Negar Kiyavash and Joris Mooij
Program Chairs

Dear Chair members,

This rebuttal letter is in response to the comments of the reviewers on our manuscript entitled "Unsupervised Feature Selection towards Maximizing Pattern Discrimination Power". We would like to express our sincere gratitude to the reviewers for their valuable comments and suggestions. We have carefully considered all the comments.

Sincerely,

Authors of the manuscript

Answers to Reviewers

Title: Unsupervised Feature Selection towards Maximizing Pattern Discrimination Power

We are grateful for the constructive comments of the reviewers. Below, we provide specific answers and explanations with regard to those comments.

1. Point-by-point response to Reviewer GMsc

(Comment 1-1) "...search for a feature subset by optimizing": What is the motivation for maximizing the entropy? Only explanation is found later.
(Answer 1-1) The primary motivation behind maximizing entropy is to fully utilize the distinctiveness of patterns, leading to improved information retrieval [8], such as enhancing the process of taxonomy construction for each pattern. This method aligns with information theory principles by selecting features that offer maximal informational value, thereby improving pattern discrimination power. Specifically, constructing the taxonomic tree [4] starting from the root node and distributing patterns into the tree based on prioritized features with the highest entropy enables a significant reduction in the maximum depth of the tree. This effectiveness arises because features selected in order of descending entropy are arranged from the root node downwards, allowing the values of each pattern associated with these features to be evenly distributed across the tree. Such an organization ensures that the tree expands in a balanced manner, simplifies analytical processes by making the data structure more compact, and enhances the clarity and efficiency of data interpretation.

(Comment 1-2) "the original feature set F should have the largest entropy": how does the large entropy leads to more discriminative power?
(Answer 1-2) A proof is provided to establish the relationship between the entropy of a feature subset and its SDP, showing that a decrease in SDP leads to a decrease in entropy.

First, SDP in the manuscript can be reformulated as:

$$SDP(W) = \frac{1}{|W|} \cdot \sum_{i=1}^{|W|} \left[\!\!\left[\left(\sum_{j=1}^{i} [\![w_i = w_j]\!] \right) = 0 \right]\!\!\right] \tag{1}$$

where W is the dataset comprised of the feature subset S, w_i is the i-th pattern, and $[\![\cdot]\!]$ yields one if the proposition stated in the brackets is true and returns zero otherwise. The maximum SDP value is 1, indicating all patterns are distinct.

Proposition 1. *For each pattern in W is distinct from each other, the joint entropy $H(S)$ is maximized.*

Proof. The joint entropy $H(S)$ of the dataset is a measure of the uncertainty within the feature subset S. It is quantified as $H(S) = -\sum_{i=1}^{|W|} P(w_i) \log P(w_i)$, where $P(w_i)$ represents the probability of occurrence of the i-th pattern within the dataset. Here, each $P(w_i)$ can be denoted as y_i, simplifying our expression to function terms where $f(y) = -y \log y$ with $y = P(w_i)$.

Given the concavity of the function $f(y) = -y \log y$, Jensen's Inequality [2] allows us to derive an upper bound for the sum $\sum_{i=1}^{|W|} f(y_i)$, which is directly applicable to our entropy calculation:

$$\sum_{i=1}^{|W|} f(P(w_i)) = \sum_{i=1}^{|W|} -P(w_i) \log P(w_i) \leq |W| \cdot f\left(\frac{1}{|W|}\right) \tag{2}$$

This formulation bounds the joint entropy as:

$$-\sum_{i=1}^{|W|} P(w_i) \log P(w_i) \leq \log |W| \tag{3}$$

When all patterns in W are distinct, this condition indicates a uniform distribution of occurrences, with $P(w_i) = \frac{1}{|W|}$ for all i, thereby transforming our inequality into an equality:

$$H(S) = \log |W| \tag{4}$$

Therefore, when each pattern in the dataset W is distinct, the joint entropy $H(S)$ of the dataset is maximized. □

Because the proposed score function aims to maximize $H(S)$, denoted in Equation (2) in the original manuscript, the theoretical proof is provided to establish the relationship between the entropy of a feature subset and SDP, showing that a decrease in SDP leads to a decrease in entropy. First, let $k = (1 - SDP(W)) \cdot |W|$ as the number of patterns within the dataset W that are

identical to at least one other pattern, considering each pattern w_n and comparing it with all other patterns w_m where $m < n$. For instance, given $k = 5$ and $|W| = 20$, then a minimum of 6 patterns are identical from to each other, as illustrated by a sequence such as $\{1, 2, 3, 4, \ldots, 14, 15, 15, 15, 15, 15, 15\}$, where the last six entries are indistinctive. On the other hand, the maximum case of 10 indistinctive patterns within the dataset could be exemplified by $\{1, 2, 3, 4, \ldots, 10, 11, 11, 12, 12, 13, 13, 14, 14, 15, 15\}$. Based on the above examples, the upper and lower bounds of $H(S)$ can be represented as:

Lemma 1. *The joint entropy $H(S)$ of a dataset W can be bounded as follows:*

$$-\frac{|W|-(k+1)}{|W|}\cdot\log\frac{1}{|W|}-\frac{k+1}{|W|}\cdot\log\frac{k+1}{|W|}\leq H(S) \tag{5}$$

$$\leq -\frac{|W|-2k}{|W|}\cdot\log\frac{1}{|W|}-\frac{2k}{|W|}\cdot\log\frac{2}{|W|} \tag{6}$$

where $k = (1 - SDP(W)) \cdot |W|$.

