Expert Systems With Applications 309 (2026) 131177

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

LIMFS: Label interaction-aware multi-label feature selection

Ying Yu [a,b,*], Bowen Li [b], Jin Qian [b], Wenhao Shu [b], Duoqian Miao [c]

[a] State Key Laboratory of Safety and Resilience of Civil Engineering in Mountain Area, East China Jiaotong University, Jiangxi, Nanchang, 330013, China
[b] School of Information and Software Engineering, East China Jiaotong University, Nanchang, Jiangxi, 330013, China
[c] School of Information and Software Engineering, Tongji University, Shanghai, Shanghai, 201804, China

ARTICLE INFO

Keywords:
Multi-label feature selection
Dynamic label interaction
Information theory
Feature weighting
High-dimensional data

ABSTRACT

Multi-label feature selection (MLFS) plays a critical role in addressing high-dimensional data challenges by identifying relevant features and minimizing redundancy to improve classification performance. However, traditional MLFS approaches often assess feature–label correlations independently, thereby ignoring the dynamic, feature-conditioned interactions among labels. To address this limitation, we propose a novel filter-style framework called Label Interaction-aware Multi-label Feature Selection (LIMFS), which explicitly incorporates dynamic label interactions into the feature evaluation process through three core components: Feature-conditioned Interaction Strength (FIS), Interaction-Enhanced Relevance (IER), and Label Interaction Enhancement (LIE). Specifically, FIS quantifies how a candidate feature dynamically modifies the dependency between label pairs, providing a quantitative foundation for interaction-aware feature scoring. IER and LIE serve as complementary relevance metrics from global and local perspectives, respectively. IER adjusts the estimation of global feature relevance by integrating FIS-aware weights, ensuring the selected features align with the overall structure of the label space. In contrast, LIE captures the local positive information gain derived from enhancing meaningful label pairs, focusing on the fine-grained value brought by feature-driven label interaction optimization. Extensive experiments on eight benchmark datasets demonstrate that LIMFS consistently outperforms nine state-of-the-art multi-label feature selection methods across multiple evaluation metrics, confirming its effectiveness in capturing feature-sensitive label interactions to improve feature selection performance.

1. Introduction

In recent years, multi-label learning (Liu et al., 2022) has gained significant attention due to its broad applicability in diverse real-world scenarios, such as text classification (Zhou et al., 2024a), image analysis (Feng et al., 2025; Singh et al., 2024), video analysis (Soykök & Güvenir, 2025), and gene function classification (Zheng et al., 2024). Unlike traditional single-label learning, where each instance is associated with only one label, a multi-label instance can be associated with multiple labels simultaneously. For example, a gene may be involved in multiple biological processes, or an image could be annotated with both 'city' and 'crowd'.

Similar to single-label problems, multi-label problems often suffer from the so-called *curse of dimensionality*, where a high-dimensional feature space contains many irrelevant or redundant attributes. This issue not only degrades classification performance (e.g., causing overfitting and sensitivity to noisy features) but also increases computational and storage costs. To address these challenges, multi-label feature selection (MLFS) (Qian et al., 2023) has been introduced as a solution. It aims to identify compact yet informative feature subsets that preserve the discriminative power across all relevant labels while eliminating redundancy, thereby enhancing both predictive accuracy and operational efficiency.

Existing MLFS approaches can be categorized into three main types (Pereira et al., 2018): filter, wrapper, and embedded methods. Filter methods (Hancer et al., 2024; Ma et al., 2025; Wang et al., 2024; Yu et al., 2024a) evaluate feature relevance and redundancy based on intrinsic data properties, independent of any specific classifier. This independence makes them scalable and fast. Wrapper methods (Dong et al., 2020), on the other hand, assess feature subsets based on their performance under a chosen classifier, often achieving higher accuracy but at the cost of greater computational expense. Embedded methods (Hao et al., 2025; Zou et al., 2024) incorporate feature selection directly into the model training process, offering a balance between effectiveness and

* Corresponding author.
E-mail addresses: 2837@ecjtu.edu.cn, yuyingjx@163.com (Y. Yu), 1305519027@qq.com (B. Li), qjqjlqyf@163.com (J. Qian), shuwenhao@126.com (W. Shu), dqmiao@tongji.edu.cm (D. Miao).

https://doi.org/10.1016/j.eswa.2026.131177
Received 9 October 2025; Received in revised form 17 December 2025; Accepted 7 January 2026
Available online 10 January 2026
0957-4174/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

efficiency. Among these, filter-based methods have attracted considerable attention due to their simplicity, scalability, and independence from classifiers.

