymphoma	**1.50 ± 0.09**	1.30 ± 0.10	1.30 ± 0.08	1.37 ± 0.19	1.37 ± 0.08	1.23 ± 0.16	0.74 ± 0.06	1.20 ± 0.15	1.21 ± 0.23
Mushrooms	0.69 ± 0.01	0.27 ± 0.23	0.28 ± 0.22	**0.73 ± 0.00**	0.64 ± 0.12	0.07 ± 0.06	0.11 ± 0.06	0.23 ± 0.18	0.02 ± 0.01
Nci9	**2.27 ± 0.14**	1.59 ± 0.14	1.01 ± 0.16	1.14 ± 0.33	1.04 ± 0.12	0.52 ± 0.61	0.92 ± 0.07	1.74 ± 0.23	0.64 ± 0.28
ORL	**2.76 ± 0.10**	2.45 ± 0.19	2.50 ± 0.14	2.58 ± 0.14	2.58 ± 0.10	2.34 ± 0.11	2.57 ± 0.09	1.90 ± 0.46	2.41 ± 0.11
Promoters	**0.76 ± 0.07**	0.25 ± 0.07	0.30 ± 0.08	0.26 ± 0.10	0.28 ± 0.10	0.29 ± 0.11	0.27 ± 0.07	0.32 ± 0.10	0.25 ± 0.05
Semeion	0.92 ± 0.04	0.64 ± 0.07	0.85 ± 0.11	0.71 ± 0.07	0.86 ± 0.21	0.79 ± 0.11	0.70 ± 0.14	**1.07 ± 0.17**	0.90 ± 0.15
Splice	**1.06 ± 0.01**	0.07 ± 0.08	0.13 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.15 ± 0.08	0.05 ± 0.13	0.12 ± 0.08
Umist	**2.21 ± 0.11**	2.00 ± 0.47	1.78 ± 0.16	1.66 ± 0.33	1.92 ± 0.25	1.65 ± 0.28	1.89 ± 0.13	1.43 ± 0.47	1.69 ± 0.11
WarpAR10P	**2.02 ± 0.13**	1.59 ± 0.19	1.54 ± 0.08	1.64 ± 0.15	1.60 ± 0.20	1.33 ± 0.24	1.61 ± 0.18	0.67 ± 0.09	1.43 ± 0.14
WarpPIE10P	**1.93 ± 0.14**	1.36 ± 0.22	1.55 ± 0.10	1.34 ± 0.10	1.62 ± 0.18	0.88 ± 0.23	1.29 ± 0.22	1.52 ± 0.16	1.13 ± 0.19
Avg. Rank	**1.56**	5.48	3.88	4.96	4.20	7.08	5.60	5.72	6.52

Table 7. Comparison of NMI values obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	0.16 ± 0.02	0.16 ± 0.04	0.16 ± 0.03	0.14 ± 0.03	0.15 ± 0.04	**0.21 ± 0.04**	0.18 ± 0.02	0.15 ± 0.05	0.19 ± 0.06
CLL.SUB.111	0.18 ± 0.03	0.18 ± 0.02	0.17 ± 0.01	**0.19 ± 0.03**	0.17 ± 0.03	**0.19 ± 0.03**	0.18 ± 0.02	**0.19 ± 0.02**	0.17 ± 0.02
GLIOMA	**0.49 ± 0.04**	0.45 ± 0.04	0.39 ± 0.09	0.34 ± 0.09	0.43 ± 0.03	0.43 ± 0.09	0.44 ± 0.02	0.45 ± 0.07	0.23 ± 0.05
Prostate.GE	0.16 ± 0.02	0.13 ± 0.02	0.13 ± 0.03	**0.17 ± 0.04**	0.13 ± 0.02	0.05 ± 0.03	0.14 ± 0.01	0.15 ± 0.11	0.08 ± 0.02
SCADI	0.52 ± 0.02	**0.53 ± 0.18**	0.36 ± 0.06	0.11 ± 0.04	0.41 ± 0.13	0.46 ± 0.11	0.33 ± 0.10	0.52 ± 0.16	0.23 ± 0.08
SPECT	**0.10 ± 0.01**	0.07 ± 0.02	0.09 ± 0.02	0.08 ± 0.01	0.09 ± 0.01	0.06 ± 0.01	0.08 ± 0.01	0.08 ± 0.02	**0.10 ± 0.01**
Tic-Tac-Toe	**0.15 ± 0.01**	0.02 ± 0.00	0.03 ± 0.02	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.02	0.02 ± 0.02	0.03 ± 0.01	0.14 ± 0.02
Arcene	**0.14 ± 0.02**	0.08 ± 0.03	0.11 ± 0.03	0.10 ± 0.02	0.12 ± 0.02	0.05 ± 0.03	0.11 ± 0.01	0.07 ± 0.03	0.05 ± 0.03
Audiology	0.47 ± 0.04	0.05 ± 0.05	0.53 ± 0.06	0.12 ± 0.23	0.12 ± 0.02	0.06 ± 0.02	0.08 ± 0.03	0.07 ± 0.07	**0.62 ± 0.06**
Ba	0.29 ± 0.02	0.35 ± 0.05	0.29 ± 0.03	**0.37 ± 0.04**	**0.37 ± 0.04**	0.23 ± 0.02	0.25 ± 0.05	**0.37 ± 0.04**	0.24 ± 0.02
Coil20	**0.54 ± 0.02**	0.25 ± 0.07	0.45 ± 0.02	0.21 ± 0.01	0.21 ± 0.01	0.26 ± 0.05	0.37 ± 0.05	0.21 ± 0.01	0.41 ± 0.01
Colon	**0.21 ± 0.03**	0.14 ± 0.04	0.16 ± 0.03	0.15 ± 0.04	0.17 ± 0.04	0.15 ± 0.03	0.10 ± 0.01	0.12 ± 0.03	0.11 ± 0.03
Isolet	0.26 ± 0.03	0.39 ± 0.06	0.29 ± 0.04	**0.42 ± 0.05**	0.36 ± 0.03	0.38 ± 0.07	0.22 ± 0.03	0.31 ± 0.03	0.21 ± 0.02
Leukemia	**0.17 ± 0.03**	0.11 ± 0.03	0.14 ± 0.03	0.13 ± 0.04	0.14 ± 0.04	0.12 ± 0.05	0.08 ± 0.01	0.15 ± 0.04	0.15 ± 0.06
LSVT	0.11 ± 0.02	0.08 ± 0.02	0.11 ± 0.03	0.07 ± 0.01	**0.12 ± 0.03**	0.05 ± 0.01	**0.12 ± 0.02**	0.08 ± 0.04	0.09 ± 0.03
Lymphoma	**0.41 ± 0.02**	0.40 ± 0.03	0.37 ± 0.02	**0.41 ± 0.05**	0.39 ± 0.02	**0.41 ± 0.05**	0.27 ± 0.02	0.37 ± 0.04	0.40 ± 0.07
Mushrooms	0.25 ± 0.00	0.17 ± 0.13	0.18 ± 0.10	0.37 ± 0.00	**0.39 ± 0.06**	0.09 ± 0.06	0.05 ± 0.02	0.17 ± 0.11	0.03 ± 0.01
Nci9	**0.59 ± 0.03**	0.47 ± 0.03	0.36 ± 0.05	0.38 ± 0.07	0.36 ± 0.03	0.19 ± 0.17	0.31 ± 0.02	0.49 ± 0.05	0.27 ± 0.08
ORL	**0.54 ± 0.02**	0.53 ± 0.02	0.50 ± 0.02	0.53 ± 0.02	0.52 ± 0.02	0.50 ± 0.02	0.52 ± 0.01	0.45 ± 0.06	0.49 ± 0.02
Promoters	**0.27 ± 0.02**	0.11 ± 0.03	0.13 ± 0.03	0.13 ± 0.05	0.12 ± 0.04	0.14 ± 0.05	0.12 ± 0.03	0.14 ± 0.05	0.13 ± 0.03
Semeion	0.22 ± 0.01	0.24 ± 0.02	0.22 ± 0.03	0.26 ± 0.02	0.26 ± 0.04	0.24 ± 0.02	0.19 ± 0.04	**0.30 ± 0.04**	0.25 ± 0.04
Splice	**0.35 ± 0.00**	0.04 ± 0.04	0.05 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01	0.06 ± 0.03	0.02 ± 0.06	0.07 ± 0.05
Umist	0.48 ± 0.02	**0.49 ± 0.09**	0.40 ± 0.03	0.42 ± 0.06	0.47 ± 0.04	0.42 ± 0.05	0.42 ± 0.03	0.37 ± 0.09	0.38 ± 0.02
WarpAR10P	**0.50 ± 0.03**	0.44 ± 0.05	0.41 ± 0.02	0.44 ± 0.03	0.42 ± 0.05	0.41 ± 0.06	0.44 ± 0.03	0.25 ± 0.03	0.42 ± 0.02
WarpPIE10P	**0.49 ± 0.03**	0.39 ± 0.05	0.41 ± 0.02	0.37 ± 0.02	0.43 ± 0.04	0.29 ± 0.06	0.37 ± 0.04	0.48 ± 0.04	0.33 ± 0.05
Avg. Rank	**2.36**	5.16	5.08	4.92	4.68	6.12	5.80	5.12	5.76