Proof. The number of indistinctive patterns in W ranges from $k + 1$ to $2k$, as illustrated in the provided example. Therefore, the entropy of $|W| - 2k$ distinctive patterns, $-\frac{|W|-2k}{|W|} \cdot \log \frac{1}{|W|}$, is constant. Considering the remaining $2k$ patterns, the entropy is maximized when the patterns are uniformly distributed, as mentioned in Proposition 1. This condition is achieved when there are k pairs of patterns, with each pair being identical, yielding $-\frac{2k}{|W|} \cdot \log \frac{2}{|W|}$. Equation (6) represents the upper bound of the joint entropy $H(S)$ in this scenario. Conversely, the entropy is minimized when the distribution is skewed, as in the case of $k + 1$ indistinctive patterns and $|W| - k - 1$ distinctive patterns, leading to the lower bound in Equation (5). □

According to Lemma 1, the relationship between the SDP and entropy can be established, demonstrating that a decrease in SDP, which corresponds to an increase in k, leads to a decrease in entropy. Given the uncertainty regarding the precise number of indistinctive patterns as k increases, we construct our proof by focusing on the upper bounds of $H(S)$ for both k and $k + 1$.

Theorem 1. *The upper bound of $H(S)$ for k is greater than the upper bound of $H(S)$ for $k + 1$.*

Proof. By applying Lemma 1, the upper bound of $H(S)$ for $k+1$ represents as follows:

$$-\frac{|W|-2(k+1)}{|W|}\cdot\log\frac{1}{|W|}-\frac{2(k+1)}{|W|}\cdot\log\frac{2}{|W|} \tag{7}$$

Subtracting the upper bound of $H(S)$ for k, as detailed in Equation 5, from the upper bound of $H(S)$ for $k+1$, we obtain:

$$\begin{aligned}
&-\frac{|W|-2k}{|W|}\cdot\log\frac{1}{|W|}-\frac{2k}{|W|}\cdot\log\frac{2}{|W|}\\
&-\left(-\frac{|W|-2(k+1)}{|W|}\cdot\log\frac{1}{|W|}-\frac{2(k+1)}{|W|}\cdot\log\frac{2}{|W|}\right)\\
&=\frac{2}{|W|}\\
&\geq 0
\end{aligned} \tag{8}$$

□

Because the upper bound of $H(S)$ for k is greater than that for $k+1$, as illustrated in Theorem 1, a decrease in SDP implies a decrease in entropy.

(Comment 1-3) It is not clear what objective of unsupervised FS is. Is it a subset that preserve the most information, instead of discriminative power?
(Answer 1-3) We provide theoretical proof of the relationship between the entropy of a feature subset and its SDP, showing that a decrease in SDP implies a decrease in entropy, as elaborated in Answer 1-2.

(Comment 1-4) The theoretical proof seem to be borrowed from existing papers, without explanation. It is not clear if anything new here. At best, they are not self contained.
(Answer 1-4) We provide a detailed explanation of Equations (6) and (7) in the manuscript to clarify the theoretical proof. Equations (6) [3] and (7) [5] in the manuscript are pivotal for the decomposition of joint entropy in the objective function of our proposed UFS method. First, Equation (6) represents the upper and lower bound of $U_k(S')$ for a given S and the cardinality k. Specifically, the upper bound is determined by the lower cardinality $k-1$ of S', which can be expressed as $U_{k-1}(S')$. Therefore, the high-dimensional joint

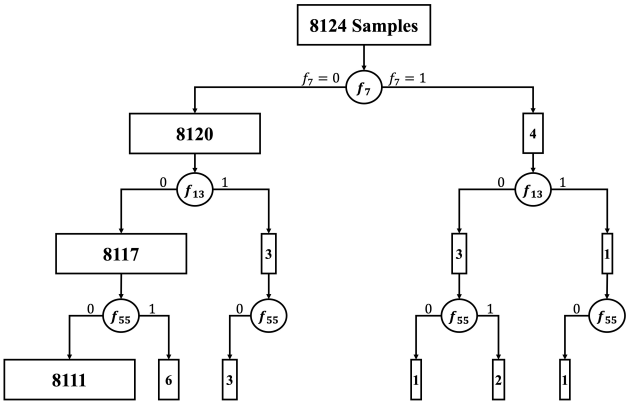
entropy $U_k(S')$ can be decomposed into the lower-dimensional joint entropy by applying Equation (6) recursively as follows:

$$H(S) \leq \frac{1}{n-1} U_{n-1}(S') \leq \frac{1}{n-1} \cdot \frac{1}{n-2} U_{n-2}(S') \leq \cdots \tag{9}$$

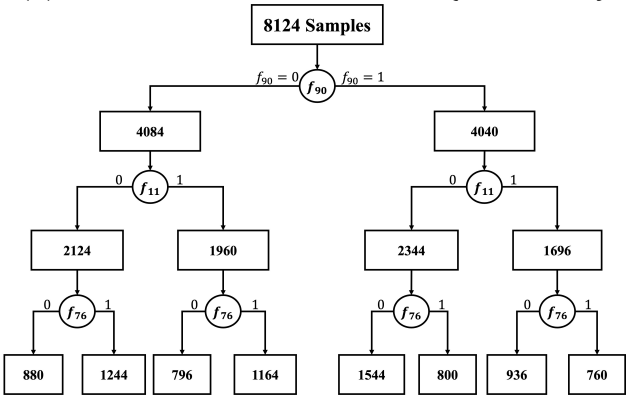
Equation (7) represents the general form of the above recursive decomposition that can directly calculate the upper bound of $H(S)$ with k-cardinality. It is important to highlight that the score function we propose is specifically designed for the UFS method, in contrast to the approaches cited in our manuscript. While these referenced papers focus on minimizing the entropy of the feature subset to reduce feature redundancy, our approach aims to maximize the entropy of the feature subset, thereby amplifying the pattern discrimination power. We would like to add this explanation to the revised manuscript to enhance the clarity of the theoretical proof.

