On Evaluation of Unsupervised Feature Selection for Pattern Classification

Gyu-Il Kim[1], Dae-Won Kim[2], Jaesung Lee[1]*
[1]Department of Artificial Intelligence, Chung-Ang University, Republic of Korea
[2]School of Computer Science and Engineering, Chung-Ang University, Republic of Korea
{gyu6491, dwkim, curseor}@cau.ac.kr

Abstract

Unsupervised feature selection aims to identify a compact subset of features that captures the intrinsic structure of data without supervised label. Most existing studies evaluate the performance of methods using the single-label dataset that can be instantiated by selecting a label from multi-label data while maintaining the original features. Because the chosen label can vary arbitrarily depending on the experimental setting, the superiority among compared methods can be changed with regard to which label happens to be selected. Thus, evaluating unsupervised feature selection methods based solely on single-label accuracy is unreasonable for assessing their true discriminative ability. This study revisits this evaluation paradigm by adopting a multi-label classification framework. Experiments on 21 multi-label datasets using several representative methods demonstrate that performance rankings differ markedly from those reported under single-label settings, suggesting the possibility of multi-label evaluation settings for fair and reliable comparison of unsupervised feature selection methods.

1 Introduction

Unsupervised Feature Selection (UFS) aims to identify a compact yet informative subset of features that effectively represents the intrinsic structure of data without any label information. Because real-world datasets often include a large number of redundant or irrelevant features, UFS serves as a crucial preprocessing step for improving interpretability and subsequent learning performance [11].

The prevailing evaluation paradigm in UFS research implicitly assumes the superiority among UFS methods can be evaluated under the single-label setting [10]. In practice, real-world data are often multi-label in nature, where feature sets may correspond to multiple valid label combinations. In this regard, a single-label dataset can be viewed as an instantiation of selecting a label from a multi-label dataset based on some intention while maintaining the original feature set, where the discarded labels are unknown. Consequently, the performance reported under single-label evaluation may not reflect the true representational capability of the selected feature subset but may rather depend on the luck regarding the arbitrarily chosen label.

To address this overlooked issue, this study revisits the evaluation framework of UFS by adopting a multi-label evaluation paradigm. We investigate how existing UFS models perform when evaluated in a multi-label classification environment. By employing representative multi-label evaluation measures such as Hamming Loss, Ranking Loss, One-Error, and Multi-Label Accuracy, we systematically analyze whether the relative performance rankings of existing UFS methods remain consistent when the evaluation shifts from single-label to multi-label settings.

*Corresponding author.

AI for Tabular Data workshop at EurIPS 2025

2 Related Work

Unsupervised FS has been widely studied as a preprocessing strategy for dimensionality reduction and data interpretation. Existing approaches are generally categorized into graph-based, information-theoretic, and evolutionary frameworks. Graph-based methods, such as Laplacian Score and Multi-Cluster Feature Selection (MCFS), preserve local manifold structures by constructing affinity graphs that capture neighborhood similarities, while methods like Unsupervised Discriminative Feature Selection (UDFS) introduce spectral regularization to enhance discriminative capacity [4, 25].

Information-theoretic approaches aim to maximize feature dependency or joint entropy to identify informative subsets, and evolutionary or memetic methods such as Robust Unsupervised Feature Selection (RUFS) and Nonnegative Discriminative Feature Selection (NDFS) employ stochastic search and iterative refinement for global optimization [16, 12]. Our previous work on Pattern Discrimination Power (PDP)-based FS demonstrated that maximizing joint entropy enhances the intrinsic discriminability of data without supervision [20].

Despite these advances, the performance of most methods has been evaluated only under single-label settings, typically by combining selected features with simple classifiers such as k-Nearest Neighbor, Naive Bayes, or Decision Tree. This evaluation scheme assumes that each instance belongs to a single class, disregarding that real-world data often exhibit multi-label associations, where one instance can correspond to multiple categories simultaneously. Consequently, single-label measures such as accuracy or NMI cannot fully reflect the structural generalization or representational robustness of the selected features.

To the best of our knowledge, no prior study has systematically examined FS performance under a multi-label evaluation framework. This study fills this gap by analyzing representative methods using multi-label measures and reveals how the evaluation paradigm itself can influence the perceived superiority of existing FS methods.