Tables 8–10 present the Entropy Ratio, PDP, and Gini Impurity obtained from the feature subsets selected by the proposed and baseline methods, respectively. Entropy Ratio, PDP, and Gini Impurity differ from the preceding measures in one respect: their evaluation does not require class labels. They describe retained entropy, the number of observed feature patterns, and the distribution of those patterns. Each entry reports the mean and standard deviation over repeated runs, and the best result for each dataset is highlighted in bold. The last row reports the average rank across the 25 datasets. For the Entropy Ratio, the proposed method achieves the highest value on all 25 datasets and obtains the best possible average rank of 1.00. This result spans datasets with different feature counts and data types. The reported Entropy Ratio gaps are especially visible on SCADI, Audiology, Coil20, Colon, Mushrooms, Nci9, Semeion, and Splice. For example, it achieves Entropy Ratio values of 0.73, 0.62, and 0.83 on SCADI, Audiology, and Colon, respectively, compared with the corresponding second-best values of 0.63, 0.53, and 0.75. Across these datasets, the selected subsets have a larger Entropy Ratio, consistent with the entropy-based selection objective. For PDP, its average rank is 1.10; in the reported table, it has the highest or tied-highest value on every dataset. The improvements are particularly evident on GLIOMA, SCADI, Colon, Nci9, and Promoters, for which the proposed method obtains PDP values of 0.67, 0.40, 0.57, 0.59, and 0.35, respectively. The corresponding second-best values are 0.56, 0.30, 0.46, 0.42, and 0.25. On Ba, Isolet, Mushrooms, and Splice, PDP values are similarly small for several methods, including the proposed method, because $32/N$ is itself small for these large datasets. On Ba, Coil20, Isolet, ORL, Semeion, Splice, and Umist, the proposed method reaches the bound in Equation (48), observing all 32 patterns in every split, which is why its standard deviation is zero on these datasets; several baselines show the same behavior on some of them. While the Entropy Ratio measures retained entropy, we use PDP to count the distinct patterns observed among samples. We then examine Gini Impurity to see how evenly those patterns occur. On this measure, the proposed method records the highest or tied-highest value for every dataset, giving it an average rank of 1.00; most of its values fall between 0.93 and 0.97, close to the upper bound of approximately 0.969 in Equation (48). Because the values are near this ceiling, their variation across splits is small, and many standard deviations round to 0.00. For SCADI, Audiology, and Coil20, it scores 0.94, 0.95, and 0.96, compared with second-best values of 0.91, 0.90, and 0.94. We also observe larger declines for some baselines on Mushrooms and Splice. Taken together, these Gini results suggest that the patterns retained by our method are less concentrated in a few states. The three intrinsic measures thus complement the classification, clustering, MI, and NMI results: they describe properties of the selected subsets without using labels, alongside performance on the tasks evaluated here. They do not, by themselves, guarantee performance on an arbitrary future task.

Figure 1 summarizes the arithmetic mean feature-selection time of each method across the 25 benchmark datasets, while Table 11 reports the mean time on each dataset and the average rank. The figure uses a logarithmic horizontal axis to accommodate the large differences in running time. Times are wall-clock times of the feature-selection step alone, averaged over the 10 splits, measured in MATLAB on a single machine with the same settings for all methods; preprocessing and evaluation are excluded. On the smallest datasets, SPECT and Tic-Tac-Toe, the proposed method finishes in about 1 ms and 3 ms. The proposed method has the lowest overall mean time at 22.11 s, followed by CNAFS at 40.04 s, and also obtains the best average time rank of 1.84, compared with 2.72 for CNAFS. It runs faster than UDFS, BLSFS, BGLR, RAFG, and SDAE on all 25 datasets; the latter three methods have overall mean times of 3437.84 s, 13,506.40 s, and 90.76 s, respectively. The ordering is not identical on every dataset: CNAFS is faster than the proposed method on 11 datasets, whereas the proposed method is faster on the other 14. The difference is

particularly pronounced on Mushrooms and Splice, where the proposed method requires 0.07 s and 0.25 s, compared with 146.99 s and 13.51 s for CNAFS. These results indicate favorable running-time performance under the evaluated settings, without implying that the proposed method is the fastest on every individual dataset.

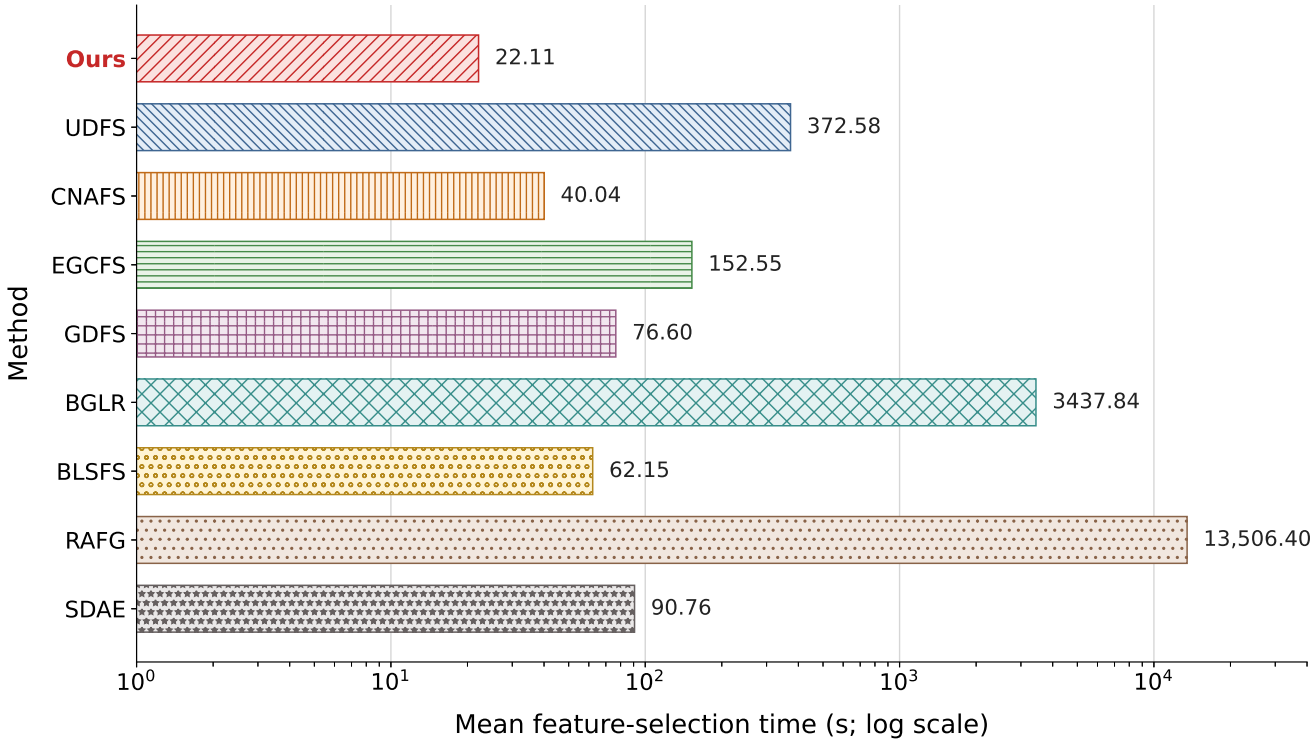


Figure 1. Mean feature-selection time of the proposed method and eight baseline methods across 25 benchmark datasets. The horizontal axis uses a logarithmic scale; lower values indicate faster execution. Each bar corresponds to the method labeled on the vertical axis.