(Comment 1-5) Eq 9 to 12: are they the same approximation or further approximations from each equation above?
(Answer 1-5) We appreciate the reviewer's comment to clarify the approximation process of the score function. Equation (10) represents an additional simplification of Equation (9) by omitting the coefficient $\frac{1}{|S|}$ and the constant term $U_2(S')$. Subsequently, Equation (11) retains the same level of approximation as Equation (10) since $U(f^+, S')$ encompasses the sum of the joint entropies between f^+ and every feature in S', as delineated in Definition 1. Finally, Equation (12) fundamentally parallels the approximation in Equation (11), as when $S = \emptyset$, it results in the selection of only f^+, the feature with the maximum entropy value.

(Comment 1-6) Experiment evaluation metric: authors did not clarify the objective, hence i'm not sure what would maximizing SDP and entropy would help in any downstream task. What are some applications? Authors should also test the selected features' performance on these applications.
(Answer 1-6) The concept of maximizing $H(S)$ can be used to build an effective taxonomic tree for a system because it can reduce the number of discriminators. The proposed method selects features iteratively based on their scores from the score function that maximizes the entropy of the feature subset. Considering that the highest entropy is observed in a uniform distribution when patterns are distributed using the selected features in a tree structure, this strategy enables the construction of a tree that closely



(a) Entropy Minimization $S = \{f_7, f_{13}, f_{55}\}$



(b) Entropy Maximization $S = \{f_{90}, f_{11}, f_{76}\}$

Figure 1: Taxonomic trees constructed by selected features from the proposed method (Entropy Maximization) and the entropy minimization method.

approximates a balanced tree. This aspect can be especially useful in information retrieval systems that balanced binary tree reduces the number of comparisons [6, 7].

We conducted experiments on the binary Mushrooms dataset to assess the effectiveness of taxonomic trees constructed via the proposed entropy maximization method compared to the conventional entropy minimization approach. For the entropy minimization method, features were selected using the same algorithm as in the original manuscript but with the objective of minimizing entropy within the feature subset. Both strategies selected three features from the dataset, and constructed the trees starting from the root node in the order of selection, resulting in trees with a depth of three levels.

Patterns were then assigned to nodes based on their feature values.

Figure 1 depicts the constructed taxonomic trees. The numbers in squares indicate the number of assigned patterns at each node, whereas the circles highlight the feature at each division. To visually represent the quantity of patterns per node, the width of each node was adjusted logarithmically in proportion to the number of patterns it contains. The tree derived from the entropy minimization method revealed a significantly skewed structure, with the majority of patterns assigned in the leftmost node. In contrast, the tree resulting from the proposed entropy maximization method exhibited a more balanced structure, approximating a uniform distribution of patterns across the nodes. Given the skewed nature of the tree resulting from the entropy minimization method, which may require additional comparisons to locate a specific pattern, the proposed entropy maximization method presents itself as a potentially more efficient solution, such as information retrieval systems.

(Comment 1-7) What is SDP and entropy for the entire feature set? In addition, it is not clear if the baseline methods also seek to maximize the entropy or have different objectives.
(Answer 1-7) Both the SDP and entropy exhibit monotonic increases with the inclusion of additional features. The monotonic increase of entropy is substantiated by the findings of Artstein et al. [1]. As for the SDP, its monotonicity can be straightforwardly demonstrated: the number of distinct patterns remains unchanged only in instances where the newly selected feature is identical to one of the selected features. Conversely, when a new distinct feature is selected, there is an increase in the number of distinct patterns. Therefore, the SDP and entropy of the entire feature set are maximized when all features are selected.

To assess the efficacy of our proposed method, we performed a comparative analysis against four baseline methods. EGC and U2 methods select features based on manifold learning to preserve the dataset's local structure, taking into account the weight of each feature in the learning process. Similarly, VCSD employs a conventional UFS framework similar to that of EGC and U2, yet it utilizes variance-correlation distance between the entire feature set and a subspace derived from a feature subset to identify statistically significant features. Meanwhile, FMI shares the same objective of using entropy maximization as our proposed method, yet its score function is derived from a heuristic method. In this research, the proposed method has demonstrated its effectiveness in maximizing the SDP through comparative analyses with

majorly researched manifold learning-based UFS methods and with entropy maximization scoring functions devised heuristically.

2. Point-by-point response to Reviewer aJ19

(Comment 2-1) The phrase "For compared methods, all datasets are normalized to the range of [0, 1] before the FS process, as the authors suggested" doesn't seem supported when reading the original papers for the other methods. They used normalized versions of the metrics sometimes but that doesn't mean that the data are normalized, or at least I don't interpret that. I am concerned that the superiority of the performance is because of this transformation, so i suggest to apply the other methods (EGC, U2, VCSD and FMI) with the original data, non-normalize and see if they behave better.
(Answer 2-1) We conducted additional experiments on the five numerical datasets to evaluate the performance of the compared methods without normalization. Furthermore, as in Comment 2-5, all datasets are discretized using the equal-width binning method with 10 bins for the proposed method.

Table 1: Comparison results of five UFS methods in terms of Entropy, SDP without normalization and equal-width binning method with 10 bins.