Although filter-based MLFS algorithms have demonstrated success in various contexts, most existing approaches treat label dependencies as static and evaluate the relevance between feature and label independently. In reality, however, correlations between labels are not globally fixed or static. Rather, they are highly feature-dependent and vary with the specific features present in the input instance. Taking news classification as an example, an article discussing how Apple's new smartwatch impacts healthy lifestyles and stock market performance might be assigned labels 'Technology', 'Finance', and 'Health'. When features such as 'stock prices', 'financial reports', and 'investment' are dominant, the labels 'Technology' and 'Finance' exhibit frequent co-occurrence, resulting in strong correlation between them. Conversely, when features like 'heart rate monitoring', 'calorie counting', and 'sleep tracking' are prominent, the correlation between 'Technology' and 'Health' becomes significantly high, while the label 'Finance' becomes nearly independent from both 'Technology' and 'Health'. If the dynamic and contextual interactions among labels are ignored, this oversimplification can lead to suboptimal feature selection, especially in cases where feature-induced label interactions are prominent. Therefore, MLFS approaches should be capable of capturing and modeling such dynamic, feature-conditioned label interactions, rather than simply relying on a static global correlation matrix.

To overcome the aforementioned limitation, we propose a novel filter-based MLFS method named Label Interaction-aware Multi-label Feature Selection(LIMFS). Unlike existing MLFS approaches that neglect the dynamic inter-label dependencies, LIMFS explicitly models such dependencies through a newly defined Feature-conditioned Interaction Strength (FIS) measure. Then, label interaction weights are constructed based on FIS to adaptively adjust the contribution of each label to feature relevance scoring, enabling a more accurate and context-sensitive evaluation of features. Specifically, LIMFS integrates dynamic label interactions through three complementary components. FIS quantifies how candidate features conditionally influence the dependencies between labels. IER (Interaction-Enhanced Relevance) leverages these interactions to assess the global relevance of features, while LIE (Label Interaction Enhancement) further captures the local information gain brought by enhancing interactions. By jointly exploiting these components, LIMFS effectively models complex label dependencies and achieves more precise feature selection.

The main contributions of this paper are summarized as follows:

1. We introduce the FIS metric to capture feature-conditioned, dynamic dependencies between label pairs. Unlike static correlation measures, FIS reveals whether the presence of one label suppresses, enhances, or has no effect on the relevance of a candidate feature to another label, thereby capturing feature-sensitive label interactions.
2. we design the IER component to refine feature relevance estimation by incorporating interaction-aware weights derived from FIS. It adaptively amplifies the importance of features under enhancing interactions and reduces their importance under suppressive interactions, thereby yielding a more accurate evaluation of features from a global perspective.
3. We propose the LIE component to explicitly quantify the positive information gain of candidate features under enhancing interactions from a local perspective, complementing IER with fine-grained modeling of higher-order label dependencies.
4. We validate the effectiveness of LIMFS algorithm through extensive experiments on multiple benchmark datasets, where it consistently outperforms nine state-of-the-art MLFS methods.

2. Background

2.1. Information theory

Information theory (Cover & Thomas, 2006; Shannon, 2001) provides a powerful foundation for quantifying dependencies among random variables. Owing to their ability to capture both linear and nonlinear relationships, information-theoretic criteria are particularly effective for handling complex, high-dimensional multi-label problems (Lee & Kim, 2016; Li et al., 2017). In multi-label feature selection, it is widely utilized to evaluate feature relevance, redundancy, and label interactions (Doquire & Verleysen, 2011; Pereira et al., 2018; Spolaôr et al., 2016). This section briefly introduces the essential information-theoretic concepts and formulations employed in this study.