3 Methodology

This study redefines the evaluation framework of UFS to address its practical inconsistency under multi-label conditions. We analyze how representative UFS methods perform when their selected features are evaluated using a multi-label classification setting.

Traditional UFS evaluation is typically performed under a single-label classification setting. In the real world, an object x can typically be assigned to multiple different labels $L = \{l_1, \ldots, l_{|L|}$. According to this nature, a multi-label dataset that consists of multiple features F and labels L can be created. Thus, multiple single-label datasets can be derived by including the original feature set F as it is while selecting a label $l^* \in L$. Because l^* will be chosen by some intention that may be unknown to the observer, this process can be a random process. This randomness may result in biased evaluation or misleading conclusions. For example, a feature subset appearing to perform well under one label might yield incompetent performance if another label is considered.

All UFS methods are trained in a fully unsupervised manner without using label information. For each dataset, a fixed number of top-k features is selected based on each method's scoring criterion. The selected features are then evaluated through a multi-label classification model, such as the ML-kNN classifier, which predicts multiple label memberships simultaneously. Each classifier is trained and tested under conditions to ensure fair comparison across FS methods.

To quantify the performance of UFS under multi-label conditions, we adopt four representative measures widely used in multi-label learning. The definition of Hamming Loss is given as follows.

$$hloss(h) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{q} \, |h(\boldsymbol{x}_i) \Delta Y_i|. \tag{1}$$

Hamming Loss measures the proportion of misclassified labels among all possible label assignments. A value of 0 indicates that every label of each instance is correctly predicted, while 1 represents complete disagreement between prediction and ground truth. A smaller Hamming Loss implies better multi-label classification performance. The definition of One-Error is given as follows.

$$one - error(f) = \frac{1}{p} \sum_{i=1}^{p} [\![[\arg\max_{y \in \mathcal{Y}} f(\boldsymbol{x}_i, y)] \notin Y_i]\!]. \tag{2}$$

One-Error evaluates how often the top-ranked predicted label is not included in the set of true labels. A smaller One-Error value indicates better predictive performance, as it implies that the most confident prediction of the model corresponds to a relevant (true) label more frequently. The definition of Ranking Loss is given as follows.

$$rloss(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}|\{(y', y'') \mid f(\boldsymbol{x}_i, y') \tag{3}$$

$$\leq f(\boldsymbol{x}_i, y''),\ (y', y'') \in Y_i \times \bar{Y}_i)\}| \tag{4}$$

Ranking Loss measures the average fraction of label pairs that are incorrectly ordered across all instances. It quantifies how often irrelevant labels are ranked above relevant ones in the prediction output. A smaller Ranking Loss value indicates better performance, implying that the model successfully assigns higher scores to true labels than to false labels in most cases. The definition of Multi-Label Accuracy is given as follows.

$$\text{Multi-label Accuracy}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(\mathbf{x}_i) \cap Y_i|}{|h(\mathbf{x}_i) \cup Y_i|}. \tag{5}$$

Multi-Label Accuracy measures the overlap between the predicted and true label sets for each instance and then averages the result over all samples. It is equivalent to the Jaccard similarity coefficient, capturing how similar the prediction and ground truth label sets are. A value of 1 indicates perfect label matching, while 0 represents complete disagreement. Hence, a larger Multi-Label Accuracy value reflects better performance in predicting the correct combination of labels.

Lower values of Hamming Loss, Ranking Loss, and One-Error indicate better performance, whereas higher Multi-label Accuracy implies superior predictive capability of the selected feature subset. Unlike traditional single-label accuracy, these measures collectively reflect both label dependency and prediction consistency, offering a more comprehensive view of feature quality in multi-label environments.

4 Experiment

To demonstrate the validity of the proposed evaluation framework, experiments were conducted on 21 publicly available multi-label datasets from diverse domains, including text, biology, image, and signal processing. The datasets were obtained from the Multi-Label Learning Resources repository provided by the University of Córdoba. The names of the datasets are as follows: Inter3000 [9], CHD49 [21], GpositiveGO [24], GpositivePseAAC [24], PlantGO [24], PlantPseAAC [24], VirusGO [24], Waterquality [1], Birds [3], CAL500 [23], Emotions [22], Enron [18], Flags [8], Foodtruck [19], Genbase [6], Image [27], Langlog [17], Medical [15], Scene [2], Coffee [5]], and Yeast [7]. These datasets cover a wide range of label cardinalities and feature dimensions, allowing for a comprehensive comparison across different data characteristics.