Dataset	Entropy					SDP				
	Proposed	EGC	U2	VCSD	FMI	Proposed	EGC	U2	VCSD	FMI
Alzheimer	**5.76**	0.29	3.36	4.63	5.41	**0.44**	0.06	0.13	0.31	**0.44**
Lsvt	**6.50**	0.46	1.79	3.13	2.50	**0.78**	0.08	0.09	0.15	0.13
Umist	**9.13**	9.03	9.10	9.12	9.07	**0.98**	0.94	0.97	0.97	0.95
WarpAR10P	**7.02**	**7.02**	6.99	6.50	6.90	**1.00**	**1.00**	0.98	0.87	0.95
WarpPIE10P	**7.71**	7.54	6.89	7.25	7.67	**1.00**	0.93	0.77	0.89	0.98
Avg. Rank	**1.10**	3.90	3.80	3.20	3.00	**1.20**	3.90	3.80	3.20	2.90

Table 1 presents the comparative outcomes of five UFS methods relative to entropy and the SDP, excluding the influence of normalization. Even under these conditions, the proposed method consistently surpasses other methods regarding both entropy and SDP across five datasets. As shown in the table, the proposed method achieves the best performance across all datasets in terms of entropy and SDP, underscoring its effectiveness in maximizing pattern discrimination power. Nevertheless, we acknowledge the reviewer's comments for the rigorous evaluation of the proposed method and will include the results in the revised manuscript.

(Comment 2-2) There is no cpu time reported and not an exhaustive complexity study. I think reporting execution time would be of interest to compare all the alternative techniques.

Table 2: Execution time in seconds for conducted UFS methods.

Dataset	Proposed	EGC	U2	VCSD	FMI
ALLAML	375.65	221.69	5978.10	1138.32	1486.96
Alzheimer	1165.01	662.63	23696.24	3816.60	5357.24
Arcene	303.85	149.41	4204.49	741.45	1587.23
Audiology	0.18	0.01	1.96	0.46	0.74
Ba	0.06	0.06	1.05	0.04	0.17
Chess	2.28	0.10	7.33	3.87	13.82
CLL_SUB_111	775.53	440.78	15112.69	2839.68	3555.26
Coil20	0.08	0.07	0.07	0.27	4.89
Colon	7.83	3.59	16.22	6.59	650.64
Leukemia	0.20	27.34	9.36	12.26	514.60
Lsvt	92.26	4.93	49.98	20.90	8496.83
Lung	266.99	136.40	3663.49	768.40	911.85
Lymphoma	388.49	218.35	5872.40	1102.46	1600.35
Madelon	0.99	0.15	6.03	2.46	4.38
Mushrooms	162.93	22.47	569.18	166.40	1085.49
Nci9	148.46	40.09	954.09	263.30	660.76
Nursery	34.69	21.69	60.47	19.09	12031.08
Pdspeech	2.06	32.62	28.36	52.79	5279.50
Promoters	659.48	441.33	16076.21	2555.64	3328.74
Prostate_GE	0.40	1389.47	184.80	194.09	22083.95
SCADI	18.67	1.17	19.37	10.07	712.49
Semeion	0.51	0.03	3.97	1.28	1.85
SPECT	5.58	3.64	11.89	5.70	992.32
Splice	14.00	29.64	55.27	21.94	7801.44
Tox171	426.87	120.82	3212.45	637.41	2417.68
Tic-Tac-Toe	0.04	0.66	1.16	0.68	5.77
Umist	16.96	0.59	14.52	7.02	479.49
WarpAR10P	68.45	7.59	191.55	78.93	264.45
WarpPIE10P	99.19	7.64	207.88	80.65	491.89
Yaleb	185.28	18.93	127.98	31.47	35409.53
Avg. Rank	2.20	1.47	4.07	2.70	4.57

(Answer 2-2) In Table 2, we present a comparative analysis of the execution times for conducted UFS methods. Execution times are reported in seconds and reflect the computational effort required to process each dataset, and the

average rank for each method is provided at the bottom of the table. The experiments were conducted on a system equipped with a 13th Generation Intel® Core™ i9-13900K processor, clocked at 3.00 GHz. As summarized in Table 2, experimental results indicate that the execution times vary significantly across different datasets and algorithms. The proposed method ranks as the second-fastest method on average after EGC, which underscores its efficiency and potential applicability in a wide range of datasets. Furthermore, for algorithms such as EGC, which rely on specific parameter settings, it is crucial to highlight that the time to achieve optimal results can vary significantly depending on the number of parameters adjusted. In some instances, the execution time can increase multiple times depending on the complexity of the parameter space being navigated.

(Comment 2-3) I would appreciate that the authors indicate if the following references are appropriate to be included:
(Answer 2-3) We appreciate the relevant references to enhance the quality of the manuscript, and we are likely to include the references in the revised manuscript.

(Comment 2-4) When you say :'In our experiments, we set k to two because it is the minimum value for the score function being a multivariate feature filter 2', I would like to know the implications (in terms of the quality of the results and on complexity) of using values of k larger than 2.
(Answer 2-4) We provide a detailed explanation of the implications of using values of k larger than 2 by giving a concrete example by introducing a newly instantiated score function when $k = 3$. First, the score function for the proposed method can be rewritten from Equation (7) in the original manuscript as:

$$\begin{aligned} J &\approx \arg\max_{f^+} \left(\sum_{i=1}^{b} \frac{i}{|S| + 1 - i} \right) U_3(\{S', f^+\}') \\ &= \arg\max_{f^+} \frac{1}{|S| \cdot (|S| - 1)} U_3(\{S', f^+\}'), \end{aligned} \tag{10}$$

where $b = min(|S| + 1 - 3, 3 - 1) = 2$. Equation (10) can be rewritten as:

$$J \approx \arg\max_{f^+} \sum_{f_i \in S} \sum_{f_j \in S} H(f^+, f_i, f_j). \tag{11}$$

by the identical process in the original manuscript. Because the newly instantiated score function requires at least two features in S, the first and second features are selected based on Equations (11) and (12) in the original manuscript.