Information entropy (Cover & Thomas, 2006): Entropy measures the uncertainty or information content of a discrete random variable X. It is defined as:

$$\mathrm{H}(X) = -\sum_{x\in\chi} \mathrm{p}(x)\log \mathrm{p}(x), \tag{1}$$

where X is a discrete variable over the finite set χ, and $\mathrm{p}(x)$ denotes the probability mass function of X.

Mutual information (MI) (Cover & Thomas, 2006): MI measures the shared information between two random variables X and Y. It reflects the reduction in the uncertainty of X given knowledge of Y, and can be expressed as:

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \sum_{x\in\chi}\sum_{y\in\gamma} \mathrm{p}(x,y)\log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\,\mathrm{p}(y)}. \tag{2}$$

Conditional mutual information (CMI) (Cover & Thomas, 2006): CMI quantifies the dependency between X and Y given a third variable Z. It captures the remaining association after accounting for the influence of Z:

$$\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y,Z). \tag{3}$$

Standard chain rules and conditional identities are summarized in (Cover & Thomas, 2006).

Joint mutual information (JMI) (Cover & Thomas, 2006): JMI evaluates the collective information that a pair of variables (X,Y) shares with another variable Z:

$$\mathrm{I}(X,Y;Z) = \mathrm{H}(X,Y) - \mathrm{H}(X,Y \mid Z). \tag{4}$$

JMI and related criteria have been used as feature selection scores; see Li et al. (2014) and Lin et al. (2015) for applications in multi-label feature selection.

Interaction information (II) (Cover & Thomas, 2006): II characterizes higher-order interaction among three variables X, Y, and Z. It measures whether the shared information between X and Y is strengthened or weakened in the presence of Z:

$$\mathrm{I}(X;Y;Z) = \mathrm{I}(X;Z) + \mathrm{I}(Y;Z) - \mathrm{I}(X,Y;Z). \tag{5}$$

Approaches exploiting II in multi-label feature selection are described in Lee and Kim (2015).

2.2. Related work

Multi-label feature selection (MLFS) methods are commonly categorized into three main paradigms based on their interaction with the learning model (Li et al., 2017; Liu et al., 2022; Pereira et al., 2018): Filter, Wrapper, and Embedded methods. Wrapper methods treat MLFS as a search problem, employing heuristic strategies to evaluate feature subsets by directly training and testing a classifier. Despite typically achieving higher performance by tailoring subsets to the classifier, this process is computationally prohibitive due to repeated model retraining, severely limiting its use for large-scale or high-dimensional data. Hence, wrapper methods are uncommon in multi-label feature selection.

Embedded methods seek to strike a balance between the speed of Filter methods and the performance of Wrapper methods. They integrate the feature selection process directly into the learning model's training objective function. This is typically achieved by applying a sparse regularization constraint (such as the ℓ_1-norm or $\ell_{2,1}$-norm) to the feature weight matrix, which forces the weights of irrelevant features to zero. Thus feature selection is performed synchronously with the model parameter learning by solving a single optimization problem. This unified framework allows embedded methods to consider the classifier's inductive bias while maintaining a much higher efficiency than Wrapper methods. Both MFS-FR (Zhou et al., 2024b) and FLFS (Zhang et al., 2025a) exemplify the modern use of optimization and sparsity constraints in embedded MLFS.

Filter methods are characterized by their complete independence from the specific learning algorithm. Widely recognized as one of the most mainstream methods in the field of multi-label feature selection, these methods generally evaluate the intrinsic goodness of a feature based on statistical metrics or information theory measures, such as mutual information, correlation, or feature variance, to assess its relevance to the labels and its redundancy with other features.

To improve computational scalability with large label sets, Lee and Kim (2017) proposed a filter-style multi-label feature selection algorithm SCLS. It introduces a redundancy scaling term based on the entropy of candidate features:

$$J_{\text{SCLS}}(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in S} \frac{I(f_k; f_j)}{H(f_k)} \sum_{l_i \in L} I(f_k; l_i) = \left(1 - \sum_{f_j \in S} \frac{I(f_k; f_j)}{H(f_k)}\right) \sum_{l_i \in L} I(f_k; l_i) \tag{6}$$