Figure 2 plots K-means, Naive Bayes, and Decision Tree accuracy against the number of selected features, from 1 to 50, for Coil20 and Promoters. On Coil20, the proposed method achieves rapid improvements with relatively few features and outperforms the baselines over most of the plotted range, although BGLR becomes comparable or slightly better near 50 features. On Promoters, the proposed method maintains a clear advantage in both classification accuracies throughout the range. Its K-means accuracy peaks at approximately 13 selected features and then gradually declines, but remains higher than that of the baselines. On these two datasets, the method performs well with relatively few features; adding further features does not improve every plotted measure.

Figure 3 compares nine methods over 25 datasets using CD diagrams for the eight evaluation metrics, and Tables 12 and 13 report the corresponding Friedman and Wilcoxon–Holm tests. The proposed method achieved the lowest average rank on every metric, and the Friedman test rejects equal performance for every measure except K-means accuracy ($p = 0.735$). For MI, NMI, Entropy Ratio, PDP, and Gini Impurity, its rank differences from most baseline methods exceeded the CD value of 2.40, and all pairwise Wilcoxon–Holm differences are significant. For classification accuracy, the Nemenyi test separates the proposed method only from UDFS, BGLR, and BLSFS in Naive Bayes accuracy and from UDFS in Decision Tree accuracy. Because the Nemenyi test controls the error rate over all 36 pairs, it is conservative for comparisons against a single method. The Wilcoxon–Holm comparison shows that the proposed method is significantly better than seven of the eight baselines in Decision Tree accuracy (all except RAFG, $p = 0.107$) and than all eight baselines

in Naive Bayes accuracy. For K-means accuracy, no pairwise difference is significant after correction, so the proposed method is competitive with, but not demonstrably better than the baselines on this measure. Because PDP, Entropy Ratio, and Gini are closely aligned with the entropy objective, we regard the classification and NMI results as the more informative evidence that the retained information transfers to downstream tasks. Overall, the proposed method shows significant advantages in information preservation and classification accuracy while remaining competitive in clustering accuracy; these results do not establish superiority on every downstream task.

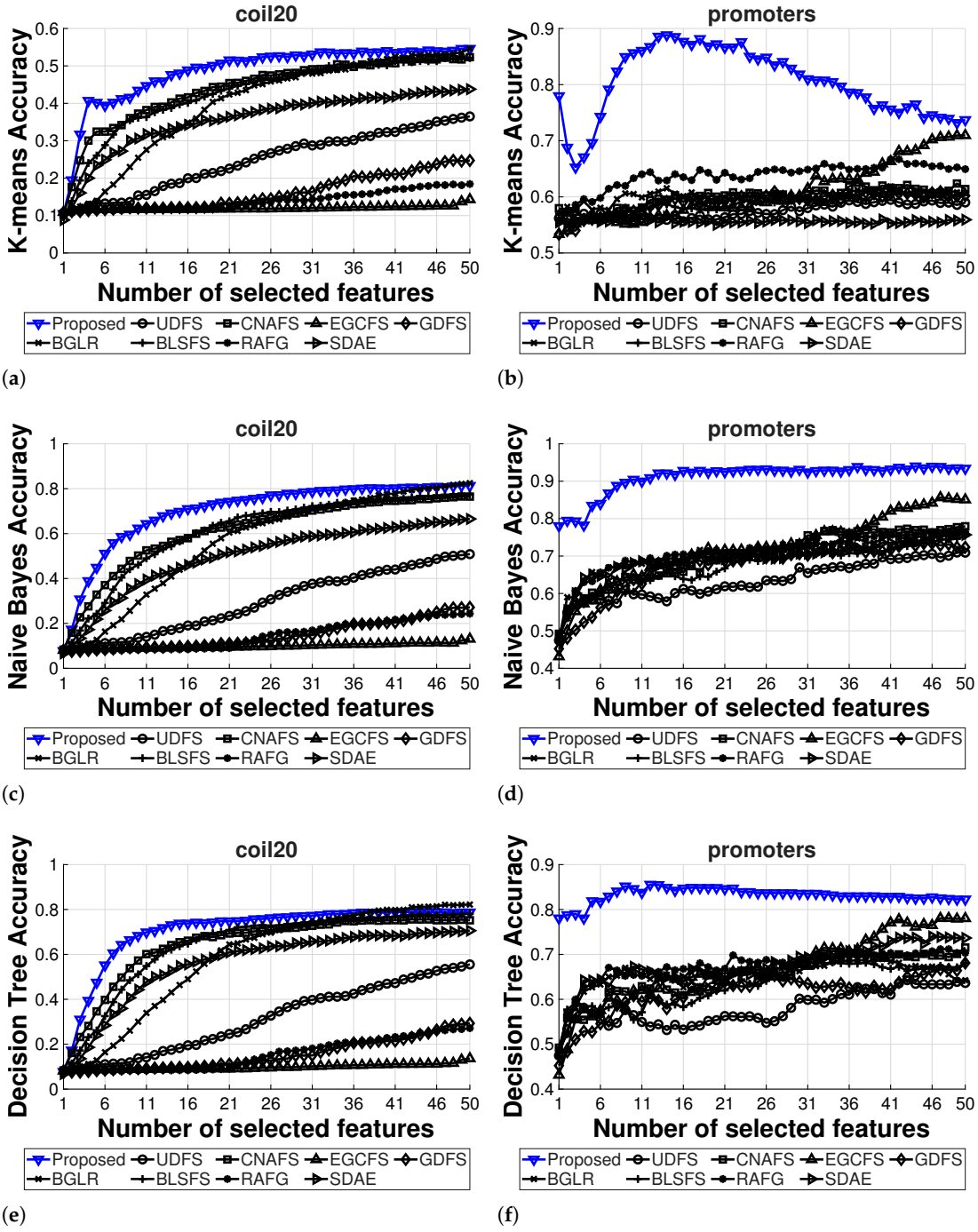


Figure 2. Classification and clustering accuracy according to the number of selected features on the Coil20 and Promoters datasets: (**a**,**c**,**e**) Coil20; (**b**,**d**,**f**) Promoters. Panels (**a**,**b**) show K-means accuracy; panels (**c**,**d**) show Naive Bayes accuracy; panels (**e**,**f**) show Decision Tree accuracy.