Algorithm 1 Incremental Search for the Proposed Method when $k = 3$

1: $f^+ \leftarrow \arg\max_{f^+ \in F} H(f^+)$ ▷ Select the first feature
2: $S \leftarrow \{f^+\}$
3: $f^+ \leftarrow \arg\max_{f^+ \in F-S} \sum_{f \in S} H(f^+, f)$ ▷ Select the second feature
4: $S \leftarrow \{f^+\}$
5: **while** $|S| < n$ **do**
6: $\quad f^+ \leftarrow \arg\max_{f^+ \in F-S} \sum_{f_i \in S} \sum_{f_j \in S} H(f^+, f_i, f_j)$
7: $\quad S \leftarrow S \cup \{f^+\}$
8: **end while**

The Algorithm 1 depicts the incremental search process of the proposed method when $k = 3$. The computational complexity of the proposed method expands to $O(n + n^2 + n^3) = O(n^3)$ because n, n^2, n^3 unit times are consumed for calculating entropy values. Compared to the originally proposed method with $k = 2$, the proposed method with $k = 3$ requires additional computational resources to calculate the entropy values because Equation (11) calculates joint entropies among three features. The new score function can capture more complex relationships among features because it calculates joint entropies of candidate features with all pairs of selected features. However, the computational complexity increases significantly compared to the original proposed method with $k = 2$, which is $O(n^2)$. Furthermore, an estimation of the joint entropy between high-dimensional features often requires a large number of patterns to achieve reliable approximations [3].

(Comment 2-5) When you decide to use this discretization: 'equal-width binning method [Talukdar et al.,2018] where the number of bins is set to the number of classes for each numerical dataset.', two issues are implicitly raised:

- 1.- why this discretization method? why not equal-frequency? Why don't you use a (distinct) rule to select the number of bins? This decision can change the results, so proper justification has to be given

- 2.- aren't you somehow transforming the problem into supervised? I mean, in the moment the number of bins for the discretization to be applied to the numerical variables is decided by the number of classes, we are not using an Unsupervised approach anymore.

(Answer 2-5) The reason we avoid the equal-frequency binning method is due to its effect of allocating an equal number of patterns to each bin, which results in a uniformly distributed transformation of the dataset. Therefore, when selecting the first feature during the incremental search, scores for all features become identical which can degrade the quality of S. We appreciate the valuable comment on the discretization method, and we acknowledge that the conducted discretization method can be considered as a supervised approach. Therefore, we will adjust the discretization method to ensure that the proposed method remains a UFS method in the revised manuscript. Currently, we report the results of the proposed method with the equal-width binning method with 10 bins in Answer 2-1.

(Comment 2-6) Could you provide some insight on using values for k in Equation (7) bigger than 2?
(Answer 2-6) We provided a detailed explanation of the implications of using values of k bigger than 2 in Answer 2-4.

3. Point-by-point response to Reviewer EZKd

(Comment 3-1) The toy example (Table I) is interesting, but it should present some quantitative evaluation.
(Answer 3-1) We describe a quantitative evaluation of the toy example.

Table 3: Joint entropies of all feature pairs in Table 1 from the original manuscript.

Selected Features	Entropy	min/max
$[f_1, f_2]$	1.252	min
$[f_1, f_3]$	1.918	
$[f_1, f_4]$	1.459	
$[f_2, f_3]$	2.252	
$[f_2, f_4]$	1.459	
$[f_3, f_4]$	2.585	max

Table 3 presents the joint entropies of all feature subsets for $|S| = 2$ within the toy dataset from the original manuscript. The selected features $[f_1, f_2]$ exhibit the lowest joint entropy, indicating that this feature pair is the least informative and discriminative. As a result, four distinct patterns from the toy dataset become indistinguishable when selecting these features. Conversely, the selected features $[f_3, f_4]$ reveal the highest joint entropy, and all patterns remain distinguishable when these features are selected.

(Comment 3-2) Another point concerns the experimental section. The authors could consider a labeled dataset and treat it as unlabeled. After finding the most representative features, a supervised classification algorithm could be used to compute the accuracy, which is way more insightful than an entropy value.
(Answer 3-2) To evaluate the classification performance of the selected features, we conducted experiments on labeled datasets and attached them in the Appendix section.

Table 4 summarizes the classification accuracy of features selected by both the proposed and compared methods, evaluated using naïve Bayes and decision tree classifiers. Specifically, the classification accuracy is measured by the 10-fold cross-validation with the naïve Bayes classifier and the decision tree classifier, which were trained by the data composed of the selected features. In the case of the naïve Bayes classifier, the proposed method achieved superior classification accuracy on 14 out of 30 datasets, with an

Table 4: Comparison results of classification accuracy performance based on the feature subsets selected by the five UFS methods.