Zhang et al. (2019) introduced LRFS, a redundancy-aware filter-style method based on conditional mutual information, while Lin et al. (2015) extended the classical mRMR approach to the multi-label setting with their MDMR algorithm, jointly modeling relevance and redundancy:

$$J_{\text{LRFS}}(f_k) = \sum_{f_j \in S} \sum_{l_i \in L} \left[I\left(f_k; l_i\right) - I\left(f_k, l_i; f_j\right) \right]. \tag{7}$$

To address the inadequate computational efficiency and poor adaptability in complex label spaces, Zhang and Gao (2021) proposed Dual Conditional Relevance Multi-label Feature Selection (DCR-MFS), which evaluates candidate features by considering both their relevance to labels and their redundancy with already selected features:

$$J_{\text{DCR-MFS}}(f_k) = \sum_{f_s \in S} \sum_{l_j \in L} I(f_k; l_j \mid f_s) + \sum_{l_j \in L} \sum_{l_q \in L \setminus \{l_j\}} I(f_k; l_j \mid l_q) - \sum_{f_s \in S} I(f_k; f_s). \tag{8}$$

Zhang et al. (2022) proposed WFRFS, a dynamically weighted multi-label feature selection method that combines label-specific uncertainty with feature–label relevance to produce a weighted relevance score for each feature. In this approach, labels are weighted according to their remaining uncertainty conditioned on a set of reference features, and the final score penalizes redundancy with those reference features. For a candidate feature f_k, the WFRFS score is given by:

$$J_{\text{WFRFS}}(f_k) = \sum_{l_j \in L} \left(\sum_{f_s \in S} \frac{H(l_j \mid f_s)}{H(l_j)} \right) I(f_k; l_j) - \sum_{f_s \in S} I(f_k; f_s), \tag{9}$$

Recently, Ma et al. (2025) introduced the LIWR-LDR method, which quantifies the dependency between the feature and label using uncertainty coefficients and incorporates label importance-weighted relevance and label-dependency redundancy:

$$J_{\text{LIWR-LDR}}(f_k) = \sum_{l_j \in L} I(f_k; l_j) \sum_{l_q \in L} I(l_j; l_q) - \sum_{f_s \in S} \sum_{l_j \in L} \frac{H(f_s) - H(f_s \mid l_j)}{H(f_s)} I(f_k; f_s). \tag{10}$$

Dai et al. (2024) propose SRLG-LMA, a filter-style method that augments MI-based relevance with label-label mutual aid and a clipped strongly-relevant gain. For a candidate feature f_k, the score is

$$J_{\text{SRLG-LMA}}(f_k) = \sum_{l_i \in L} \left[I(f_k; l_i) + \sum_{l_j \in L \setminus \{l_i\}} I\left(f_k; l_j \mid l_i\right) + \text{SRLG}(f_k, l_i, S) \right] - \sum_{f_j \in S} I(f_k; f_j), \tag{11}$$

where $\text{SRLG}(f_k, l_i, S) = \max\{0,\ I(f_k; l_i \mid S) - I(S; l_i)\}$.

Zhang et al. (2025b) propose FLIS, a filter-style MLFS method that augments plain feature–label relevance with *supplementary* information from label-label and selected-feature-label interactions, while penalizing redundancy with the already-selected features. For a candidate feature f_k and label set L, FLIS scores f_k can be calculated by

$$J_{\text{FLIS}}(f_k) = \sum_{l_i \in L} \left\{ I(f_k; l_i) + \sum_{l_j \in L \setminus \{l_i\}} \max\left\{0,\ I(f_k; l_i; l_j) + I(f_k; l_i; f_j)\right\} \right\} - \sum_{f_j \in S} I(f_k; f_j). \tag{12}$$

LSRIFS (Han et al., 2025) represents a novel strategy in information-theoretic filtering that emphasizes the micro-distribution of relevance. Traditional greedy approaches focus on maximizing the overall mutual information but overlook how this relevance is distributed across individual labels. LSRIFS resolves this by introducing a label-specific relevance weight (θ_k), which is derived from the Cauchy-Schwarz inequality. This mechanism prioritizes features that offer highly concentrated and non-uniform information to specific labels, ensuring the selection of truly discriminative features over "mediocre" features that have only weak, dispersed correlations with all labels.

