MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 4, pp. 1381–1392 (2025)

Information Theory in Multi-Label Feature Selection: An Analytical Review

Meskaoui M., Chamlal H., Ouaderhman T.

Computer Science and Systems Laboratory (LIS), Faculty of Sciences Ain Chock, Hassan II University of Casablanca, Morocco

(Received 3 August 2025; Revised 14 December 2025; Accepted 15 December 2025)

In the context of multi-label learning, feature selection (MLFS) is a key process for handling high-dimensional datasets, aiming to retain the most informative features while preserving inter-label relationships. This study presents an extensive overview of state-of-the-art MLFS approaches founded on principles from information theory. The paper first introduces the fundamental concepts of information theory, then provides a detailed review of representative MLFS methods along with their theoretical background. Performance assessments are carried out on real-world multi-label datasets, allowing us to highlight the advantages and shortcomings of each method. For comparison, we employ widely used evaluation metrics such as Hamming Loss, Accuracy, Label Ranking Loss, and F1-score. We also outline future research perspectives, including the design of a new feature relevance criterion that integrates label importance weighting and redundancy reduction based on label dependencies, with the aim of enhancing both feature selection and multi-label classification accuracy.

Keywords: *algorithm adaptation; multi-label feature selection; machine learning; information theory; interaction information; problem transformation.*

2010 MSC: 94A15, 68T05, 62H30, 68Q32, 62F07 **DOI:** 10.23939/mmc2025.04.1381

1. Introduction

In the field of machine learning, two primary learning paradigms are commonly distinguished: *supervised learning* [1, 2], in which models are built using datasets containing input–output pairs, and *unsupervised learning*, where the aim is to identify patterns or structures from data without predefined labels [3]. Supervised learning is a core component of statistical learning and artificial intelligence, with the central goal of constructing predictive models based on annotated training examples. Traditionally, it operates under the assumption that each data instance corresponds to a single label. However, in many real-world applications, outputs can be complex, with individual instances linked to multiple, potentially dependent, labels. This situation has given rise to *multi-label learning*, an extension of the supervised framework designed to manage cases where an observation may belong to several categories simultaneously [4].

The motivation for multi-label learning stems from the need to accurately model problems characterized by overlapping and interdependent concepts. For example, in document categorization, a single article can be simultaneously classified as “science”, “health”, and “AI”. In bioinformatics, a gene may be involved in several biological processes, and in multimedia annotation or recommendation systems, items are typically associated with multiple relevant tags or attributes [4, 5]. This framework has therefore become central in diverse domains including natural language processing, image annotation, medical diagnosis, and information retrieval.

A major challenge in multi-label learning is the high dimensionality and complexity of the data, which can adversely affect both model interpretability and computational efficiency. Feature selection [6], a fundamental step in statistical learning, aims to identify and retain only the most informative features while removing irrelevant or redundant ones [4, 7]. This process reduces the risk of overfitting, enhances model interpretability, and decreases computational cost.

Feature selection approaches are generally grouped into three categories [4]:

— **Filter methods** [8, 9] evaluate feature relevance based on intrinsic data properties, such as correlation or mutual information [7], and are independent of the learning algorithm.
— **Wrapper methods** assess feature subsets by training and validating a learning model, often resulting in better performance but at higher computational cost.
— **Embedded methods** integrate feature selection directly into the model training process, exemplified by decision trees or regularized linear models.

While these strategies have shown success in single-label settings, their direct application to multi-label problems is limited by the presence of label dependencies and the increased complexity of output spaces.

To overcome these challenges, research on multi-label feature selection (MLFS) has generally progressed in two main directions [4]. The first, known as *problem transformation*, converts a multi-label problem into one or more single-label problems, for instance through approaches such as binary relevance [5] or label powerset [10,11]. The second, *algorithm adaptation*, modifies existing feature selection techniques so they can directly operate on multi-label datasets and account for label dependencies. In addition, a more recent and increasingly popular research avenue is the application of *information-theoretic approaches*, which offer a rigorous mathematical framework to quantify relationships between features, individual labels, and entire label sets [3, 12–14].

Within this framework, key concepts from information theory — including entropy, mutual information, conditional mutual information, and interaction information — are used to evaluate feature relevance, redundancy, and higher-order label dependencies [3, 12]. These measures have shown particular effectiveness in multi-label learning, where preserving label correlations and maximizing predictive information are essential for achieving both accurate feature selection and strong classification results [12–14].

The rest of the paper is organized as follows. Section 2 introduces the notations and core principles of information theory. Section 3 categorizes multi-label feature selection techniques, covering problem transformation, algorithm adaptation, and information-theoretic methods. Section 4 outlines the evaluation metrics used in multi-label classification. Section 5 presents a comparative study of advanced MLFS approaches on benchmark datasets. Finally, Section 6 summarizes the contributions and highlights possible avenues for future work.