Dataset	Naïve Bayes					Decision Tree				
	Proposed	EGC	U2	VCSD	FMI	Proposed	EGC	U2	VCSD	FMI
ALLAML	0.57 ± 0.22	**0.72 ± 0.20**	0.67 ± 0.11	0.65 ± 0.24	0.71 ± 0.10	0.56 ± 0.23	**0.68 ± 0.21**	0.61 ± 0.13	0.65 ± 0.24	0.59 ± 0.19
Alzheimer	**0.76 ± 0.12**	0.41 ± 0.08	0.52 ± 0.14	0.71 ± 0.09	0.68 ± 0.13	**0.75 ± 0.09**	0.42 ± 0.09	0.45 ± 0.13	0.65 ± 0.08	0.69 ± 0.12
Arcene	0.60 ± 0.14	0.45 ± 0.16	**0.62 ± 0.15**	0.56 ± 0.14	0.58 ± 0.12	0.58 ± 0.08	0.49 ± 0.09	0.56 ± 0.15	0.55 ± 0.14	**0.60 ± 0.12**
Audiology	**0.63 ± 0.14**	0.52 ± 0.10	0.50 ± 0.10	0.35 ± 0.12	0.30 ± 0.12	**0.66 ± 0.10**	0.52 ± 0.11	0.49 ± 0.10	0.33 ± 0.11	0.38 ± 0.12
Ba	**0.25 ± 0.03**	0.07 ± 0.02	0.07 ± 0.03	0.21 ± 0.03	0.24 ± 0.05	0.21 ± 0.04	0.08 ± 0.02	0.08 ± 0.02	0.21 ± 0.03	**0.25 ± 0.03**
Chess	**0.75 ± 0.03**	0.71 ± 0.02	0.67 ± 0.01	0.63 ± 0.02	0.62 ± 0.03	**0.82 ± 0.02**	0.73 ± 0.02	0.70 ± 0.02	0.67 ± 0.01	0.66 ± 0.02
CLL_SUB_111	0.62 ± 0.19	0.32 ± 0.12	0.46 ± 0.12	0.57 ± 0.16	**0.65 ± 0.16**	0.56 ± 0.18	0.38 ± 0.14	0.33 ± 0.15	**0.59 ± 0.12**	0.49 ± 0.13
Coil20	**0.85 ± 0.03**	0.10 ± 0.02	0.36 ± 0.05	0.78 ± 0.04	0.82 ± 0.04	0.81 ± 0.04	0.10 ± 0.02	0.45 ± 0.03	0.79 ± 0.03	**0.81 ± 0.04**
Colon	0.57 ± 0.20	0.53 ± 0.22	**0.66 ± 0.22**	0.59 ± 0.15	0.53 ± 0.22	0.42 ± 0.15	0.50 ± 0.24	0.44 ± 0.21	0.32 ± 0.10	**0.60 ± 0.13**
Leukemia	**0.80 ± 0.15**	0.54 ± 0.17	0.67 ± 0.19	0.65 ± 0.14	0.69 ± 0.15	0.44 ± 0.04	0.13 ± 0.02	0.28 ± 0.03	0.29 ± 0.04	**0.47 ± 0.04**
LSVT	**0.77 ± 0.06**	0.66 ± 0.15	0.66 ± 0.15	0.70 ± 0.12	0.60 ± 0.10	**0.73 ± 0.13**	0.48 ± 0.22	0.63 ± 0.19	0.69 ± 0.12	0.67 ± 0.16
Lung	0.80 ± 0.07	0.74 ± 0.08	0.75 ± 0.08	0.80 ± 0.08	**0.86 ± 0.06**	0.69 ± 0.11	0.66 ± 0.15	0.67 ± 0.15	**0.79 ± 0.07**	0.60 ± 0.13
Lymphoma	0.58 ± 0.18	0.51 ± 0.22	0.43 ± 0.11	0.75 ± 0.15	**0.80 ± 0.13**	0.80 ± 0.12	0.66 ± 0.10	0.62 ± 0.09	0.76 ± 0.07	**0.83 ± 0.08**
Madelon	0.56 ± 0.02	0.47 ± 0.02	0.49 ± 0.02	0.47 ± 0.02	**0.62 ± 0.03**	0.50 ± 0.17	0.47 ± 0.22	0.48 ± 0.16	0.58 ± 0.16	**0.64 ± 0.13**
Mushrooms	0.89 ± 0.01	**0.93 ± 0.01**	**0.93 ± 0.01**	0.80 ± 0.02	0.90 ± 0.03	0.54 ± 0.02	0.47 ± 0.02	0.51 ± 0.03	0.47 ± 0.02	**0.69 ± 0.03**
Nci9	**0.17 ± 0.16**	0.07 ± 0.09	**0.17 ± 0.14**	0.10 ± 0.09	**0.17 ± 0.16**	**0.98 ± 0.00**	0.94 ± 0.01	0.94 ± 0.01	0.86 ± 0.01	0.96 ± 0.01
Nursery	0.76 ± 0.01	0.42 ± 0.01	0.51 ± 0.01	**0.76 ± 0.01**	0.76 ± 0.01	0.18 ± 0.12	0.10 ± 0.12	**0.20 ± 0.15**	0.15 ± 0.12	0.13 ± 0.17
Pdspeech	**0.75 ± 0.03**	**0.75 ± 0.03**	0.74 ± 0.03	0.74 ± 0.04	0.69 ± 0.06	**0.79 ± 0.01**	0.41 ± 0.01	0.54 ± 0.01	0.76 ± 0.01	0.78 ± 0.01
Promoters	**0.93 ± 0.11**	0.67 ± 0.19	0.48 ± 0.16	0.92 ± 0.13	0.88 ± 0.13	0.37 ± 0.07	0.36 ± 0.05	0.15 ± 0.07	0.35 ± 0.08	**0.45 ± 0.06**
Prostate_GE	0.74 ± 0.10	**0.79 ± 0.14**	0.64 ± 0.13	0.57 ± 0.15	0.57 ± 0.24	0.70 ± 0.08	0.75 ± 0.03	0.74 ± 0.03	0.75 ± 0.03	**0.79 ± 0.04**
SCADI	**0.80 ± 0.17**	0.66 ± 0.14	**0.80 ± 0.15**	0.74 ± 0.11	0.73 ± 0.16	**0.90 ± 0.13**	0.72 ± 0.13	0.53 ± 0.12	0.88 ± 0.11	0.86 ± 0.13
Semeion	0.50 ± 0.03	0.53 ± 0.07	0.28 ± 0.03	0.27 ± 0.04	**0.54 ± 0.04**	0.72 ± 0.13	**0.79 ± 0.13**	0.54 ± 0.17	0.61 ± 0.13	0.55 ± 0.20
SPECT	**0.81 ± 0.06**	0.76 ± 0.06	0.76 ± 0.06	0.81 ± 0.05	0.80 ± 0.07	**0.80 ± 0.15**	0.73 ± 0.18	**0.80 ± 0.15**	0.76 ± 0.17	0.77 ± 0.10
Splice	**0.93 ± 0.02**	0.52 ± 0.02	0.52 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.53 ± 0.05	0.54 ± 0.07	0.31 ± 0.04	0.28 ± 0.03	**0.56 ± 0.05**
Tox171	0.45 ± 0.15	0.30 ± 0.08	0.46 ± 0.11	**0.51 ± 0.15**	0.47 ± 0.08	0.81 ± 0.08	0.80 ± 0.08	0.80 ± 0.08	**0.84 ± 0.08**	0.81 ± 0.07
Tic-Tac-Toe	0.72 ± 0.06	0.65 ± 0.04	0.65 ± 0.04	0.69 ± 0.05	**0.72 ± 0.05**	0.93 ± 0.02	0.52 ± 0.02	0.52 ± 0.02	0.93 ± 0.01	**0.93 ± 0.01**
Umist	0.63 ± 0.08	0.69 ± 0.06	0.83 ± 0.03	0.60 ± 0.07	**0.83 ± 0.07**	**0.52 ± 0.12**	0.36 ± 0.11	0.44 ± 0.13	0.40 ± 0.06	0.48 ± 0.11
WarpAR10P	0.40 ± 0.14	0.18 ± 0.09	0.27 ± 0.15	0.26 ± 0.17	**0.46 ± 0.15**	0.94 ± 0.02	0.69 ± 0.04	0.69 ± 0.04	0.82 ± 0.04	**0.95 ± 0.02**
WarpPIE10P	**0.49 ± 0.11**	0.31 ± 0.06	0.45 ± 0.10	0.24 ± 0.09	0.37 ± 0.10	0.64 ± 0.10	0.69 ± 0.03	**0.80 ± 0.07**	0.67 ± 0.05	0.77 ± 0.08
Yaleb	0.08 ± 0.02	**0.09 ± 0.02**	0.06 ± 0.01	0.07 ± 0.02	0.07 ± 0.01	0.38 ± 0.08	0.20 ± 0.07	0.33 ± 0.13	0.41 ± 0.19	**0.47 ± 0.09**
Avg. Rank	**1.97**	3.70	3.17	3.30	2.57	**2.09**	3.73	3.61	3.09	2.27