In summary, existing MLFS methods have made significant progress in modeling feature relevance and redundancy. However, they typically treat label dependencies as globally static and do not explicitly model how these dependencies vary with different candidate features. In this work, we propose a dynamic label interaction-aware MLFS framework that addresses this limitation by explicitly modeling how candidate features influence label relationships during feature scoring.

3. Proposed method: LIMFS

In multi-label learning, label correlations play a crucial role in guiding effective feature selection. However, most existing multi-label feature selection (MLFS) methods regard label dependencies as static and context-independent, assuming that the label relationships remain unchanged regardless of which candidate feature is being evaluated. This assumption may be at odds with reality and the introduction of a candidate feature can dynamically alter the semantic structure among labels. For example, in a medical diagnosis task, the presence of a feature 'high blood glucose' may significantly strengthen the correlation between the labels 'diabetes' and 'obesity' while weakening the association with other labels such as 'hypertension'. This means that the same feature can simultaneously amplify certain label dependencies and weaken others, depending on the underlying data distribution. Ignoring such feature-induced structural variation can lead to inaccurate estimations of label relevance and compromise the selection of truly discriminative features.

To overcome this limitation, we propose a dynamic modeling framework LIMFS that conditions inter-label dependencies on each candidate feature. This design enables the model to capture differential interaction between labels when specific features are present, and to model both asymmetric and feature-sensitive label relations. In this section, we detail the proposed Label Interaction-aware Multi-label Feature Selection (LIMFS) framework. We begin in Section 3.1 by introducting necessary notations and formulating the multi-label feature selection objective under the Maximum Relevance Minimum Redundancy (MRMR) principle.

Subsequently, we elaborate on the three core components of LIMFS: Feature-condifitoned Interaction Strength (FIS), Interaction-Enhanced Relevance (IER), and Label Interaction Enhancement (LIE). These components collectively enable LIMFS to explicitly incorporate the dynamic influence of feature-conditioned label interactions into the feature evaluation process.

3.1. Problem formulation and overall objective

Let $D = \{(x_i, L_i)\}_{i=1}^{n}$ represent a multi-label dataset, where $x_i \in \mathbb{R}^d$ denotes the feature vector of the i-th instance and $L_i \in \{0,1\}^q$ is the corresponding label vector over the label set $\mathcal{L} = \{l_1, \ldots, l_q\}$. Let $\mathcal{F} = \{f_1, \ldots, f_d\}$ be the complete set of features, and Let K be the desired size of the feature subset to be selected. Traditional MLFS can be regarded as an instance of the Maximum Relevance Minimum Redundancy (MRMR) framework, aiming to identify a feature subset $S \subseteq \mathcal{F}$ with $|S| = K$ that maximizes relevance to the label set while minimizing redundancy among the selected features. Formally, this objective can be expressed as

$$\begin{aligned} S^{\star} &= \arg\max_{S:|S|=K} \Big[\mathrm{I}(S;\mathcal{L}) - \lambda\,\mathrm{R}(S)\Big], \\ &= \arg\max_{S:|S|=K} \Big[I(S;\mathcal{L}) - \lambda \sum_{f_i,f_j\in S} I(f_i;f_j)\Big], \end{aligned} \tag{13}$$

where mutual information $\mathrm{I}(S,\mathcal{L})$ measures the relevance between the candidate features and the label set, and $\mathrm{R}(S)$ penalizes redundancy among the features in S. λ is a balancing parameter.

Since the relevance term $I(S;\mathcal{L})$ is computationally intractable, especially in the presence of high-dimensional feature subsets, a greedy selection strategy is commonly adopted. At each iteration, the feature $f \notin S$ that maximizes the marginal gain $\Delta(f)$ would be selected.

$$\Delta I(f) = I(f;\mathcal{L} \mid S) \tag{14}$$

Using the chain rule of mutual information, the relevance between a feature f and the label set $\mathcal{L}$ can be decomposed as:

$$I(f;\mathcal{L}) = \sum_{i=1}^{q} I(f;l_i) - \sum_{i<j} I(f;l_i;l_j) + \sum_{i<j<k} I(f;l_i;l_j;l_k) - \cdots. \tag{15}$$

In order to simplify the problem, we omit S here. As can be seen from Eq. (15), the first term captures independent feature-label relevance. The higher-order interaction terms describe how the relevance of f to one label is modulated by the presence of other labels. Therefore, traditional MLFS methods implicitly approximate $I(f;\mathcal{L})$ using only the first-order term $\sum_i I(f;l_i)$, which ignores the rich dependencies among labels.