2. Notations and fundamental concepts

2.1. Notations of information theory

Information theory [3] serves as a fundamental framework for studying dependencies and interactions between several variables. Within the scope of multi-label feature selection, it offers powerful mechanisms to assess both how relevant and how redundant certain features are. Consider three discrete random variables: $X = \{x_1, x_2, \ldots, x_p\}$, $Y = \{y_1, y_2, \ldots, y_q\}$, and $Z = \{z_1, z_2, \ldots, z_l\}$, each representing a finite set of possible values.

Entropy. In information theory, *entropy* measures the degree of uncertainty or unpredictability associated with a random variable. Formally, the entropy of X is expressed as:

$$H(X) = -\sum_{i=1}^{p} p(x_i) \log p(x_i),$$

where $p(x_i)$ refers to the probability mass assigned to the event $X = x_i$.

Joint entropy extends the concept of entropy to the case of two or more random variables, and can be expressed as:

$$H(X, Y) = -\sum_{i=1}^{p} \sum_{j=1}^{q} P(x_i, y_j) \log P(x_i, y_j),$$

where $P(x_i, y_j)$ denotes the probability mass function of the pair (X, Y).

Conditional entropy describes the remaining uncertainty of one random variable when the value of another is known. It indicates the average unpredictability of X once Y is observed. The conditional entropy $H(X|Y)$ is given by:

$$H(X|Y) = -\sum_{i=1}^{p}\sum_{j=1}^{q} P(x_i, y_j)\log\frac{P(x_i, y_j)}{P(y_j)}.$$

Mutual information measures the dependency between two random variables by quantifying the reduction in uncertainty of one variable given knowledge of the other. It is defined as:

$$I(X;Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P(x_i, y_j)\log\frac{P(x_i, y_j)}{P(x_i)P(y_j)}.$$

Conditional mutual information extends mutual information by measuring the dependency between two variables given the values of a third variable. It is given by:

$$I(X;Y|Z) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{l} P(x_i, y_j, z_k)\log\frac{P(x_i, y_j \mid z_k)}{P(x_i \mid z_k)\,P(y_j \mid z_k)}.$$

Interaction information generalizes mutual information to capture dependencies among multiple variables. The three-way interaction information is expressed as:

$$\begin{aligned} I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \\ &= I(Y;Z) - I(Y;Z|X) \\ &= I(X;Z) - I(X;Z|Y). \end{aligned}$$

3. Multi-label feature selection

Multi-label feature selection has been the subject of significant research efforts aimed at pinpointing the most informative attributes in complex multi-label datasets. Generally, these approaches are divided into two principal strategies [4]: problem transformation and algorithm adaptation.

3.1. Problem transformation approach

Problem transformation methods reformulate a multi-label classification task by breaking it down into one or more equivalent single-label problems. This reformulation makes it possible to apply conventional single-label feature selection techniques. After transformation, the selection process is performed separately on each generated single-label problem, and the outcomes are combined to produce the final set of features relevant to the original multi-label classification.

The main techniques frequently employed in this approach include:

— **Label Powerset (LP) [10]:** Assigns a unique class to each possible combination of labels, effectively transforming a multi-label problem into a multi-class problem.

— **Pruned Problem Transformation (PPT) [11]:** An extension of LP that removes infrequent label combinations based on a predefined threshold, improving efficiency and reducing complexity.

— **Binary Relevance (BR) [5]:** Decomposes the multi-label task into multiple binary classification problems, one for each label, and applies single-label feature selection separately to each subtask. The selected features are then merged to form a final subset.

While problem transformation methods are widely used, they come with significant challenges, such as risking the loss of inter-label dependencies and encountering problems due to unbalanced label frequencies.

3.2. Algorithm adaptation approach

To address these drawbacks, algorithm adaptation approaches adjust conventional single-label feature selection methods so they can better manage the complexities involved in multi-label learning. Such approaches maintain dependencies between labels and enhance the overall process of selecting relevant features.

Beyond the algorithm adaptation approaches mentioned earlier, methods based on information theory have emerged as some of the most prevalent techniques for feature selection in multi-label classification. These methods apply principles from information theory to pinpoint important features, ensuring label dependencies are preserved while also accounting for redundancy. The next section outlines several notable techniques grounded in information theory.

3.3. Information theory-based methods

Given a set of features $F = \{f_1, f_2, \ldots, f_d\}$ and a set of labels $L = \{l_1, l_2, \ldots, l_q\}$, let $S \subset F$ denote the subset of features already selected, and consider $f_i \in F \setminus S$ as a candidate for selection.

Approaches grounded in information theory typically evaluate each candidate through a scoring criterion composed of two parts:

1. the *relevance* of the candidate feature with respect to a given label l,
2. the *redundancy* of the candidate with respect to the features that have already been chosen.

This trade-off between maximizing relevance and minimizing redundancy forms the basis of the *Maximum Relevance Minimum Redundancy (mRMR)* framework [7]. The score assigned to f_i can be expressed as:

$$\mathrm{mRMR}(f_i) = I(f_i; l) - \frac{1}{|S|} \sum_{f_j \in S} I(f_i; f_j),$$

where $I(f_i; l)$ quantifies the mutual information between the candidate f_i and label l, and the second term measures its average redundancy with the selected subset.