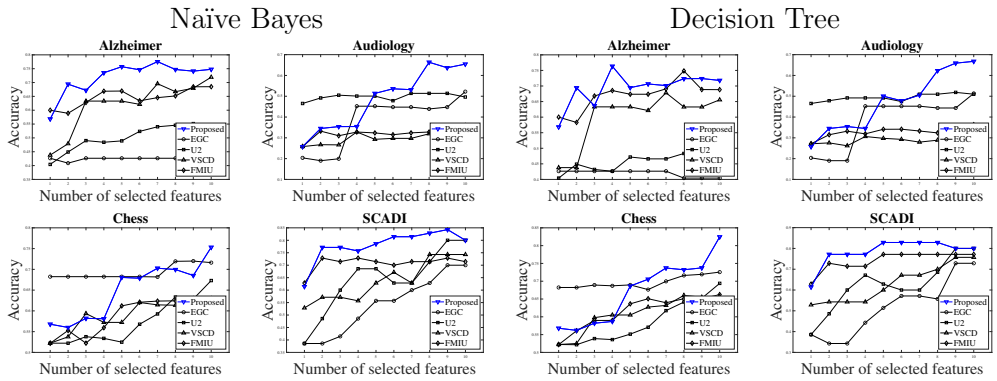


Figure 2: Comparison results of classification accuracy performance according to the number of features selected by the five UFS methods

average rank of 1.97. This performance outperformed that of the next most effective method, FMI, which garnered an average rank of 2.57. For the decision tree classifier, the proposed method also led the field, securing the highest classification accuracy on 9 out of the 30 datasets and an average rank of 2.09. This surpassed the second-best method, EGC, which obtained an average rank of 2.27.

Figure 2 illustrates the classification accuracy achieved by the proposed and compared methods across varying feature sizes on the Alzheimer, Audiology, Chess, and SCADI datasets. In the figure, the blue line symbolizes the proposed method, while the black line represents the comparative methods. Notably, the classification accuracy of the proposed method surpassed that of the compared methods across all four datasets as the feature size increased. This superior performance can be attributed to the feature subset that is optimized towards pattern discrimination power, thereby contributing to improved classification accuracy.

4. Point-by-point response to Reviewer Rd5f

(Comment 4-1) The research motivation and main contributions should be more prominent in the abstract and introduction sections.
(Answer 4-1) We revised the original manuscript to enhance the research motivation and main contributions in the abstract and introduction sections.