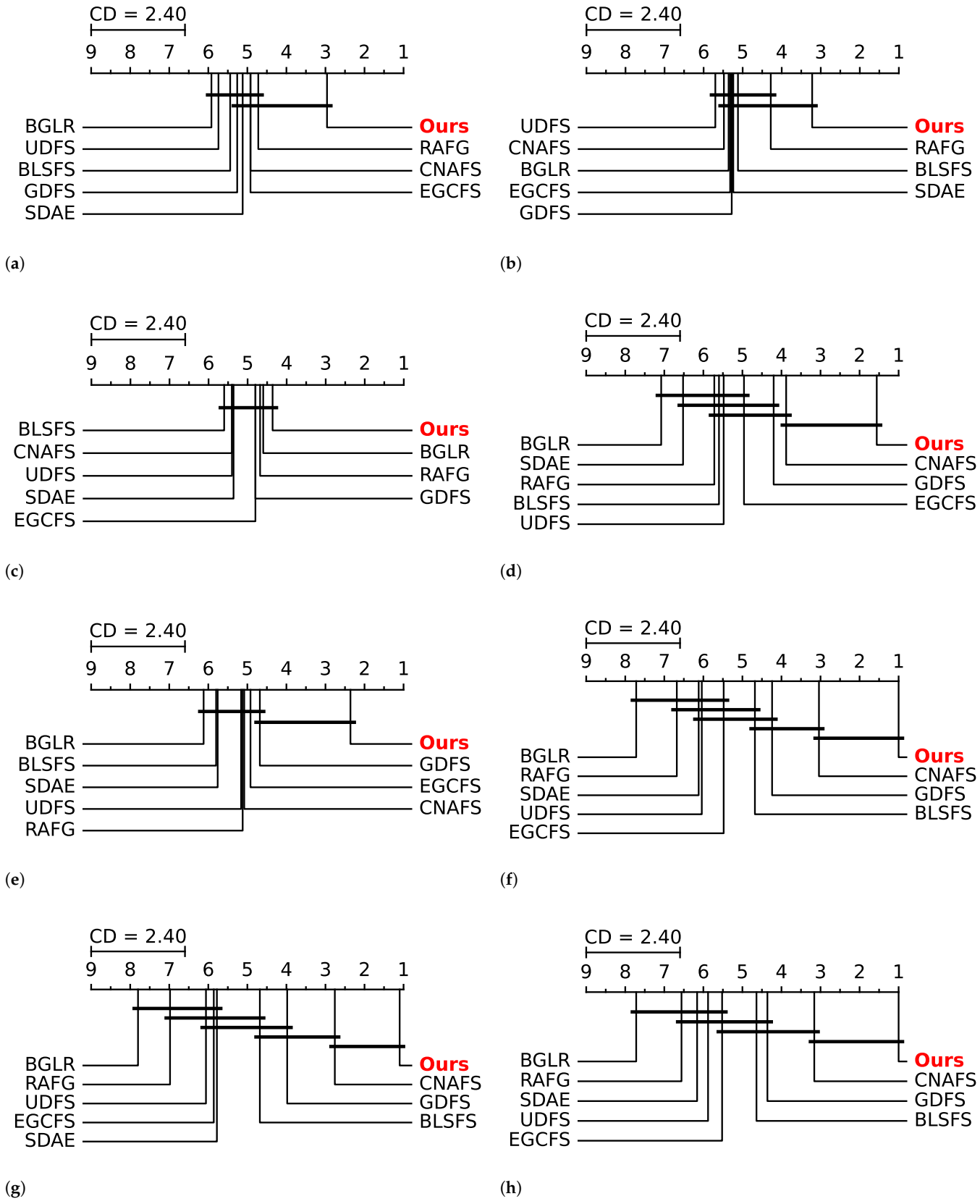


Figure 3. Critical-difference diagrams for the eight evaluation metrics: (**a**) Naive Bayes accuracy; (**b**) Decision Tree accuracy; (**c**) K-means accuracy; (**d**) MI; (**e**) NMI; (**f**) Entropy Ratio; (**g**) PDP; (**h**) Gini Impurity. Lower average ranks indicate better performance. Methods connected by a horizontal bar are not significantly different (Nemenyi test, CD = 2.40).

Table 8. Comparison of Entropy Ratio values obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	**0.81 ± 0.02**	0.72 ± 0.04	0.74 ± 0.02	0.75 ± 0.03	0.76 ± 0.01	0.61 ± 0.07	0.77 ± 0.01	0.63 ± 0.06	0.44 ± 0.09
CLL.SUB.111	**0.75 ± 0.01**	0.68 ± 0.02	0.71 ± 0.01	0.69 ± 0.02	0.71 ± 0.02	0.60 ± 0.06	0.68 ± 0.01	0.65 ± 0.02	0.54 ± 0.06
GLIOMA	**0.87 ± 0.02**	0.72 ± 0.05	0.75 ± 0.10	0.73 ± 0.07	0.80 ± 0.02	0.50 ± 0.10	0.79 ± 0.01	0.60 ± 0.05	0.32 ± 0.08
Prostate.GE	**0.76 ± 0.01**	0.62 ± 0.05	0.69 ± 0.03	0.67 ± 0.03	0.67 ± 0.02	0.16 ± 0.07	0.51 ± 0.01	0.47 ± 0.11	0.20 ± 0.08
SCADI	**0.73 ± 0.04**	0.36 ± 0.07	0.63 ± 0.04	0.10 ± 0.03	0.38 ± 0.13	0.29 ± 0.06	0.41 ± 0.09	0.33 ± 0.12	0.27 ± 0.10
SPECT	**0.59 ± 0.04**	0.49 ± 0.05	0.53 ± 0.05	0.54 ± 0.00	0.54 ± 0.03	0.47 ± 0.03	0.50 ± 0.04	0.51 ± 0.05	0.53 ± 0.03
Tic-Tac-Toe	**0.50 ± 0.00**	0.32 ± 0.00	0.42 ± 0.02	0.32 ± 0.00	0.38 ± 0.00	0.32 ± 0.00	0.33 ± 0.01	0.40 ± 0.04	0.49 ± 0.00
Arcene	**0.76 ± 0.01**	0.48 ± 0.17	0.70 ± 0.02	0.63 ± 0.06	0.70 ± 0.02	0.13 ± 0.10	0.64 ± 0.03	0.41 ± 0.05	0.13 ± 0.05
Audiology	**0.62 ± 0.01**	0.02 ± 0.02	0.53 ± 0.03	0.07 ± 0.13	0.04 ± 0.01	0.03 ± 0.01	0.06 ± 0.03	0.03 ± 0.02	0.49 ± 0.02
Ba	**0.49 ± 0.00**	0.38 ± 0.05	0.45 ± 0.02	0.37 ± 0.04	0.41 ± 0.04	0.36 ± 0.03	0.39 ± 0.05	0.36 ± 0.04	0.43 ± 0.03
Coil20	**0.46 ± 0.00**	0.08 ± 0.03	0.43 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.11 ± 0.04	0.32 ± 0.06	0.06 ± 0.00	0.42 ± 0.01
Colon	**0.83 ± 0.02**	0.59 ± 0.07	0.74 ± 0.03	0.70 ± 0.06	0.75 ± 0.03	0.45 ± 0.09	0.46 ± 0.02	0.57 ± 0.07	0.34 ± 0.09
Isolet	**0.48 ± 0.00**	0.17 ± 0.03	0.46 ± 0.01	0.22 ± 0.01	0.18 ± 0.02	0.28 ± 0.07	0.43 ± 0.05	0.21 ± 0.02	0.45 ± 0.02
Leukemia	**0.80 ± 0.02**	0.64 ± 0.03	0.73 ± 0.03	0.64 ± 0.07	0.74 ± 0.04	0.26 ± 0.10	0.51 ± 0.01	0.60 ± 0.05	0.38 ± 0.11
LSVT	**0.73 ± 0.01**	0.43 ± 0.16	0.69 ± 0.01	0.10 ± 0.03	0.51 ± 0.10	0.27 ± 0.11	0.68 ± 0.03	0.22 ± 0.07	0.64 ± 0.09
Lymphoma	**0.77 ± 0.01**	0.66 ± 0.03	0.72 ± 0.02	0.66 ± 0.04	0.74 ± 0.01	0.57 ± 0.05	0.49 ± 0.00	0.65 ± 0.03	0.58 ± 0.05
Mushrooms	**0.40 ± 0.00**	0.19 ± 0.06	0.16 ± 0.05	0.25 ± 0.00	0.20 ± 0.01	0.06 ± 0.04	0.28 ± 0.03	0.12 ± 0.05	0.02 ± 0.01
Nci9	**0.84 ± 0.02**	0.66 ± 0.04	0.47 ± 0.05	0.52 ± 0.10	0.49 ± 0.04	0.23 ± 0.24	0.52 ± 0.01	0.71 ± 0.06	0.29 ± 0.12
ORL	**0.59 ± 0.00**	0.48 ± 0.05	0.57 ± 0.01	0.54 ± 0.02	0.56 ± 0.02	0.49 ± 0.02	0.56 ± 0.02	0.36 ± 0.10	0.55 ± 0.01
Promoters	**0.72 ± 0.01**	0.56 ± 0.03	0.61 ± 0.01	0.48 ± 0.03	0.57 ± 0.02	0.48 ± 0.04	0.56 ± 0.03	0.57 ± 0.02	0.46 ± 0.03
Semeion	**0.48 ± 0.00**	0.19 ± 0.03	0.42 ± 0.02	0.21 ± 0.02	0.30 ± 0.06	0.30 ± 0.06	0.38 ± 0.04	0.37 ± 0.03	0.38 ± 0.03
Splice	**0.41 ± 0.00**	0.25 ± 0.05	0.34 ± 0.01	0.02 ± 0.06	0.00 ± 0.00	0.04 ± 0.07	0.34 ± 0.01	0.06 ± 0.09	0.16 ± 0.07
Umist	**0.55 ± 0.00**	0.43 ± 0.07	0.52 ± 0.01	0.41 ± 0.06	0.45 ± 0.04	0.40 ± 0.05	0.53 ± 0.01	0.37 ± 0.09	0.52 ± 0.02
WarpAR10P	**0.72 ± 0.00**	0.58 ± 0.04	0.62 ± 0.02	0.62 ± 0.03	0.65 ± 0.02	0.48 ± 0.07	0.61 ± 0.06	0.29 ± 0.03	0.52 ± 0.06
WarpPIE10P	**0.62 ± 0.01**	0.50 ± 0.04	0.57 ± 0.02	0.53 ± 0.02	0.57 ± 0.02	0.36 ± 0.06	0.48 ± 0.08	0.41 ± 0.07	0.49 ± 0.02
Avg. Rank	**1.00**	6.04	3.04	5.48	4.24	7.72	4.68	6.68	6.12