In practice, the feature subset is constructed by *maximizing* this score, aiming for features that are both highly relevant and minimally redundant.

The mRMR formulation aligns with the maximum dependency principle for first-order selection. Its stepwise selection process circumvents the complexity of full multivariate density estimation. Furthermore, mRMR can be integrated with other selection paradigms, such as wrapper-based methods, to produce compact yet informative subsets while keeping the computational burden low.

In their work, Lee and Kim [15] introduced the *Pairwise Multi-label Utility* (PMU) approach, a technique for feature selection in multi-label settings. The core idea of PMU is to enhance the dependency between features and labels by incorporating *three-way interaction* analysis. This method estimates the usefulness of a feature by jointly considering dependencies among features, between features and labels, and among labels themselves. The corresponding score function for PMU is formulated as:

$$\mathrm{PMU}(f_i) = \sum_{l_j \in L} I(f_i; l_j) - \sum_{f_s \in S} \sum_{l_j \in L} I(f_i; f_s; l_j) - \sum_{l_j \in L} \sum_{l_k \in L \setminus \{l_j\}} I(f_i; l_j; l_k).$$

This formulation ensures that selected features exhibit strong individual relevance while minimizing redundancy with other features and considering pairwise label interactions.

Lee and Kim [12] introduced a new score function, designed to evaluate the relevance of candidate features in multi-label feature selection. Their approach leverages mutual information and interaction information to capture dependencies between features and labels, with a particular focus on relationships between label pairs. The D2F evaluation criterion is defined as:

$$\mathrm{D2F}(f_i) = \sum_{l_j \in L} I(f_i; l_j) - \sum_{l_j \in L} \sum_{l_k \in L} I(f_i; l_j; l_k).$$

This formulation ensures that both individual feature-label relevance and interactions between labels are considered in the feature selection process.

Lee and Kim [16] proposed the *Scalable Criterion for Large Label Set* (SCLS), an advanced approach for multi-label feature selection. This method is designed to efficiently approximate dependency measures, enabling more effective evaluation of feature relevance. By improving computational scalability, SCLS becomes particularly suitable for scenarios involving datasets with a large number of labels. The formal definition of the SCLS evaluation criterion is given as:

$$\mathrm{SCLS}(f_i) = \sum_{l_j \in L} I(f_i; l_j) - \sum_{f_s \in S} \frac{I(f_i; f_s)}{H(f_i)} \sum_{l_j \in L} I(f_i; l_j).$$

By incorporating an efficient redundancy correction term, SCLS balances feature relevance and redundancy, ensuring the selection of informative yet non-redundant features in large-scale multi-label datasets.

Building on the concept of PMU, Zhang et al. [13] addressed the issue of label redundancy in multi-label feature selection. Instead of incorporating label redundancy directly into the redundancy term, they utilized conditional mutual information within the relevance term to estimate the mutual information between candidate features and labels. This method, referred to as *Label Redundancy-aware Feature Selection* (LRFS), enhances feature evaluation by more effectively exploiting label dependencies. The LRFS evaluation function is expressed as:

$$\mathrm{LRFS}(f_i) = \sum_{l_k \in L} \left\{ \sum_{l_j \neq l_k} I(f_i; l_j \mid l_k) - \frac{1}{|S|} \sum_{f_j \in S} I(f_i; f_j) \right\}.$$

Zhang, Liu et al. [14] introduced a multi-label feature selection method that incorporates label complementarity, known as LSMFS. This method first evaluates the relationships between label pairs to determine whether they are independent, redundant, or interdependent. To quantify label complementarity, LSMFS utilizes the term $\min(0, I(f_i; l_j; l_k))$ which measures the additional information gained from label l_k when computing the mutual information (MI) between a candidate feature f_i and a label l_j.

The feature relevance term in LSMFS consists of two components:

— The mutual information (MI) between the feature and the label set.
— The complementary information derived from label relationships.

The feature redundancy term accounts for the mutual information between the candidate feature and the selected features. The LSMFS evaluation function is given by:

$$\mathrm{LSMFS}(f_i) = \sum_{l_j \in L} \left[I(f_i; l_j) + \sum_{l_k \in L} \big| \min(0, I(f_i; l_j; l_k)) \big| \right] - \sum_{f_j \in S} I(f_i; f_j).$$

By incorporating label complementarity, LSMFS enhances feature selection by better capturing the informative relationships between labels, leading to improved multi-label classification performance.