To evaluate the effectiveness of the selected feature subsets, Multi-Label k-Nearest Neighbor (MLkNN) [28], where the number of neighbors was fixed to $k = 10$ Each experiment was repeated ten times under a hold-out cross-validation scheme. For each run, 80% of the instances were randomly selected for training, while the remaining 20% were used as the test set to assess classification performance.

The predicted labels of the test samples were evaluated using four standard multi-label measures. Among them, Multi-Label Accuracy serves as the primary metric for assessing classification performance, while the complementary loss-based measures, Hamming Loss, Ranking Loss, and One-Error, are provided in Appendix A for reference. A higher Multi-Label Accuracy or lower loss-based values indicate better performance, reflecting each model's ability to capture the structural label dependencies within the datasets.

The experimental results summarized in Table 1 demonstrate the comparative performance of the Entropy Maximization UFS (EMUFS) [20] and several representative unsupervised FS methods across multi-label datasets. Overall, the EMUFS achieves competitive or superior performance, ranking first in Multi-Label Accuracy followed by MCFS [4], Fast Sparse Discriminative K-means (FSDK) [14], Robust Unsupervised Feature Selection with Local Preserving (RUSLP) [13], Convex Nonnegative Matrix Factorization with Adaptive Graph Constraint (CNAFS) [26], and Novel Unsupervised Feature

Table 1: Comparison of EMUFS [20] and representative unsupervised FS methods on 21 multi-label datasets using evaluation measure Multi-Label Accuracy. The highest values for accuracy are highlighted in bold. "Avg. Rank" represents the average ranking of each method across all datasets, where a lower value indicates better overall performance.

Datasets	**EMUFS**	CNAFS	EGCFS	FSDK	MCFS	RUSLP
Inter3000	0.189±0.047	**0.191±0.030**	0.154±0.034	0.182±0.036	0.167±0.044	0.174±0.042
CHD49	0.445±0.033	0.446±0.052	0.469±0.025	**0.474±0.040**	0.441 ±0.047	0.461±0.039
GpositiveGO	**0.814±0.029**	0.350±0.063	0.297±0.046	0.810±0.047	0.808±0.030	0.336±0.062
GpositivePseAAC	**0.618±0.059**	0.500±0.071	0.433±0.048	0.621±0.024	0.599±0.045	0.540±0.065
PlantGO	**0.658±0.035**	0.141±0.019	0.101±0.008	0.584±0.053	0.643±0.041	0.121±0.015
PlantPseAAC	**0.265±0.037**	0.139±0.032	0.117±0.020	0.224±0.059	0.236±0.039	0.214±0.047
VirusGO	**0.680±0.088**	0.456±0.086	0.282±0.050	0.634±0.116	0.709±0.067	0.456±0.090
Water quality	0.401±0.014	0.397±0.017	**0.402±0.017**	0.399±0.014	0.395±0.004	0.406±0.012
Birds	0.371±0.145	0.365±0.091	0.265±0.091	0.459±0.121	**0.485±0.032**	0.288±0.082
CAL500	0.173±0.004	0.179±0.007	0.173±0.004	**0.185±0.005**	0.179±0.006	0.179±0.008
Emotions	0.508±0.034	**0.540±0.024**	0.528±0.022	0.530±0.031	0.521±0.014	0.53±0.031
Enron	**0.269±0.027**	0.175±0.023	0.203±0.029	0.258±0.032	0.182±0.013	0.087±0.003
Flags	**0.525±0.040**	0.513±0.023	0.520±0.031	0.506±0.034	0.499±0.041	0.514±0.034
Foodtruck	0.254±0.023	0.245±0.023	**0.275±0.022**	0.246±0.020	0.267±0.034	0.254±0.023
Genbase	0.318±0.039	0.199±0.059	**0.358±0.091**	0.274±0.093	0.221±0.065	0.375±0.208
Image	0.495±0.030	**0.544±0.025**	0.505±0.011	0.475±0.032	0.527±0.022	0.505±0.012
Llog	**0.042±0.005**	0.022±0.006	0.024±0.005	0.064±0.058	0.050±0.048	0.017±0.001
Medical	**0.433±0.083**	0.032±0.003	0.033±0.002	0.373±0.129	0.404±0.142	0.031±0.001
Scene	0.518±0.027	0.645±0.020	0.559±0.016	0.545±0.025	**0.668±0.022**	0.576±0.017
Coffee	**0.067±0.013**	0.027±0.007	0.026±0.003	0.059±0.013	0.065±0.015	0.018±0.003
Yeast	0.412±0.013	0.410±0.029	0.374±0.016	0.399±0.035	0.427±0.031	**0.435±0.028**
Avg.Rank	**2.76**	3.86	4.24	3.05	3.05	3.67