First, we emphasize the research motivation in the abstract.
[**Revised Manuscript**]
Abstract: The goal of unsupervised feature selection is to identify a feature subset based on the intrinsic characteristics of a given dataset without user-guided information such as class variables. To achieve this, score functions based on information measures can be used to identify essential features. The major research direction of conventional information-theoretic unsupervised feature selection is to minimize the entropy of the final feature subset. Although the opposite way, i.e., maximization of the joint entropy, can also lead to novel insights, studies in this direction are rare. Specifically, in the field of information retrieval, selecting features to maximize entropy enables pattern discrimination within a dataset with fewer features, thereby facilitating the construction of an effective taxonomy and significantly improving retrieval efficiency. In this work, we first demonstrate how two feature subsets, each obtained by minimizing/maximizing the joint entropy, respectively, are different based on a toy dataset. By comparing these two feature subsets, we show that the maximization of the joint entropy enhances the pattern discrimination power of feature subset. Then, we derive a score function by remedying joint entropy calculation; high-dimensional joint entropy calculation is circumvented by using the low-order approximation. The experimental results on 30 public datasets indicate that the proposed method yields superior performance in terms of pattern discrimination power-related measures.

Next, we highlight the main contributions in the introduction section.
[**Revised Manuscript**] **Introduction:** (End of the introduction section) Finally, we validated the performance of the proposed UFS method using 30 public datasets and confirmed its superiority in terms of pattern discrimination power-related measures. The main contributions of this work are summarized as follows:

- An information-theoretic UFS method is introduced, which maximizes entropy to identify an effective feature subset, thereby significantly enhancing pattern discrimination power within datasets.

- A comparative analysis demonstrates our approach using a toy dataset, illustrating the differences between feature subsets obtained through entropy minimization and maximization, and highlighting the enhanced pattern discrimination power achieved with entropy maximization.
- To tackle the computational challenge of high-dimensional joint entropy, we introduce a novel score function for UFS based on joint entropy decomposition. This approach effectively decomposes high-dimensional entropy into the sum of lower-dimensional terms, offering a practical approximation for entropy calculation.
- The efficacy of the proposed method is validated through extensive testing on 30 public datasets, which confirms its superiority in improving pattern discrimination power over existing UFS methods.

(Comment 4-2) Each symbol in the methods section should be given an explanation, for example the letter d in section 3.2.
(Answer 4-2) The letter d in Section 3.2 represents the number of entire features in the dataset. We will carefully revise the manuscript to provide a detailed explanation of the symbols. For example, d can be clarified as $|F|$ in the revised manuscript.
[**Original Manuscript**]
Let $W \in \mathcal{R}^d$ be the original dataset with d features $F = \{f_1, f_2, \cdots, f_d\}$ and the goal of the UFS is to identify a feature subset S consisting of n features with the optimal pattern discrimination power where n is the number of features to be selected.
[**Revised Manuscript**]
Let $W \in \mathcal{R}^{|F|}$ be the original dataset with d features $F = \{f_1, f_2, \cdots, f_{|F|}\}$ and the goal of the UFS is to identify a feature subset S consisting of n features with the optimal pattern discrimination power where n is the number of features to be selected.

(Comment 4-3) Formulas in the experimental section should also be numbered.
(Answer 4-3) We appreciate the valuable comment to enhance the readability of the manuscript. We will number the formulas in the experimental section like the following example.
(Comment 4-4) The display of experimental results can be richer.

Original Manuscript

Entropy $= H(S)$.

Next, We employed the self-discrimination power test (SDP) that measures the portion of the discriminable data patterns in the dataset based on the feature subset [Clocksin, 2004]. The SDP can be represented as

$$\text{SDP} = \frac{\text{\# of discriminable patterns}}{|W|}.$$

Revised Manuscript

Entropy $= H(S)$. (13)

Next, We employed the self-discrimination power test (SDP) that measures the portion of the discriminable data patterns in the dataset based on the feature subset [Clocksin, 2004]. The SDP can be represented as

$$\text{SDP} = \frac{\text{\# of discriminable patterns}}{|W|}. \quad (14)$$

(Answer 4-4) We will improve the display of experimental results by adjusting the resolutions of the figures and tables in the revised manuscript. We appreciate the valuable comment to enhance the quality of the manuscript.

References

[1] Shiri Artstein, Keith Ball, Franck Barthe, and Assaf Naor. Solution of shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[2] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[3] Jaesung Lee and Dae-Won Kim. Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, 42(4):2013–2025, 2015.

[4] Derek Reiman, Ahmed A Metwally, Jun Sun, and Yang Dai. Popphy-cnn: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE journal of biomedical and health informatics*, 24(10):2993–3001, 2020.

[5] Wangduk Seo, Dae-Won Kim, and Jaesung Lee. Generalized information-theoretic criterion for multi-label feature selection. *IEEE Access*, 7: 122854–122863, 2019.

[6] Zhihua Xia, Xinhui Wang, Xingming Sun, and Qian Wang. A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. *IEEE Trans. Parallel Distrib. Syst.*, 27(2):340–352, 2015.

[7] Ying-Si Zhao and Qing-An Zeng. Secure and efficient product information retrieval in cloud computing. *IEEE Access*, 6:14747–14754, 2018.

[8] Jiaofei Zhong, Weili Wu, Yan Shi, and Xiaofeng Gao. Energy-efficient tree-based indexing schemes for information retrieval in wireless data broadcast. In *International Conference on Database Systems for Advanced Applications*, pages 335–351. Springer, 2011.