To correct this flaw, we approximate $\Delta I(f)$ as

$$I(f;\mathcal{L} \mid S) \approx \mathrm{IER}(f) + \mathrm{LIE}(f), \tag{16}$$

where Interaction-Enhanced Relevance (IER) captures the basic, global correlation between the feature f and each label l, while Label Interaction Enhancement (LIE) captures local positive interaction gains. Then, by integrating the relevance approximation and redundancy penalty, the final evaluation function of LIMFS is defined as follows, with $\lambda = 1$ implicitly set.

$$J(f) = \mathrm{IER}(f) + \mathrm{LIE}(f) - \sum_{f_j\in S} I(f;f_j), \tag{17}$$

which serves as a greedy approximation to the marginal objective $I(f;\mathcal{L} \mid S) - \sum_{f_j\in S} I(f;f_j)$. The final term $\sum_{f_j\in S} I(f;f_j)$ penalizes redundancy between the candidate feature f and the previously selected features. Therefore, $J(f)$ is an information-theoretic criterion (similar to MRMR) where the 'Max-Relevance' part is theoretically augmented by 'IER + LIE' to address the inherent structural dependencies of the label space, making it a more principled surrogate for $I(f;\mathcal{L}|S)$ than existing MLFS filters.

The final scoring function $J(f)$ assigns equal weights to the IER and LIE terms. This design is deliberate and well-justified. Both terms serve as complementary components within a unified information-theoretic objective, namely approximating the conditional mutual information $I(f;L \mid S)$. IER captures relevance from a global, interaction-weighted perspective, whereas LIE recovers the strictly positive second-order gains from enhancing label pairs. Since both are expressed in mutual-information units, their direct summation provides a natural and unbiased surrogate for the overall relevance of the candidate feature. Introducing an additional weighting parameter would not only necessitate costly tuning but also raise the risk of over-fitting to dataset-specific patterns. Equal weighting avoids these drawbacks, keeping the method fully unsupervised, stable, and reproducible, in line with the filter-based feature-selection paradigm.

The following subsections detail the design of the three core components, namely Directed Interaction Strength (FIS), Interaction-Enhanced Relevance (IER), and Label Interaction Enhancement (LIE).

3.2. Feature-conditioned label interactions modeling

As mentioned above, conventional MLFS methods usually assume that the relationships between labels are static and invariant. However, in real-world scenarios, the semantic structure among labels is not fixed but evolves dynamically with changes in features. Such feature-induced dynamic interaction property between labels cannot be effectively modeled by conventional measures. To address this limitation, we propose the Feature-conditioned Interaction Strength (FIS) metric, which is specifically designed to capture such dynamic inter-label interactions. This directed and bounded measure is capable of quantifying how the mutual information between a candidate feature f and a target label l_i is influenced by the presence of another label l_j.

3.2.1. Feature-conditioned Interaction Strength (FIS)

Definition 1 (Feature-conditioned Interaction Strength, FIS)**.** Given a candidate feature f and two distinct labels l_i and l_j, the $FIS(f;l_i;l_j)$ metric is defined as:

$$FIS(f;l_i;l_j) = \frac{I(f;l_i) - I(f;l_i \mid l_j)}{I(f;l_i) + I(f;l_i \mid l_j)}, \tag{18}$$

where $I(f;l_i)$ denotes the mutual information between f and l_i, while $I(f;l_i \mid l_j)$ denotes the conditional mutual information between f and l_i given l_j. The core motivation for normalizing the difference $I(f;l_i) - I(f;l_i|l_j)$ by the sum $I(f;l_i) + I(f;l_i|l_j)$ is to ensure the boundedness and standardization of the FIS metric. The FIS score lies in the interval $[-1,1]$, where the sign reflects the interaction direction (enhancing or suppressive), and the magnitude indicates the degree of interaction disturbance caused by label l_j.