Selection via Adaptive Graph Learning and Constraint (EGCFS) [29] and showing the lowest average ranks in Hamming Loss, Ranking Loss, and One-Error EMUFS, MCFS, FSDK, CNAFS, RUSLP, EGCFS. These results highlight the robustness of the EMUFS across diverse domains. While FSDK and MCFS sometimes exhibit coPrevious studies have commonly reported that recently developed methods, such as FSDK, RUSLP, and CNAFS, outperform traditional graph-based methods like MCFS in single-label environments.amming Loss, Ranking Loss, and One-Error) are provided in Appendix A, whereas Table 1 focuses on Multi-Label Accuracy as the primary indicator of overall effectiveness.

Previous studies have commonly reported that recently developed methods such as FSDK, RUSLP, and CNAFS outperform traditional graph-based methods like MCFS in single-label environments. These findings were largely derived under evaluation frameworks that assume a one-to-one correspondence between features and class variables, favoring information-theoretic or sparse-representation–based approaches that optimize discriminability for individual labels. However, when reexamined under a multi-label evaluation setting, this relative superiority no longer holds consistently. In particular, MCFS, despite being a comparatively classical method, demonstrates competitive or even superior performance across multiple evaluation criteria. This observation suggests that single-label evaluations may have overestimated the generalization capability of newer methods by failing to reflect inter-label dependencies inherent in real-world data. Therefore, the conventional assumption may be an artifact of the single-label paradigm, emphasizing the need to reassess FS methods under a unified multi-label perspective. These results confirm that performance rankings reported in traditional single-label evaluations do not necessarily hold under multi-label conditions.

5 Conclusion

This study revisited the evaluation paradigm of UFS by examining existing methods under a multi-label classification framework. While previous UFS research predominantly relied on single-label evaluations, our findings on 21 diverse datasets revealed that feature subsets produce substantially different results when assessed under multi-label conditions. This discrepancy indicates that the reported superiority of UFS methods in earlier studies may be influenced by random label assignments rather than their true structural representation capability. The results emphasize the necessity of employing multi-label evaluation protocols to more accurately reflect the generalization and robustness of FS methods in real-world data scenarios.

Acknowledgments and Disclosure of Funding

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligent Graduate School Program (Chung-Ang University)].

References

[1] Hendrik Blockeel, Sašo Džeroski, and Jasna Grbović. Simultaneous prediction of multiple chemical parameters of river water quality with tilde. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 32–40. Springer, 1999.

[2] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[3] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z Fern, Raviv Raich, Sarah JK Hadley, Adam S Hadley, and Matthew G Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012.

[4] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342, 2010.

[5] Francisco Charte and David Charte. Working with multilabel datasets in r: The mldr package. 2015.

[6] Sotiris Diplaris, Grigorios Tsoumakas, Pericles A Mitkas, and Ioannis Vlahavas. Protein classification with multiple algorithms. In *Panhellenic conference on informatics*, pages 448–456. Springer, 2005.

[7] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

[8] Eduardo Corrêa Gonçalves, Alexandre Plastino, and Alex A Freitas. A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In *2013 IEEE 25th international conference on tools with artificial intelligence*, pages 469–476. IEEE, 2013.

[9] Derek Greene and Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 423–438. Springer, 2009.

[10] Saeed Karami, Farid Saberi-Movahed, Prayag Tiwari, Pekka Marttinen, and Sahar Vahdati. Unsupervised feature selection based on variance–covariance subspace distance. *Neural Networks*, 166:188–203, 2023.

[11] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 1026–1032, 2012.

[12] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 1026–1032, 2012.