Table 9. Comparison of PDP values obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	**0.50 ± 0.03**	0.39 ± 0.03	0.42 ± 0.04	0.43 ± 0.04	0.44 ± 0.02	0.30 ± 0.05	0.47 ± 0.02	0.32 ± 0.04	0.21 ± 0.05
CLL.SUB.111	**0.35 ± 0.01**	0.30 ± 0.03	0.33 ± 0.01	0.31 ± 0.01	0.32 ± 0.02	0.25 ± 0.04	0.29 ± 0.01	0.28 ± 0.02	0.23 ± 0.04
GLIOMA	**0.67 ± 0.05**	0.46 ± 0.07	0.50 ± 0.12	0.45 ± 0.10	0.56 ± 0.04	0.26 ± 0.07	0.53 ± 0.02	0.33 ± 0.06	0.21 ± 0.04
Prostate.GE	**0.37 ± 0.01**	0.27 ± 0.04	0.32 ± 0.03	0.30 ± 0.03	0.31 ± 0.02	0.05 ± 0.02	0.21 ± 0.01	0.18 ± 0.07	0.10 ± 0.03
SCADI	**0.40 ± 0.05**	0.13 ± 0.03	0.30 ± 0.04	0.09 ± 0.02	0.16 ± 0.04	0.11 ± 0.02	0.17 ± 0.03	0.11 ± 0.05	0.12 ± 0.03
SPECT	**0.13 ± 0.01**	0.12 ± 0.01	**0.13 ± 0.01**	**0.13 ± 0.00**	**0.13 ± 0.01**	0.11 ± 0.01	0.12 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
Tic-Tac-Toe	**0.04 ± 0.00**	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	**0.04 ± 0.00**
Arcene	**0.39 ± 0.01**	0.17 ± 0.08	0.33 ± 0.02	0.28 ± 0.04	0.33 ± 0.02	0.05 ± 0.02	0.29 ± 0.03	0.11 ± 0.03	0.07 ± 0.02
Audiology	**0.16 ± 0.01**	0.02 ± 0.01	0.13 ± 0.01	0.02 ± 0.02	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.01 ± 0.01	0.10 ± 0.01
Ba	**0.03 ± 0.00**	**0.03 ± 0.00**	**0.03 ± 0.00**	**0.03 ± 0.00**	**0.03 ± 0.00**	0.02 ± 0.00	**0.03 ± 0.00**	**0.03 ± 0.00**	**0.03 ± 0.00**
Coil20	**0.03 ± 0.00**	0.01 ± 0.00	**0.03 ± 0.00**	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	**0.03 ± 0.00**
Colon	**0.57 ± 0.03**	0.29 ± 0.07	0.44 ± 0.03	0.40 ± 0.07	0.46 ± 0.04	0.25 ± 0.06	0.23 ± 0.02	0.26 ± 0.06	0.20 ± 0.05
Isolet	**0.03 ± 0.00**	0.01 ± 0.00	**0.03 ± 0.00**	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	**0.03 ± 0.00**
Leukemia	**0.50 ± 0.03**	0.30 ± 0.04	0.40 ± 0.05	0.33 ± 0.07	0.42 ± 0.04	0.13 ± 0.04	0.16 ± 0.00	0.30 ± 0.03	0.18 ± 0.06
LSVT	**0.31 ± 0.01**	0.13 ± 0.08	0.29 ± 0.02	0.07 ± 0.01	0.18 ± 0.04	0.07 ± 0.03	0.28 ± 0.03	0.07 ± 0.02	0.25 ± 0.07
Lymphoma	**0.41 ± 0.01**	0.30 ± 0.02	0.36 ± 0.02	0.30 ± 0.04	0.37 ± 0.02	0.24 ± 0.02	0.12 ± 0.00	0.30 ± 0.02	0.26 ± 0.05
Mushrooms	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
Nci9	**0.59 ± 0.04**	0.37 ± 0.05	0.23 ± 0.04	0.26 ± 0.06	0.26 ± 0.02	0.14 ± 0.12	0.18 ± 0.01	0.42 ± 0.08	0.16 ± 0.05
ORL	**0.10 ± 0.00**	0.09 ± 0.01	**0.10 ± 0.00**	**0.10 ± 0.00**	**0.10 ± 0.00**	0.09 ± 0.00	**0.10 ± 0.00**	0.08 ± 0.01	**0.10 ± 0.00**
Promoters	**0.35 ± 0.01**	0.21 ± 0.01	0.25 ± 0.01	0.20 ± 0.02	0.24 ± 0.02	0.18 ± 0.04	0.22 ± 0.03	0.21 ± 0.02	0.19 ± 0.02
Semeion	**0.03 ± 0.00**	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Splice	**0.01 ± 0.00**	0.00 ± 0.00	**0.01 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.01 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00
Umist	**0.07 ± 0.00**	0.06 ± 0.01	**0.07 ± 0.00**	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	**0.07 ± 0.00**	0.05 ± 0.01	**0.07 ± 0.00**
WarpAR10P	**0.30 ± 0.01**	0.22 ± 0.04	0.24 ± 0.02	0.24 ± 0.03	0.26 ± 0.02	0.15 ± 0.03	0.24 ± 0.03	0.08 ± 0.01	0.22 ± 0.03
WarpPIE10P	**0.18 ± 0.01**	0.13 ± 0.02	0.17 ± 0.01	0.15 ± 0.02	0.16 ± 0.01	0.08 ± 0.02	0.14 ± 0.02	0.10 ± 0.02	0.14 ± 0.02
Avg. Rank	**1.10**	6.06	2.76	5.86	3.98	7.80	4.68	6.98	5.78