LIWR-LDR [17]: Conventional approaches often neglect the varying significance of individual labels within the complete label set and ignore the influence of selected features on the labels. To address these issues, the LIWR-LDR method incorporates the concept of *Label Importance Weight* (LIW) to measure the importance of each label. This importance score is used to define a label importance-weighted relevance (LIWR) term for assessing feature relevance, while redundancy is modeled through a label-dependency redundancy (LDR) term computed using the uncertainty coefficient. The objective is to maximize LIWR while minimizing LDR. The scoring function is expressed as:

$$J_{\text{LIWR-LDR}}(f_i) = \sum_{l_j \in L} I(f_i; l_j) \sum_{l_k \in L} I(l_j; l_k) - \sum_{f_s \in S} \sum_{l_j \in L} \frac{H(f_s) - H(f_s | l_j)}{H(f_s)} I(f_i; f_s).$$

In this formulation, the first term captures the relevance of feature f_i to each label, adjusted according to overall label dependencies. The second term penalizes redundancy between f_i and the already selected features, weighted by the extent to which features depend on the labels.

LSRIFS [18]: Conventional information-theoretic approaches to feature selection typically employ greedy algorithms, which can neglect how relevant information is distributed and underestimate the influence of redundancy. LSRIFS addresses these challenges by introducing a label-specific relevance weight that evaluates feature relevance both from macro and micro perspectives. This weight amplifies the importance of features strongly correlated with individual labels. Redundancy is handled similarly to mRMR. The scoring function is given by:

$$J_{\text{LSRIFS}}(f_i) = \frac{1}{|L|} \exp\left(\frac{\sum_{l_j \in L} I(f_i;l_j)^2}{\sum_{f_k \in F}\sum_{l_j \in L} I(f_k;l_j)^2}\right) \sum_{l_j \in L} I(f_i;l_j) - \frac{1}{|S|}\sum_{f_s \in S} I(f_i;f_s).$$

The exponential term emphasizes features with high and varying relevance across labels, while the redundancy term ensures selection of less redundant features.

FLIS [19]: In multi-label feature selection, effectively modeling the intricate and evolving dependencies between features and labels can significantly enhance performance. However, conventional techniques often overlook the way in which selected features dynamically influence label interrelations. The FLIS approach (Feature and Label Information Supplementation) addresses this by incorporating both conditional mutual information and mutual information to capture relationships involving selected features, candidate features, and labels. Furthermore, FLIS identifies specific label relationships that contribute additional label-related information. By dynamically assessing the influence of different features on label dependencies, this method aims to improve classification accuracy. Its scoring function is:

$$J_{\text{FLIS}}(f_k, L) = \sum_{l_i \in L}\left\{ I(f_k;l_i) + \sum_{l_j \in L\setminus\{l_i\}} \max\left[0, I(f_k;l_i;l_j) + I(f_k;l_i;f_j)\right]\right\} - \sum_{f_j \in S} I(f_k;f_j).$$

The term includes classic feature-label mutual information and higher-order interactions involving label pairs and previously selected features, considering only positive contributions to relevance.

These MLFS methods improve multi-label classification by balancing feature relevance, redundancy reduction, and label dependencies, leading to better feature selection and improved model performance.

A large number of advanced multi-label feature selection techniques grounded in information theory employ a similar greedy selection strategy. The general procedure, outlined in Algorithm 1, iteratively chooses the feature that optimizes a scoring criterion, balancing its relevance to the target labels against its redundancy with the features that have already been selected.

Algorithm 1 Generic information-theoretic feature selection.

Require: Feature set F, Label set L, Desired number of features K
Ensure: Selected feature subset S with $|S| = K$

$S \leftarrow \varnothing$
for $t = 1$ to K
  for all $f \in F \setminus S$
    Compute **Relevance**(f, L)
    Compute **Redundancy**(f, S)
    Compute Score(f) = Relevance(f, L) − Redundancy(f, S)
  Select $f^* = \arg\max_f \text{Score}(f)$
  $S \leftarrow S \cup \{f^*\}$
return S

4. Multi-label evaluation metrics

In multi-label classification, several evaluation metrics are employed to measure how well a model performs. These metrics are typically grouped into two main categories: label-based and example-based [4, 20],

- **Label-based metrics** measure performance at the individual label level, assigning a separate score to each label and averaging the results over all labels.
- **Example-based metrics** evaluate performance at the instance level by comparing predicted and actual label sets for each instance, followed by averaging these differences across all instances.

Let $D = \{(x_i, L_i) \mid i = 1, 2, \ldots, n\}$ be a multi-label test set, where n is the total number of instances, x_i denotes the d-dimensional feature vector of the i-th instance, and $L_i \subseteq L$ is its ground-truth label set. The set $\hat{L}_i \subseteq L$ represents the labels predicted for the i-th instance.

Hamming loss measures the proportion of misclassified labels across all instances:

$$\text{Hamming Loss}(D) = \frac{1}{n}\sum_{i=1}^{n}\frac{|L_i \triangle \hat{L}_i|}{q},$$

where q is the total number of labels, and $\triangle$ denotes the symmetric difference.

Macro-F1 is calculated as the arithmetic mean of the F1 scores across all labels:

$$\text{Macro-F1} = \frac{1}{q}\sum_{i=1}^{q}\frac{2\text{TP}_i}{2\text{TP}_i + \text{FP}_i + \text{FN}_i},$$

where TP_i, FP_i, and FN_i represent the number of true positives, false positives, and false negatives for the i-th label, respectively.