[13] Chuan Luo, Jian Zheng, Tianrui Li, Hongmei Chen, Yanyong Huang, and Xi Peng. Orthogonally constrained matrix factorization for robust unsupervised feature selection with local preserving. *Information Sciences*, 586:662–675, 2022.

[14] Feiping Nie, Zhenyu Ma, Jingyu Wang, and Xuelong Li. Fast sparse discriminative k-means for unsupervised feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9943–9957, 2023.

[15] John Pestian, Chris Brew, Pawel Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, 2007.

[16] Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *IJCAI*, pages 1621–1627, 2013.

[17] Jesse Read. *Scalable multi-label classification*. PhD thesis, University of Waikato, 2010.

[18] Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Multi-label classification using ensembles of pruned sets. In *2008 eighth IEEE international conference on data mining*, pages 995–1000. IEEE, 2008.

[19] Adriano Rivolli, Larissa C Parker, and Andre CPLF de Carvalho. Food truck recommendation using multi-label classification. In *EPIA Conference on Artificial Intelligence*, pages 585–596. Springer, 2017.

[20] Wangduk Seo and Jaesung Lee. Unsupervised feature selection towards pattern discrimination power. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[21] Huan Shao, GuoZheng Li, GuoPing Liu, and YiQin Wang. Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine. *Science China Information Sciences*, 56(5):1–13, 2013.

[22] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 workshop on mining multidimensional data (MMD'08)*, volume 21, pages 53–59, 2008.

[23] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[24] Jianhua Xu, Jiali Liu, Jing Yin, and Chengyu Sun. A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. *Knowledge-Based Systems*, 98:172–184, 2016.

[25] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. 2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI international joint conference on artificial intelligence*, 2011.

[26] Aihong Yuan, Mengbo You, Dongjian He, and Xuelong Li. Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection. *IEEE Transactions on cybernetics*, 52(6):5522–5534, 2020.

[27] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[28] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[29] Rui Zhang, Yunxing Zhang, and Xuelong Li. Unsupervised feature selection via adaptive graph learning and constraint. *IEEE Transactions on neural networks and learning systems*, 33(3):1355–1362, 2020.

A Additional Evaluation Measures

Table 2: Comparison of EMUFS [20] and representative unsupervised FS methods on 21 multi-label datasets using evaluation measure Hamming Loss. The lowest values for accuracy are highlighted in bold. "Avg. Rank" represents the average ranking of each method across all datasets, where a lower value indicates better overall performance.

Datasets	**EMUFS**	CNAFS	EGCFS	FSDK	MCFS	RUSLP
Inter3000	**0.383±0.037**	0.414±0.062	0.431±0.042	0.388±0.041	0.411±0.053	0.448±0.046
CHD49	0.397±0.055	0.404±0.068	0.367±0.021	**0.359±0.050**	0.404±0.071	0.368±0.040
GpositiveGO	0.095±0.015	0.485±0.120	0.586±0.114	**0.092±0.022**	0.099±0.021	0.492±0.152
GpositivePseAAC	**0.189±0.031**	0.261±0.042	0.304±0.030	0.190±0.016	0.207±0.028	0.237±0.039
PlantGO	**0.059±0.007**	0.443±0.090	0.762±0.085	0.077±0.012	0.067±0.007	0.569±0.089
PlantPseAAC	**0.161±0.023**	0.233±0.036	0.287±0.037	0.189±0.048	0.171±0.034	0.183±0.035
VirusGO	0.121±0.038	0.228±0.059	0.456±0.085	0.138±0.039	**0.111±0.025**	0.225±0.050
Water quality	0.336±0.010	0.336±0.007	0.336±0.013	0.336±0.010	0.336±0.007	**0.333±0.010**
Birds	0.104±0.021	0.104±0.008	0.112±0.017	0.092±0.016	**0.085±0.009**	0.109±0.013
CAL500	0.336±0.011	0.324±0.012	0.333±0.007	0.316±0.014	**0.323±0.012**	0.325±0.017
Emotions	0.242±0.021	**0.231±0.014**	0.241±0.012	0.238±0.016	0.244±0.016	0.235±0.018
Enron	**0.109±0.006**	0.179±0.027	0.133±0.018	0.112±0.010	0.160±0.020	0.594±0.037
Flags	**0.340±0.036**	0.345±0.022	0.346±0.021	0.348±0.032	0.354±0.036	0.348±0.025
Foodtruck	0.331±0.038	0.333±0.022	**0.301±0.029**	0.329±0.023	0.310±0.044	0.340±0.022
Genbase	0.093±0.018	0.189±0.060	**0.083±0.021**	0.127±0.039	0.169±0.059	0.093±0.042
Image	0.232±0.015	**0.212±0.012**	0.243±0.008	0.247±0.018	0.223±0.015	0.240±0.007
Llog	**0.089±0.010**	0.170±0.048	0.121±0.024	0.098±0.025	0.091±0.010	0.851±0.023
Medical	**0.036±0.010**	0.743±0.078	0.682±0.033	0.047±0.021	0.045±0.019	0.789±0.046
Scene	0.177±0.011	0.120±0.007	0.152±0.007	0.159±0.008	**0.112±0.008**	0.147±0.007
Coffee	**0.083±0.007**	0.173±0.035	0.181±0.034	0.087±0.012	0.086±0.007	0.341±0.054
Yeast	0.306±0.010	0.312±0.033	0.350±0.023	0.318±0.033	0.294±0.034	**0.288±0.025**
Avg.Rank	**2.43**	3.81	4.29	2.95	2.76	4.10