Table 10. Comparison of Gini Impurity values obtained by the proposed and baseline methods. Results are reported as means ± standard deviations. The best result for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	**0.96 ± 0.00**	0.94 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.00	0.88 ± 0.05	0.95 ± 0.00	0.90 ± 0.03	0.70 ± 0.12
CLL.SUB.111	**0.96 ± 0.00**	0.94 ± 0.01	0.95 ± 0.00	0.95 ± 0.01	0.95 ± 0.00	0.90 ± 0.04	0.94 ± 0.00	0.93 ± 0.01	0.86 ± 0.04
GLIOMA	**0.96 ± 0.00**	0.91 ± 0.03	0.92 ± 0.05	0.92 ± 0.03	0.94 ± 0.01	0.77 ± 0.10	0.94 ± 0.00	0.85 ± 0.03	0.50 ± 0.13
Prostate.GE	**0.96 ± 0.00**	0.91 ± 0.03	0.94 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.42 ± 0.19	0.84 ± 0.01	0.80 ± 0.09	0.37 ± 0.15
SCADI	**0.94 ± 0.01**	0.69 ± 0.10	0.91 ± 0.02	0.16 ± 0.05	0.66 ± 0.21	0.59 ± 0.13	0.71 ± 0.13	0.64 ± 0.19	0.48 ± 0.16
SPECT	**0.93 ± 0.02**	0.85 ± 0.04	0.87 ± 0.05	0.89 ± 0.00	0.89 ± 0.03	0.84 ± 0.03	0.85 ± 0.04	0.87 ± 0.04	0.90 ± 0.02
Tic-Tac-Toe	**0.96 ± 0.00**	0.87 ± 0.00	0.93 ± 0.01	0.88 ± 0.00	0.87 ± 0.00	0.87 ± 0.00	0.88 ± 0.00	0.92 ± 0.02	**0.96 ± 0.00**
Arcene	**0.96 ± 0.00**	0.80 ± 0.17	0.94 ± 0.01	0.91 ± 0.04	0.94 ± 0.01	0.29 ± 0.23	0.92 ± 0.01	0.80 ± 0.05	0.25 ± 0.10
Audiology	**0.95 ± 0.00**	0.04 ± 0.04	0.90 ± 0.03	0.16 ± 0.31	0.07 ± 0.01	0.05 ± 0.01	0.10 ± 0.07	0.06 ± 0.06	0.90 ± 0.01
Ba	**0.97 ± 0.00**	0.89 ± 0.05	0.95 ± 0.01	0.89 ± 0.04	0.92 ± 0.03	0.87 ± 0.04	0.90 ± 0.04	0.88 ± 0.05	0.93 ± 0.02
Coil20	**0.96 ± 0.00**	0.28 ± 0.11	0.94 ± 0.01	0.21 ± 0.02	0.21 ± 0.01	0.33 ± 0.11	0.83 ± 0.07	0.20 ± 0.01	0.93 ± 0.01
Colon	**0.96 ± 0.00**	0.87 ± 0.04	0.93 ± 0.01	0.92 ± 0.02	0.94 ± 0.01	0.69 ± 0.12	0.77 ± 0.01	0.86 ± 0.04	0.56 ± 0.14
Isolet	**0.97 ± 0.00**	0.63 ± 0.11	0.96 ± 0.01	0.72 ± 0.06	0.67 ± 0.04	0.75 ± 0.14	0.93 ± 0.04	0.72 ± 0.02	0.95 ± 0.01
Leukemia	**0.96 ± 0.00**	0.91 ± 0.02	0.94 ± 0.01	0.90 ± 0.03	0.94 ± 0.01	0.49 ± 0.19	0.86 ± 0.01	0.88 ± 0.04	0.64 ± 0.15
LSVT	**0.96 ± 0.00**	0.80 ± 0.14	0.95 ± 0.00	0.18 ± 0.07	0.85 ± 0.08	0.59 ± 0.24	0.95 ± 0.01	0.52 ± 0.14	0.93 ± 0.03
Lymphoma	**0.96 ± 0.00**	0.93 ± 0.01	0.95 ± 0.01	0.93 ± 0.01	0.95 ± 0.00	0.87 ± 0.05	0.87 ± 0.00	0.92 ± 0.02	0.88 ± 0.03
Mushrooms	**0.95 ± 0.00**	0.69 ± 0.14	0.57 ± 0.18	0.84 ± 0.00	0.73 ± 0.04	0.24 ± 0.18	0.87 ± 0.04	0.51 ± 0.17	0.05 ± 0.02
Nci9	**0.96 ± 0.00**	0.90 ± 0.02	0.74 ± 0.04	0.80 ± 0.10	0.75 ± 0.04	0.35 ± 0.33	0.86 ± 0.00	0.92 ± 0.02	0.47 ± 0.20
ORL	**0.97 ± 0.00**	0.90 ± 0.03	0.95 ± 0.00	0.94 ± 0.01	0.95 ± 0.01	0.91 ± 0.02	0.95 ± 0.01	0.77 ± 0.12	0.94 ± 0.01
Promoters	**0.95 ± 0.00**	0.89 ± 0.03	0.90 ± 0.01	0.79 ± 0.03	0.88 ± 0.02	0.83 ± 0.04	0.88 ± 0.02	0.90 ± 0.02	0.77 ± 0.03
Semeion	**0.97 ± 0.00**	0.55 ± 0.10	0.93 ± 0.02	0.61 ± 0.05	0.78 ± 0.12	0.80 ± 0.10	0.90 ± 0.04	0.89 ± 0.03	0.90 ± 0.03
Splice	**0.95 ± 0.00**	0.84 ± 0.06	0.89 ± 0.02	0.08 ± 0.24	0.00 ± 0.00	0.15 ± 0.31	0.88 ± 0.02	0.24 ± 0.34	0.56 ± 0.22
Umist	**0.96 ± 0.00**	0.89 ± 0.06	0.95 ± 0.01	0.87 ± 0.05	0.90 ± 0.04	0.87 ± 0.06	0.95 ± 0.01	0.81 ± 0.11	0.95 ± 0.01
WarpAR10P	**0.96 ± 0.00**	0.91 ± 0.02	0.92 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.85 ± 0.05	0.91 ± 0.05	0.64 ± 0.05	0.84 ± 0.07
WarpPIE10P	**0.95 ± 0.00**	0.89 ± 0.02	0.92 ± 0.02	0.91 ± 0.02	0.93 ± 0.01	0.78 ± 0.06	0.85 ± 0.09	0.81 ± 0.07	0.87 ± 0.01
Avg. Rank	**1.00**	5.88	3.16	5.52	4.36	7.72	4.64	6.56	6.16

Table 11. Comparison of feature-selection execution time by the proposed and baseline methods. Results are reported as mean times in seconds over 10 runs; values below 1 s are given to three decimal places. Lower values indicate faster execution. The fastest method for each dataset is shown in bold.

Dataset	Proposed	UDFS	CNAFS	EGCFS	GDFS	BGLR	BLSFS	RAFG	SDAE
ALLAML	15.40	214.17	**13.02**	88.06	22.20	1073.21	25.38	7618.64	88.68
CLL.SUB.111	**39.64**	442.29	45.39	205.85	59.96	3695.54	74.71	20,717.01	215.33
GLIOMA	6.00	46.72	**3.37**	18.04	7.81	259.16	10.31	1983.06	40.22
Prostate.GE	11.59	120.31	**7.70**	46.38	14.95	537.15	21.14	3600.83	105.23
SCADI	0.019	0.025	**0.007**	0.011	0.088	3.39	0.083	0.562	0.625
SPECT	**0.001**	0.006	0.286	0.017	0.431	0.651	0.205	0.684	0.576
Tic-Tac-Toe	**0.003**	0.052	0.413	0.176	4.84	0.731	1.56	20.37	3.39
Arcene	30.46	298.27	**28.85**	150.76	43.61	2496.05	64.11	11,583.98	173.60
Audiology	**0.010**	0.020	0.031	0.020	0.373	4.14	0.257	0.850	1.45
Ba	**0.145**	0.415	1.82	0.814	10.22	2.77	4.69	55.35	23.18
Coil20	1.20	1.52	3.99	**1.18**	11.41	13.20	4.34	61.23	251.21
Colon	10.31	122.08	**6.50**	33.46	14.92	623.46	21.63	4161.17	63.71
Isolet	**0.479**	0.835	3.40	1.09	12.66	5.89	5.42	43.52	166.03
Leukemia	**147.27**	2433.63	205.91	1459.98	237.88	23,618.30	395.48	84,373.18	254.04
LSVT	0.056	0.094	0.025	**0.024**	0.159	2.79	0.149	3.04	2.09
Lymphoma	**42.24**	529.11	44.85	273.40	65.03	4583.18	100.45	19,269.87	188.45
Mushrooms	**0.075**	4.81	146.99	47.79	892.78	11.39	35.52	6601.15	64.30
Nci9	**241.37**	5082.98	469.56	1470.02	422.30	48,872.27	752.02	176,806.90	282.07
ORL	0.714	1.09	0.528	**0.435**	1.10	12.04	1.21	34.42	70.21
Promoters	0.056	0.057	**0.013**	0.031	0.118	3.10	0.120	1.96	1.24
Semeion	**0.106**	0.376	2.17	1.40	12.92	3.05	4.13	39.02	20.93
Splice	**0.248**	1.29	13.51	10.07	72.98	4.98	22.83	241.68	44.75
Umist	0.334	0.439	0.521	**0.332**	1.68	5.90	1.15	16.38	63.57
WarpAR10P	2.56	6.96	**0.939**	2.17	2.21	57.59	3.30	240.87	55.11
WarpPIE10P	2.56	6.99	**1.14**	2.35	2.43	55.98	3.62	184.29	88.96
Avg. Rank	**1.84**	4.52	2.72	3.48	4.96	7.12	4.88	8.64	6.84

Table 12. Friedman test on the average ranks over 25 datasets ($k = 9$ methods). χ_F^2: Friedman statistic; F_F: Iman–Davenport correction; the Nemenyi critical difference at $\alpha = 0.05$ is CD = 2.40.

Measure	χ_F^2	p	F_F	p	Rank (Proposed)
Decision Tree	15.97	0.043	2.08	0.039	3.22
Naive Bayes	19.75	0.011	2.63	0.009	2.96
K-means	5.21	0.735	0.64	0.742	4.36
MI	71.58	2.4×10^{-12}	13.38	2.6×10^{-15}	1.56
NMI	31.99	9.4×10^{-5}	4.57	4.1×10^{-5}	2.36
PDP	119.23	4.8×10^{-22}	35.43	$< 10^{-15}$	1.10
Entropy Ratio	111.03	2.3×10^{-20}	29.95	$< 10^{-15}$	1.00
Gini	107.16	1.5×10^{-19}	27.70	$< 10^{-15}$	1.00

Table 13. Pairwise Wilcoxon signed-rank tests between the proposed method and each baseline over 25 datasets, with Holm correction for eight comparisons. Each cell gives win/tie/loss of the proposed method (top) and the Holm-adjusted p-value (bottom); bold indicates $p < 0.05$.