Micro-F1. Similarly, Micro-F1 is a weighted average of the F1 scores across all labels:

$$\text{Micro-F1} = \frac{\sum_{i=1}^{q} 2\text{TP}_i}{\sum_{i=1}^{q}(2\text{TP}_i + \text{FP}_i + \text{FN}_i)}.$$

For these evaluation metrics, a lower Hamming Loss (HL) indicates better classification performance. Conversely, higher Macro-F1 and Micro-F1 values signify improved classification effectiveness.

Table 1. Summary of common evaluation metrics for multi-label classification.

Metric	**Description**	**Direction**
Hamming Loss	Proportion of misclassified labels among all instances and labels. Lower is better.	↓
Macro-F1	Arithmetic mean of F1 scores across all labels. Higher is better.	↑
Micro-F1	Weighted mean of F1 scores across all labels. Higher is better.	↑
Label Ranking Loss	Measures ranking quality of relevant vs. irrelevant labels. Lower is better.	↓
Coverage Error	How far one must go in the ranking to cover all true labels. Lower is better.	↓
Average Precision Score	Assesses precision at different recall levels. Higher is better.	↑
Accuracy Score	Proportion of exactly matched label sets. Higher is better.	↑
Jaccard Score	Similarity between predicted and true label sets. Higher is better.	↑

Additionally, in multi-label classification, various evaluation metrics are used to assess a model's performance from different perspectives. Among them, the following metrics provide valuable insights into prediction quality:

— **Label Ranking Loss:** This metric evaluates how well a model ranks the relevant labels higher than the irrelevant ones. A lower score indicates that the model is better at ranking the correct labels ahead of the incorrect ones for each instance.

— **Coverage Error:** This metric measures how far, on average, the model needs to go down the ranked list of labels to cover all the true labels of an instance. A lower coverage error means that relevant labels appear earlier in the ranking, which is desirable.

— **Average Precision Score:** This metric assesses how well the model predicts relevant labels by considering the precision at different recall levels. It rewards models that rank relevant labels higher and distribute predictions across different recall levels effectively.

— **Accuracy Score:** In multi-label classification, accuracy is often computed based on exact matches between predicted and true label sets for each instance. A higher accuracy indicates that the model is correctly predicting the entire label set for more instances.

— **Jaccard Score:** Also known as the Jaccard Index, this metric measures the similarity between the predicted and true label sets. It calculates the proportion of common labels in both sets relative to the total number of unique labels in either set. Higher values indicate better performance.

Each of these metrics provides a different perspective on the model's performance, making them valuable in evaluating multi-label classification algorithms.

5. Experimental comparison and analysis

This section presents a comparative analysis of advanced information-theoretic feature selection methods on two benchmark multi-label datasets: *Emotions* (music) and *Yeast* (biology). Nine state-of-the-art methods were evaluated and the results are reported across seven evaluation metrics.

All experiments were conducted in Python using `scikit-learn` and `scikit-multilearn`. Computations were performed on an Intel Core i5-2540M CPU (2.60 GHz, 8 GB RAM). The ML-KNN classifier ($k = 10$) was applied after selecting the top 50 features per method. We report the following evaluation metrics: Hamming Loss (HL), Label Ranking Loss (LRL), Coverage Error (CE), Average Precision Score (APS), F1-score, Accuracy (Acc), and Jaccard Score (Jacc).

5.1. Dataset description and experimental setup

Table 2 summarizes the main characteristics of the two datasets selected from the Mulan library. Both datasets are widely used in multi-label learning research and present distinct application challenges.

Table 2. Description of the datasets.

Name	Domain	Instances	Features	Labels	Cardinality	Density	Distinct
Emotions	Music	593	72	6	1.869	0.311	27
Yeast	Biology	2417	103	14	4.237	0.303	198

5.2. Evaluation results

Tables 3 and 4 summarize the performance of each method on the Emotions and Yeast datasets (top 50 features).

Table 3. Performance of feature selection methods on the emotions dataset (top 50 features).

Method	HL	LRL	CE	APS	F1	Acc	Jacc
SCLS	0.232	0.529	4.77	0.527	0.592	0.218	0.427
LSMFS	0.260	0.553	4.90	0.488	0.548	0.188	0.385
LRFS	0.229	0.501	4.68	0.527	0.611	0.243	0.444
IGMF	0.242	0.514	4.66	0.497	0.589	0.218	0.424
PMU	0.248	0.487	4.50	0.497	0.607	0.252	0.440
D2F	0.250	0.536	4.85	0.508	0.570	0.198	0.404
MDMR	0.250	0.536	4.85	0.508	0.570	0.198	0.404
MLSMFS	0.251	0.541	4.87	0.494	0.573	0.168	0.409
PPT_MI	0.225	0.483	4.56	0.542	0.617	0.272	0.451

Table 4. Performance of feature selection methods on the yeast dataset (top 50 features)