Appendix A presents the supplementary experimental results evaluated using loss-based multi-label measures, including Hamming Loss, Ranking Loss, and One-Error. These measures provide complementary perspectives to the main analysis, quantifying classification consistency and label-ranking stability. While the primary paper focuses on Multi-Label Accuracy as the most interpretable and representative indicator, the trends across these additional metrics are largely consistent with the overall findings, further validating the robustness of the proposed evaluation.

Table 3: Comparison of EMUFS [20] and representative unsupervised FS methods on 21 multi-label datasets using evaluation measure Ranking Loss. The lowest values for accuracy are highlighted in bold. "Avg. Rank" represents the average ranking of each method across all datasets, where a lower value indicates better overall performance.

Datasets	**EMUFS**	CNAFS	EGCFS	FSDK	MCFS	RUSLP
Inter3000	**0.418±0.045**	0.438±0.051	0.432±0.048	0.437±0.048	0.455±0.058	0.455±0.055
CHD49	0.240±0.015	0.228±0.027	**0.222±0.018**	0.231±0.021	0.235±0.015	0.232±0.023
GpositiveGO	0.075±0.018	0.270±0.062	0.279±0.036	0.074±0.020	**0.069±0.010**	0.237±0.058
GpositivePseAAC	0.145±0.023	0.222±0.052	0.239±0.023	**0.145±0.018**	0.158±0.028	0.204±0.053
PlantGO	**0.049±0.008**	0.253±0.018	0.271±0.017	0.075±0.017	0.070±0.012	0.259±0.020
PlantPseAAC	**0.188±0.007**	0.234±0.025	0.244±0.018	0.192±0.017	0.196±0.022	0.217±0.019
VirusGO	**0.076±0.032**	0.142±0.033	0.229±0.047	0.094±0.041	0.091±0.017	0.132±0.025
Waterquality	0.260±0.012	0.263±0.010	0.259±0.009	0.261±0.012	0.262±0.010	**0.253±0.006**
Birds	0.205±0.035	0.203±0.025	**0.191±0.017**	0.208±0.027	0.197±0.027	0.193±0.019
Cal500	0.188±0.004	0.185±0.003	0.185±0.005	0.184±0.007	**0.182±0.003**	0.185±0.005
Emotions	**0.166±0.017**	0.173±0.016	0.177±0.022	0.174±0.020	0.188±0.015	0.173±0.016
Enron	**0.095±0.003**	0.109±0.006	0.104±0.004	0.098±0.004	0.107±0.007	0.111±0.005
Flags	0.232±0.029	**0.230±0.022**	0.236±0.017	0.247±0.020	0.251±0.022	0.233±0.033
Foodtruck	0.170±0.022	0.169±0.035	0.170±0.020	0.168±0.013	**0.161±0.019**	0.173±0.026
Genbase	0.009±0.002	0.045±0.019	0.009±0.006	0.013±0.007	0.012±0.007	**0.006±0.005**
Image	0.213±0.021	**0.189±0.014**	0.217±0.010	0.231±0.018	0.197±0.011	0.228±0.012
Langlog	0.179±0.010	0.179±0.010	0.187±0.017	**0.178±0.011**	0.183±0.011	0.187±0.012
Medical	0.052±0.009	0.133±0.005	0.119±0.007	0.054±0.008	**0.042±0.004**	0.135±0.008
Scene	0.186±0.013	0.102±0.009	0.140±0.008	0.145±0.014	**0.094±0.006**	0.129±0.010
Stackexcoffee	0.295±0.034	0.273±0.045	0.271±0.045	0.257±0.027	**0.252±0.031**	0.305±0.047
Yeast	0.190±0.005	0.179±0.005	0.193±0.006	0.182±0.013	0.174±0.009	**0.171±0.005**
Avg.Rank	**2.90**	3.67	3.95	3.24	2.95	3.81