	Proposed vs.							
Measure	**UDFS**	**CNAFS**	**EGCFS**	**GDFS**	**BGLR**	**BLSFS**	**RAFG**	**SDAE**
Decision Tree	**19/1/5** **0.003**	**19/0/6** **0.014**	**17/1/7** **0.022**	**18/0/7** **0.011**	**19/0/6** **0.011**	**19/0/6** **0.004**	15/0/10 0.107	**18/0/7** **0.041**
Naive Bayes	**19/0/6** **0.002**	**18/0/7** **0.018**	**17/0/8** **0.037**	**19/0/6** **0.017**	**20/0/5** **0.005**	**22/0/3** **<0.001**	**15/1/9** **0.037**	**20/0/5** **0.037**
K-means	15/0/10 1.000	16/0/9 0.108	13/0/12 1.000	14/0/11 1.000	14/0/11 1.000	17/0/8 0.150	11/0/14 1.000	16/0/9 1.000
MI	**24/0/1** **<0.001**	**23/0/2** **<0.001**	**21/0/4** **<0.001**	**24/0/1** **<0.001**	**23/0/2** **<0.001**	**24/0/1** **<0.001**	**23/0/2** **<0.001**	**24/0/1** **<0.001**
NMI	**20/0/5** **0.004**	**22/0/3** **<0.001**	**18/0/7** **0.013**	**20/0/5** **0.013**	**21/0/4** **<0.001**	**23/0/2** **<0.001**	**21/0/4** **0.003**	**21/0/4** **0.001**
PDP	**25/0/0** **<0.001**	**22/3/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**24/1/0** **<0.001**
Entropy Ratio	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**
Gini	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**	**25/0/0** **<0.001**

To examine how α and the choice of surrogate affect the results, we repeated the selection on all 25 datasets with eight values of α and with six alternative surrogates. Every setting used the same 10 splits, so the settings can be compared split by split. The setting $\alpha = 1$ evaluates every pair exactly and serves as the reference.

Table 14 shows that the results saturate quickly as α increases. For $\alpha \leq 0.05$, the Entropy Ratio is significantly lower than with exact evaluation, and at $\alpha = 0.01$, Naive Bayes accuracy and MI also decrease. From $\alpha = 0.1$ onward, none of the six measures differs significantly from exact evaluation. At $\alpha = 0.1$, the averages are within 0.5% of the exact values, while the running time is 19.0 s instead of 91.0 s; the measured ratio of 0.23 is close to the theoretical fraction $2\alpha - \alpha^2 = 0.19$ of exactly evaluated pairs. The selected subsets are not identical to the exact ones (Jaccard overlap 0.78), but the replaced features have similar entropy and do not reduce downstream performance. We therefore recommend $\alpha = 0.1$ as a default. Values between 0.1 and 0.2 trade a modest increase in time for closer agreement with exact evaluation.

Table 14. Sensitivity to the landmark ratio α (geometric-mean surrogate, $k = 5$). Values are averages over 25 datasets of the per-dataset means over 10 splits. Jaccard is the mean overlap between the five selected features and those selected with exact pairwise evaluation ($\alpha = 1$) on the same split. †: significantly lower than $\alpha = 1$ (Wilcoxon signed-rank test over datasets, $p < 0.05$).

α	NB	DT	K-Means	MI	NMI	ER	Jaccard	Time (s)
0.01	0.493 †	0.493	0.439	1.185 †	0.318	0.637 †	0.36	2.91
0.02	0.503	0.497	0.445	1.205	0.323	0.639 †	0.46	4.86
0.05	0.506	0.500	0.438	1.217	0.323	0.647 †	0.63	10.55
0.1	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.04
0.2	0.506	0.501	0.437	1.214	0.320	0.652	0.89	34.03
0.3	0.506	0.502	0.438	1.215	0.321	0.653	0.95	47.46
0.5	0.507	0.503	0.438	1.216	0.321	0.653	0.98	68.98
1.0 (exact)	0.507	0.504	0.438	1.217	0.321	0.653	1.00	91.02

Table 15 compares surrogates for non-landmark pairs at $\alpha = 0.1$. The geometric, arithmetic, and harmonic means and the minimum and maximum of the two marginal entropies produce nearly identical results; their selected subsets coincide in 98–100% of the runs. The reason is that the selected features are mostly landmarks: on average 4.5 of the five selected features are landmarks, and in 64% of the runs all five are. With $k = 5$, non-landmark pairs therefore rarely decide the selection, and any surrogate that stays within the range of the two marginal entropies performs equally well. Within this family, we use the geometric mean for its analytical properties rather than for an empirical advantage. It is the unique diagonal-preserving rank-one factorization (Lemma 1), it never exceeds the true joint entropy (Lemma 2), and it is separable, so the non-landmark block need not be stored. Surrogates outside this range perform worse. The independence assumption $H(f_i) + H(f_j)$, an upper bound on the joint entropy, significantly lowers Entropy Ratio. The standard Nyström reconstruction $\mathbf{Q}_{RR} \approx \mathbf{Q}_{RL}\mathbf{Q}_{LL}^{+}\mathbf{Q}_{LR}$ significantly lowers five of the six measures, including Naive Bayes and Decision Tree accuracy, and reduces MI by up to 69% on individual datasets. This supports the observation in Section 2.3 that the joint-entropy matrix does not have the positive semidefinite structure on which the Nyström reconstruction relies.

Table 15. Comparison of surrogates for non-landmark pairs at $\alpha = 0.1$ ($k = 5$). Values are averages over 25 datasets of the per-dataset means over 10 splits. NB, DT, and ER denote Naive Bayes accuracy, Decision Tree accuracy, and Entropy Ratio, respectively. Jaccard is the mean overlap between the five selected features and those selected with exact pairwise evaluation ($\alpha = 1$) on the same split. †: significantly lower than exact evaluation ($\alpha = 1$) (Wilcoxon signed-rank test over datasets, $p < 0.05$).

Surrogate	NB	DT	K-Means	MI	NMI	ER	Jaccard	Time (s)
Geometric mean (proposed)	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.04
Arithmetic mean	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.83
Harmonic mean	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.24
$\min\{H(f_i), H(f_j)\}$	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.02
$\max\{H(f_i), H(f_j)\}$	0.507	0.503	0.438	1.215	0.321	0.651	0.78	19.47
$H(f_i) + H(f_j)$ (independence)	0.492	0.493	0.433	1.148	0.311	0.633 †	0.39	19.71
Nyström reconstruction	0.476 †	0.472 †	0.426	1.099 †	0.299 †	0.627 †	0.42	21.39
Exact ($\alpha = 1$)	0.507	0.504	0.438	1.217	0.321	0.653	1.00	91.02

5. Limitations

The subset entropy $H(S)$ measures information capacity, not relevance. By Equation (5), a large $H(S)$ is necessary for low error on any class variable but not sufficient: a feature can have high entropy and carry no class information, and after mean binarization even pure noise approaches the maximum of one bit. Datasets in which many features vary independently of the class are therefore the least favorable case. This is consistent with our results, in which the advantage is clear for the information-based measures, smaller for classification accuracy, and not significant for K-means accuracy. Moreover, Entropy Ratio, PDP, and Gini Impurity are aligned with the entropy objective, so the classification, clustering, and NMI results are the more informative evidence of downstream usefulness.

The method also inherits limitations from its estimation and search. The pairwise score ignores dependencies among three or more features, such as parity-type interactions, which supervised criteria such as R2CI [8] can model when labels are available. Entropy estimates depend on the discretization and are biased downward for small samples [34]. Greedy selection does not guarantee an optimal subset, and the landmark-based approximation, although it reduces the number of exactly evaluated pairs, still has a worst-case cost of $O(d^2N)$. Bias-corrected estimators, adaptive discretization, and the use of a small number of labels when they are available are natural directions for addressing these limitations.

6. Conclusions

We developed this method for a setting in which feature selection must take place before the downstream classification task is known. Motivated by the necessary condition implied by Fano's inequality, we used the entropy of the selected subset as a task-agnostic heuristic for information capacity, without claiming that it guarantees a lower error on a particular task. Estimating that subset's high-dimensional joint entropy directly is difficult, so our incremental search evaluates candidates through pairwise joint entropy. We further reduce pairwise computation by choosing high-marginal-entropy features as landmarks. Pairs involving a landmark receive empirical joint-entropy estimates; for the remaining pairs, we use a geometric-mean surrogate.