Method	HL	LRL	CE	APS	F1	Acc	Jacc
SCLS	0.209	0.460	11.39	0.368	0.582	0.169	0.447
LSMFS	0.220	0.486	11.57	0.356	0.552	0.159	0.419
LRFS	0.217	0.471	11.45	0.360	0.571	0.160	0.436
IGMF	0.232	0.536	12.25	0.333	0.500	0.121	0.381
PMU	0.217	0.462	11.39	0.359	0.573	0.162	0.440
D2F	0.213	0.465	11.33	0.369	0.584	0.170	0.446
MDMR	0.213	0.465	11.33	0.369	0.584	0.170	0.446
MLSMFS	0.217	0.504	12.07	0.358	0.547	0.140	0.414
PPT_MI	0.213	0.473	11.45	0.361	0.566	0.169	0.434

Analysis. The results show that all methods achieve comparable performance for most metrics, but some differences emerge. PPT_MI achieves the best overall scores, with the lowest Hamming Loss (0.225), lowest Label Ranking Loss (0.483), and highest F1-score (0.617) and Jaccard score (0.451), indicating a strong ability to select discriminative features for the *emotions* dataset. LRFS and PMU also demonstrate strong performance, particularly in terms of $F1$ and Accuracy scores. In contrast, LSMFS and MLSMFS exhibit slightly lower effectiveness on this dataset, especially for accuracy and Jaccard score.

Figure 1 provides a detailed analysis, showing the evolution of each metric as the number of selected features increases on the Emotions dataset. Generally, performance improves with more features, and the most pronounced differences between methods are observed in the ranking-based and precision-based metrics.

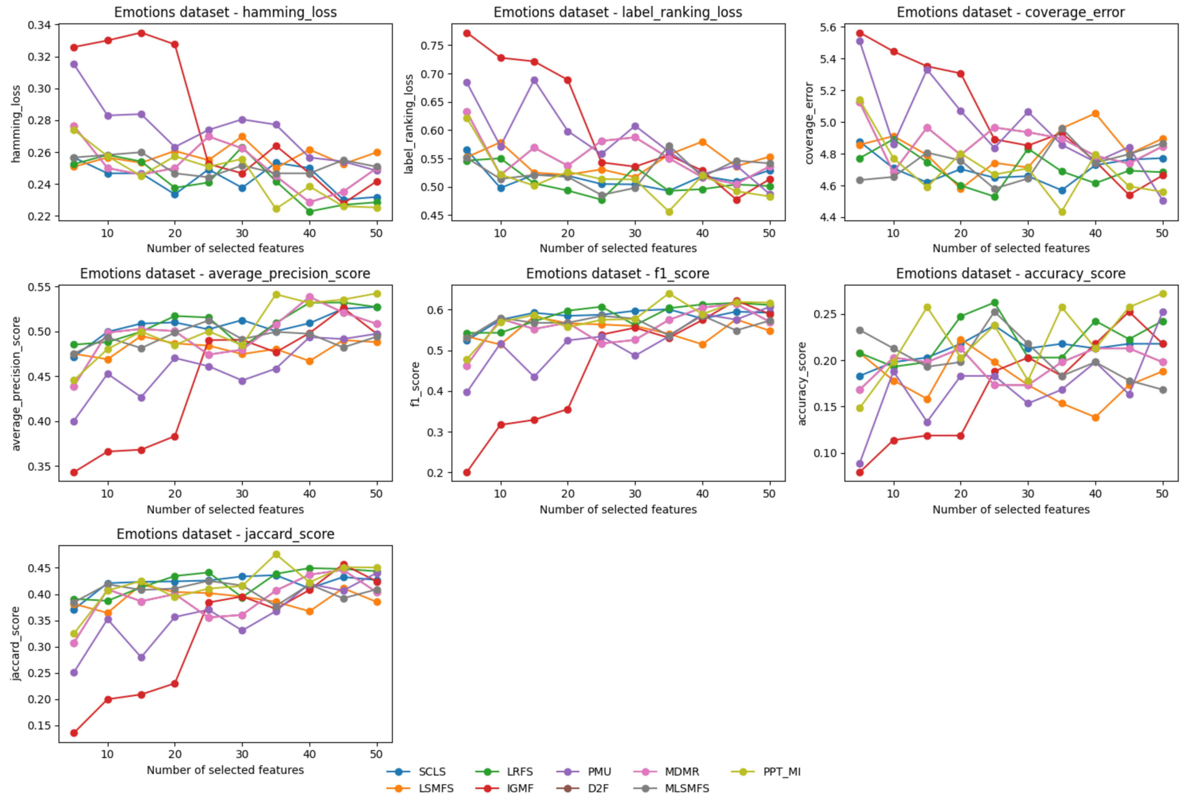

Fig. 1. Evolution of performance metrics with the number of selected features for each method on the *Emotions* dataset.