Property 1 (Boundedness): For any candidate feature f and any pair of labels l_i, l_j, the feature-conditioned interaction strength satisfies: $|FIS(f;l_i;l_j)| \le 1$. Justification: By definition, $|I(f;l_i) - I(f;l_i \mid l_j)| \le I(f;l_i) + I(f;l_i \mid l_j)$. Since both mutual information $I(f;l_i)$ and conditional mutual information $I(f;l_i \mid l_j)$ are non-negative, the denominator is always positive. Meanwhile, the numerator is bounded by the sum in the denominator: $|I(f;l_i) - I(f;l_i \mid l_j)| \le I(f;l_i) + I(f;l_i \mid l_j)$. Thus, $|FIS(f;l_i;l_j)| \le 1$ always holds; namely, the FIS score always lies within the interval of $[-1,1]$.

This boundedness guarantees that FIS provides a standardized scale for comparing interaction effects across different feature-label-label triplets, independent of absolute entropy values or underlying data distributions.

Property 2 (Asymmetry): FIS is generally asymmetric with respect to the label order: $FIS(f;l_i;l_j) \neq FIS(f;l_j;l_i)$.

Justification: This asymmetry naturally stems from the non-symmetric nature of conditional mutual information, i.e., $I(f;l_i \mid l_j) \neq I(f;l_j \mid l_i)$, unless labels l_i and l_j are conditionally symmetric with

respect to feature f, which is rarely observed in real-world multi-label data. As a result, the corresponding FIS values are computed as $FIS(f;l_i;l_j) = \frac{I(f;l_i)-I(f;l_i|l_j)}{I(f;l_i)+I(f;l_i|l_j)}$ and $FIS(f;l_j;l_i) = \frac{I(f;l_j)-I(f;l_j|l_i)}{I(f;l_j)+I(f;l_j|l_i)}$, which are generally not equal due to the asymmetry in both the numerators and denominators.

This directional property enables FIS to identify the nuanced and asymmetric dependencies among labels in the presence of a specific feature and then model directional influence, namely how the presence of label l_j affects the information relevance of f to label l_i.

Although other information-theoretic metrics like Interaction Information (II) have been used in Multi-label Feature Selection (MLFS) , the unique design of FIS is intended to address the limitations of existing methods that typically overlook dynamic, feature-conditioned label interactions.

The numerator $I(f;l_i) - I(f;l_i \mid l_j)$ of FIS captures the absolute change in mutual information, which aligns with the classical Interaction Information $I(f;l_i;l_j)$. However, using this raw difference alone is problematic because the scale of mutual information can vary widely across different feature-label pairs, making direct comparisons unfair.

Therefore, it is necessary to normalize the difference by the sum $I(f;l_i) + I(f;l_i \mid l_j)$. This denominator serves as a stable reference that reflects the total information content under both unconditional and conditional settings. The resulting ratio is bounded in $[-1, 1]$, providing a standardized measure that is comparable across diverse features and label pairs. Importantly, unlike symmetric interaction information, FIS is asymmetric with respect to l_i and l_j, allowing it to distinguish between the influence of l_j on the $f - l_i$ link and the reverse influence, which is a crucial property for modeling directed label interactions in multi-label contexts.

Alternative normalization schemes (e.g., dividing by $I(f;l_i)$ or $\sqrt{I(f;l_i) \cdot I(f;l_i \mid l_j)}$) were explored during preliminary experiments. The chosen form proved most robust in maintaining interpretability and stability when $I(f;l_i)$ or $I(f;l_i \mid l_j)$ approaches zero, while still preserving the direction and relative strength of interactions.

Note that the term 'feature-conditioned' in FIS indicates that the interaction strength depends on which candidate feature is being evaluated. It does not imply that interactions are updated dynamically during the iterative feature selection process. Each FIS value is computed statically before selection begins, capturing how a specific feature alters label dependencies relative to the full feature space.