The evaluation across 25 public benchmark datasets placed our method first in average rank on all eight metrics. For MI, NMI, Entropy Ratio, PDP, and Gini Impurity, it significantly outperformed all baselines in pairwise Wilcoxon–Holm tests. In classification accuracy, it was significantly better than seven of the eight baselines with Decision Tree and all eight with Naive Bayes. Its clustering accuracy was competitive but not significantly different from the baselines. As discussed in Section 5, entropy measures information capacity rather than relevance, so preserving information without class labels offers a useful selection criterion, not a guarantee of better results on every later task. We plan to explore adaptive landmark selection, more accurate surrogates for pairs without landmarks, bias-corrected entropy estimation, and the use of a small number of labels when they are available.

Author Contributions: Conceptualization, W.S. and J.L.; methodology, G.-I.K. and T.K.; software, G.-I.K. and T.K.; validation, G.-I.K., T.K., W.S. and J.L.; formal analysis, W.S. and J.L.; investigation, G.-I.K. and T.K.; resources, J.L.; data curation, G.-I.K. and T.K.; writing—original draft preparation, G.-I.K.; writing—review and editing, W.S. and J.L.; visualization, G.-I.K.; supervision, W.S. and J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2026-25546026; Leading Generative AI Human Resources Development) and by the Institute of Information &

Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [RS-2021-II211341, Artificial Intelligent Graduate School Program (Chung-Ang University)].

Data Availability Statement: The public benchmark datasets used in this study can be accessed through the Feature Selection Datasets repository (https://jundongl.github.io/scikit-feature/datasets.html, accessed on 1 November 2026) , the UCI Machine Learning Repository (https://archive.ics.uci.edu/datasets, accessed on 1 November 2026), and Kaggle (https://www.kaggle.com/datasets, accessed on 1 November 2026). The implementation of the proposed method is available at https://github.com/Kim-Gyuil/landmark-entropy-ufs (accessed on 1 November 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brickley, D.; Burgess, M.; Noy, N. Google dataset search: Building a search engine for datasets in an open web ecosystem. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 1365–1375.
2. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Recent advances and emerging challenges of feature selection in the context of big data. *Knowl.-Based Syst.* **2015**, *86*, 33–45. [CrossRef]
3. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 6, 1–45. [CrossRef]
4. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182. [CrossRef]
5. Puerto-Santana, C.; Larrañaga, P.; Bielza, C. Feature subset selection in data-stream environments using asymmetric hidden Markov models and novelty detection. *Neurocomputing* **2023**, *554*, 126641. [CrossRef]
6. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
7. Brown, G.; Pocock, A.; Zhao, M.-J.; Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 2, 27–66.
8. Wan, J.; Chen, H.; Li, T.; Huang, W.; Li, M.; Luo, C. R2CI: Information theoretic-guided feature selection with multiple correlations. *Pattern Recognit.* **2022**, *127*, 108603. [CrossRef]
9. Wan, J.; Li, X.; Wang, Z.; Zhao, J. Uncertainty-aware significance and interaction-enhanced feature selection: A fuzzy multi-granularity information perspective. *Inf. Process. Manag.* **2026**, *63*, 6, 104775. [CrossRef]
10. Ahmadi, N.; Majidnezhad, V.; Zarei, B. BTCOA: A two stage feature selection framework combining graph theory and trigonometric functions for HDLSS biomedical data. *J. Supercomput.* **2026**, *82*, 450. [CrossRef]
11. Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; Lu, H. Unsupervised feature selection using nonnegative spectral analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012; Volume 26, pp. 1026–1032.
12. Shang, R.; Zhang, W.; Lu, M.; Jiao, L.; Li, Y. Feature selection based on non-negative spectral feature learning and adaptive rank constraint. *Knowl.-Based Syst.* **2022**, *236*, 107749. [CrossRef]
13. Karami, S.; Saberi-Movahed, F.; Tiwari, P.; Marttinen, P.; Vahdati, S. Unsupervised feature selection based on variance–covariance subspace distance. *Neural Netw.* **2023**, *166*, 188–203. [CrossRef]
14. Feng, J.; Jiao, L.; Liu, F.; Sun, T.; Zhang, X. Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognit.* **2016**, *51*, 295–309. [CrossRef]
15. Wang, S.; Tang, J.; Liu, H. Embedded unsupervised feature selection. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29, p. 1.
16. Wang, Y. Unsupervised representative feature selection algorithm based on information entropy and relevance analysis. *IEEE Access* **2018**, *6*, 45317–45324. [CrossRef]
17. Liu, T.; Hu, R.; Zhu, Y. Completed sample correlations and feature dependency-based unsupervised feature selection. *Multimed. Tools Appl.* **2023**, *82*, 15305–15326. [CrossRef]
18. Seo, W.; Kim, D.-W.; Lee, J. Generalized information-theoretic criterion for multi-label feature selection. *IEEE Access* **2019**, *7*, 122854–122863. [CrossRef]
19. Fano, R.M.; Hawkins, D. Transmission of information: A statistical theory of communications. *Am. J. Phys.* **1961**, *29*, 11, 793–794. [CrossRef]
20. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. In Proceedings of the Advances in Neural Information Processing Systems 18, Vancouver, Canada, 5–8 December 2005; pp. 507–514.
21. Cai, D.; Zhang, C.; He, X. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 333–342.

22. Yuan, A.; You, M.; He, D.; Li, X. Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection. *IEEE Trans. Cybern.* **2020**, *52*, 6, 5522–5534. [CrossRef]
23. Zhang, R.; Zhang, Y.; Li, X. Unsupervised feature selection via adaptive graph learning and constraint. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 3, 1355–1362. [CrossRef]
24. Jiang, K.; Cao, T.; Zhu, L.; Sun, Q. Adaptive and flexible l_1-norm graph embedding for unsupervised feature selection. *Appl. Intell.* **2024**, *54*, 22, 11732–11751. [CrossRef]
25. Xiang, L.; Chen, H.; Yin, T.; Horng, S.J.; Li, T. Unsupervised feature selection based on bipartite graph and low-redundant regularization. *Knowl.-Based Syst.* **2024**, *302*, 112379. [CrossRef]
26. Xu, Y.; Xie, X.; Jiang, X.; Xiong, Y. Multi-view unsupervised feature selection based on graph discrepancy learning. *Neurocomputing* **2025**, *656*,131487 . [CrossRef]
27. Lifshitz, S.; Lindenbaum, O.; Mishne, G.; Meir, R.; Benisty, H. Unsupervised feature selection through group discovery. *arXiv* **2025**, arXiv:2511.09166.
28. Yang, Y.; Shen, H.T.; Ma, Z.; Huang, Z.; Zhou, X. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
29. Hu, Z.; Wang, J.; Zhang, K.; Pedrycz, W.; Pal, N.R. Bi-level spectral feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 6597–6611. [CrossRef]
30. Hassanieh, W.; Chehade, A. Selective deep autoencoder for unsupervised feature selection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 20–27 February 2024; Volume 38, pp. 12322–12330
31. Huang, S.-L.; Xu, X.; Zheng, L. An information-theoretic approach to unsupervised feature selection for high-dimensional data. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 157–166. [CrossRef]
32. Lim, H.; Kim, D.-W. Pairwise dependence-based unsupervised feature selection. *Pattern Recognit.* **2021**, *111*, 107663. [CrossRef]
33. Xu, H.; Xie, Y.; Liu, X. A global unsupervised feature selection method based on fuzzy mutual information. *Symmetry* **2026**, *18*, 633. [CrossRef]
34. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
35. Williams, C.K.I.; Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*; MIT Press: Cambridge, MA, USA, 2001; pp. 682–688.
36. Drineas, P.; Mahoney, M.W. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **2005**, *6*, 2153–2175.
37. Kumar, S.; Mohri, M.; Talwalkar, A. Sampling methods for the Nyström method. *J. Mach. Learn. Res.* **2012**, *13*, 981–1006.
38. Boutsidis, C.; Mahoney, M.W.; Drineas, P. An improved approximation algorithm for the column subset selection problem. In Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms, New York, NY, USA, 4–6 January 2009; pp. 968–977.
39. Mitra, P.; Murthy, C.A.; Pal, S.K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 301–312. [CrossRef]
40. Liu, W.; He, J.; Chang, S.-F. Large graph construction for scalable semi-supervised learning. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 679–686.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.