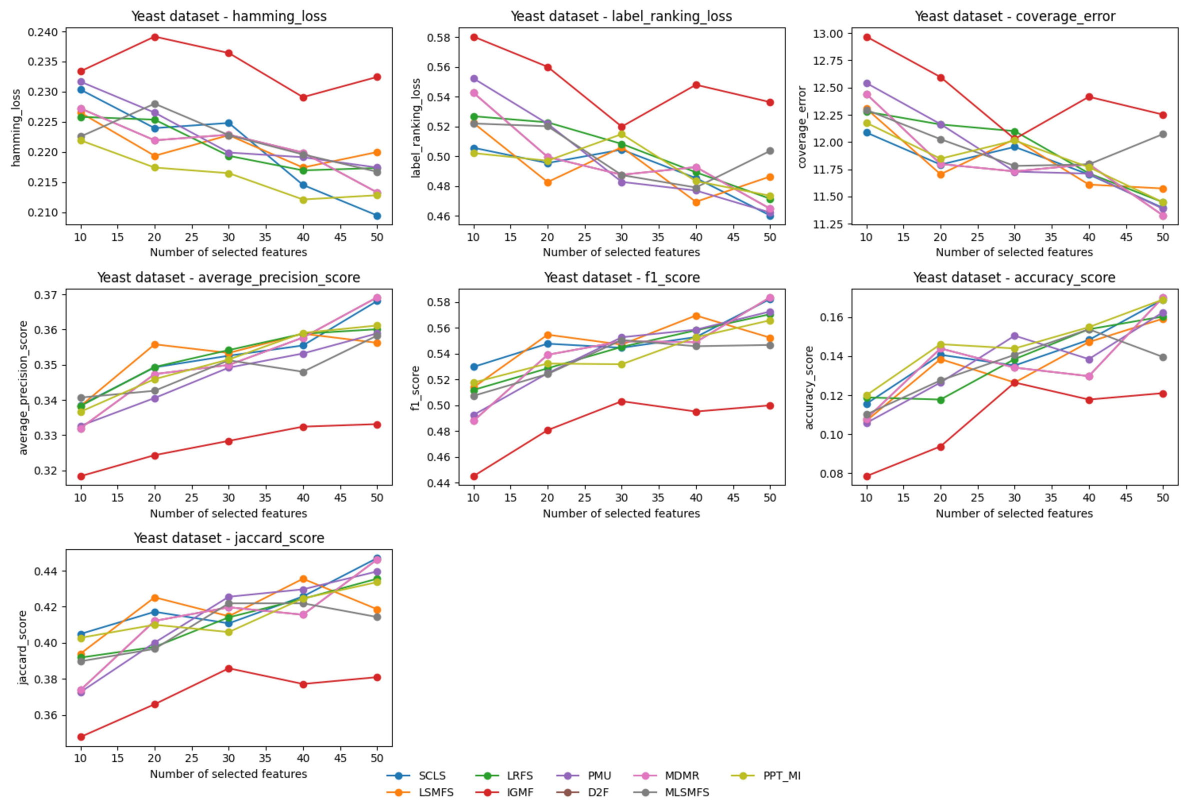

Fig. 2. Evolution of performance metrics with the number of selected features for each method on the *Yeast* dataset.

Overall, the comparative analysis highlights that while most advanced information-theoretic feature selection methods are robust, the PPT_MI and LRFS methods stand out on the Emotions benchmark. This suggests that integrating problem transformation with mutual information (PPT_MI) or leveraging label redundancy-aware selection (LRFS) is particularly effective for multi-label emotion recognition.

Analysis. Across the Yeast dataset, methods such as D2F and MDMR obtain the highest F1 and Jaccard scores, while SCLS, PMU, and PPT_MI are highly competitive, particularly on Hamming Loss and precision-oriented metrics. In contrast, IGMF yields lower scores on most metrics. The overall comparison confirms the robustness of advanced feature selection approaches on biological data, with some methods showing clear advantages on specific criteria.

Figure 2 illustrates the metric curves for the Yeast dataset as the number of selected features increases. The performance trends generally mirror those observed on the Emotions dataset, with certain methods excelling in specific metrics as feature count grows.

6. Conclusion

This study conducted a comprehensive experimental comparison of nine state-of-the-art information-theoretic feature selection methods for multi-label classification, focusing on the widely-used *Emotions* and *Yeast* datasets. The evaluation, based on seven standard metrics and the ML-KNN classifier, demonstrates that all advanced methods provide competitive results. However, certain methods such as PPT_MI and LRFS stand out, particularly in terms of F1-score, Jaccard score, and overall classification accuracy.

The results suggest that integrating problem transformation or leveraging label redundancy in the feature selection process can significantly enhance performance for multi-label problems. Furthermore, the experimental curves highlight that the benefit of adding features generally plateaus after a certain threshold, emphasizing the importance of effective feature ranking.

Overall, this benchmark contributes to a deeper understanding of the strengths and limitations of recent MLFS algorithms and provides practical guidance for method selection in real-world multi-label applications.