Table 4: Comparison of EMUFS [20] and representative unsupervised FS methods on 21 multi-label datasets using evaluation measure One-Error. The lowest values for accuracy are highlighted in bold. "Avg. Rank" represents the average ranking of each method across all datasets, where a lower value indicates better overall performance.

Datasets	**EMUFS**	CNAFS	EGCFS	FSDK	MCFS	RUSLP
Inter3000	**0.742±0.066**	0.758±0.052	0.767±0.074	0.752±0.047	0.788±0.062	0.797±0.064
CHD49	0.273±0.035	0.261±0.051	0.269±0.042	0.262±0.029	0.260±0.043	**0.258±0.037**
GpositiveGO	0.171±0.033	0.549±0.111	0.528±0.075	0.164±0.037	**0.134±0.018**	0.477±0.100
GpositivePseAAC	**0.287±0.039**	0.430±0.088	0.465±0.038	0.291±0.035	0.297±0.051	0.397±0.083
PlantGO	**0.250±0.020**	0.821±0.026	0.896±0.015	0.318±0.042	0.301±0.034	0.865±0.083
PlantPseAAC	**0.629±0.037**	0.707±0.048	0.711±0.030	0.631±0.036	0.635±0.022	0.673±0.026
VirusGO	**0.185±0.069**	0.420±0.073	0.561±0.102	0.244±0.098	0.222±0.035	0.407±0.100
Waterquality	0.312±0.030	0.320±0.032	0.298±0.025	0.307±0.036	0.312±0.025	**0.290±0.034**
Birds	0.571±0.057	0.568±0.064	0.578±0.034	0.514±0.036	**0.488±0.042**	0.534±0.063
Cal500	0.190±0.037	0.195±0.038	0.186±0.042	0.190±0.041	**0.179±0.052**	0.194±0.064
Emotions	0.304±0.029	**0.284±0.033**	0.298±0.048	0.291±0.048	0.312±0.036	0.308±0.041
Enron	**0.300±0.025**	0.423±0.042	0.329±0.027	0.337±0.032	0.378±0.048	0.442±0.046
Flags	**0.224±0.076**	0.239±0.079	0.268±0.074	0.271±0.087	0.263±0.025	0.239±0.050
Foodtruck	0.288±0.047	0.283±0.059	0.290±0.062	0.293±0.049	**0.281±0.045**	0.294±0.047
Genbase	0.021±0.013	0.355±0.131	**0.014±0.015**	0.022±0.008	0.024±0.018	0.021±0.022
Image	0.398±0.037	**0.352±0.030**	0.394±0.015	0.424±0.031	0.362±0.014	0.411±0.014
Langlog	**0.837±0.022**	0.866±0.026	0.885±0.026	0.858±0.027	0.860±0.022	0.876±0.038
Medical	0.336±0.022	0.717±0.029	0.701±0.020	0.320±0.027	**0.242±0.019**	0.718±0.012
Scene	0.432±0.029	0.308±0.021	0.386±0.020	0.402±0.034	**0.280±0.020**	0.361±0.021
Stackexcoffee	**0.698±0.049**	0.793±0.057	0.782±0.043	0.760±0.049	0.704±0.080	0.816±0.049
Yeast	**0.259±0.015**	0.263±0.020	0.277±0.021	0.263±0.018	0.272±0.013	0.264±0.024
Avg.Rank	**2.52**	3.90	4.29	3.14	2.76	4.14