Future Work: As part of our future research directions, we intend to enhance the classification of MLFS methods by evaluating their suitability for specific application contexts. This will involve exploring ensemble-based approaches, advancing information-theoretic techniques, and devising strategies that more effectively leverage label dependencies. Additionally, we plan to propose a new feature relevance measure that integrates label-importance weighting and redundancy derived from label dependencies, with the objective of further improving MLFS performance.

[1] Chamlal H., Aaboub F., Ouaderhman T. A preordonance-based decision tree method and its parallel implementation in the framework of Map-Reduce. Applied Soft Computing. **167** (A), 112261 (2024).

[2] Chamlal H., Rebbah F. E., Ouaderhman T. An ensemble classifier combining Dempster–Shafer theory and feature selection methods aggregation strategy. Applied Soft Computing. **180**, 113306 (2025).

[3] Cover T. M., Thomas J. A. Elements of Information Theory. John Wiley & Sons, Inc. (2006).

[4] Kashef S., Nezamabadi-pour H., Nikpour B. Multilabel feature selection: A comprehensive review and guiding experiments. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. **8** (2), e1240 (2018).

[5] Zhang M.-L., Li Y.-K., Liu X.-Y., Geng X. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science. **12**, 191–272 (2018).

[6] Kamalov F., Sulieman H., Alzaatreh A., Emarly M., Chamlal H., Safaraliev M. Mathematical Methods in Feature Selection: A Review. Mathematics. **13** (6), 996 (2025).

[7] Peng H., Long F., Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. **27** (8), 1226–1238 (2005).

[8] Chamlal H., Ouaderhman T., Aaboub F. Preordonance correlation filter for feature selection in the high dimensional classification problem. 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS). 1–5 (2021).

[9] Chamlal H., Ouaderhman T., Rebbah F. E. A novel filter based feature selection approach for microarray dataset. 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS). 1–6 (2021).

[10] Spolaôr N., Cherman E. A., Monard M. C., Lee H. D. A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science. **292**, 135–171 (2013).

[11] Read J. A pruned problem transformation method for multi-label classification. Proceedings of New Zealand Computer Science Research Student Conference, NZCSRSC 2008. 143–150 (2008).

[12] Lee J.-S., Kim D.-W. Mutual information-based multi-label feature selection using interaction information. Expert Systems with Applications. **42** (4), 2013–2725 (2015).

[13] Zhang P., Liu G., Gao W. Distinguishing two types of labels for multi-label feature selection. Pattern Recognition. **95**, 72–82 (2019).

[14] Zhang P., Liu G., Gao W., Song J. Multi-label feature selection considering label supplementation. Pattern Recognition. **120**, 108137 (2021).

[15] Lee J., Kim D.-W. Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters. **34** (3), 349–377 (2013).

[16] Lee J., Kim D.-W. SCLS: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition. **66**, 342–372 (2017).

[17] Ma X.-A., Liu H., Liu Y., Zhang J. Z. Multi-label feature selection considering label importance-weighted relevance and label-dependency redundancy. European Journal of Operational Research. **322** (1), 215–236 (2025).

[18] Han Q., Zhao Z., Hu L., Gao W. Enhanced multi-label feature selection considering label-specific relevant information. Expert Systems with Applications. **264**, 125819 (2025).

[19] Zhang S., Li Y., Zhang P., Gao W. Exploring multi-label feature selection via feature and label information supplementation. Engineering Applications of Artificial Intelligence. **159** (A), 111552 (2025).

[20] Pereira R. B., Plastino A., Zadrozny B., Merschmann L. H. C. Correlation analysis of performance measures for multi-label classification. Information Processing & Management. **54** (3), 359–369 (2018).

Теорія інформації в багатомітковому виборі ознак: аналітичний огляд

Мескауї М., Чамлал Х., Уадерман Т.

Лабораторія комп'ютерних наук та систем (LIS), Факультет наук Айн Чок, Університет Хасана II у Касабланці, Марокко

У контексті багатоміткового навчання, вибір ознак (MLFS) є ключовим процесом для обробки багатовимірних наборів даних, метою якого є збереження найбільш інформативних ознак, зберігаючи при цьому зв'язки між мітками. Це дослідження є широким оглядом сучасних підходів MLFS, заснованих на принципах теорії інформації. У статті спочатку подано фундаментальні концепції теорії інформації, а потім наведено детальний огляд репрезентативних методів MLFS разом з їх теоретичною базою. Оцінка ефективності проводиться на реальних багатоміткових наборах даних, що дозволяє виділити переваги та недоліки кожного методу. Для порівняння використано широко використовувані метрики оцінки, такі як втрати Хеммінга, точність, втрати ранжування міток та F1-оцінка. Також окреслено перспективи майбутніх досліджень, включаючи розробку нового критерію релевантності ознак, який інтегрує зважування важливості міток та зменшення надлишковості на основі залежностей міток, з метою підвищення точності як вибору ознак, так і багатоміткової класифікації.

Ключові слова: *адаптація алгоритму; вибір ознак за кількома мітками; машинне навчання; теорія інформації; інформація про взаємодію; трансформація задачі.*