3.2.2. Categories of label interactions

According to the previous analysis, the relevance of f to label l_i is generally influenced by other labels, namely $I(f;l_i) \neq I(f;l_i \mid l_j)$. This difference reflects feature-conditioned label interaction and corresponds to the interaction information:

$$I(f;l_i;l_j) = I(f;l_i) - I(f;l_i \mid l_j). \tag{19}$$

Therefore, FIS can be interpreted as a normalized measure of the second-order interaction term $I(f;l_i;l_j)$, indicating whether label l_j enhances or suppresses the relevance of feature f to label l_i.

The sign of FIS encodes the direction of the interaction: a positive value indicates that the presence of label l_j suppresses the relevance of feature f to label l_i, while a negative value signifies that l_j improves this relevance. Consequently, according to the value of $FIS(f;l_i;l_j)$, the label interactions between labels can be categorized into three representative types: suppressive, enhancing and neutral interactions. Such categorization enables our method not only to quantify the strength of label interactions, but also to explicitly capture their directionality and functional influence. By identifying suppressive, enhancing, and neutral interactions, the proposed framework adaptively adjusts feature relevance, thereby supporting more informed and effective feature selection in multi-label scenarios.

The sign convention of FIS, where a positive value indicates suppressive interaction and a negative value indicates enhancing interaction, arises naturally from its definition in Eq. (18). While this may initially appear counter-intuitive, it can be clearly understood through the lens of *information gain or loss*.

- Suppressive Interaction: When $FIS(f;l_i;l_j) > 0$, the numerator $I(f;l_i) - I(f;l_i|l_j)$ is positive, meaning that conditioning on l_j *reduces* the mutual information between f and l_i. In other words, l_j "suppresses" or "explains away" some of the information that f provides about l_i, leading to a *loss* of predictive power. The positive sign directly reflects this *increase* in conditional uncertainty or *reduction* in relevance.
- Enhancing Interaction: When $FIS(f;l_i;l_j) < 0$, the numerator is negative, implying that the presence of l_j *increases* the mutual information between f and l_i. Here, label l_j acts as a catalyst that "enhances" the relevance of f to l_i, resulting in an *information gain*. The negative sign thus signifies a *decrease* in conditional uncertainty or an *improvement* in relevance.
- Neutral Interaction: When $FIS(f;l_i;l_j) = 0$, the numerator is zero, indicating $I(f;l_i) = I(f;l_i|l_j)$. Label l_j has a *neutral* effect on the dependency between f and l_i. In information-theoretic terms, this implies conditional independence between f and l_i given l_j with respect to the available information, and no interaction needs to be modeled.

A geometric analogy can be drawn from vector projection: Consider $I(f;l_i)$ as the length of a vector representing the information between f and l_i. Conditioning on l_j is akin to projecting this vector onto a subspace orthogonal to l_j. If the projection is shorter ($I(f;l_i|l_j) < I(f;l_i)$), information is 'lost' due to suppression (positive FIS). If the projection is effectively longer or reveals a new component ($I(f;l_i|l_j) > I(f;l_i)$), information is 'gained' due to enhancement (negative FIS). If the length remains unchanged ($I(f;l_i|l_j) = I(f;l_i)$), the vector is already orthogonal to l_j, and no gain or loss occurs (FIS $= 0$).

This sign convention is therefore not arbitrary but is a direct mathematical consequence of measuring the *change* in mutual information upon conditioning. It ensures that the sign of FIS consistently and unambiguously indicates the *direction of change* in feature-label relevance caused by the presence of another label.

3.3. Feature relevance estimation

Conventional multi-label feature selection (MLFS) methods typically assess the importance of a feature by separately measuring its relevance to each label and then summing them up. Such approaches implicitly assume that each label contributes equally and independently to the importance of features. In reality, however, labels exhibit complex feature-dependent interactions, which have been captured by the Feature-conditioned Interaction Strength (FIS) measure. The relevance of a feature to a particular label may be suppressed or enhanced by the presence of other labels. Namely, label contributions are actually dynamic. To address this limitation, we propose a dynamically interaction-aware weighting scheme that adaptively adjusts the contribution weight of each label based on inter-label interactions quantified by FIS. This mechanism allows the importance estimation to more accurately reflect the true discriminative power of a feature .

3.3.1. Label interaction weight

Definition 2 (Label Interaction Weight)**.** Given a candidate feature f and a pair of labels l_i and l_j, the label interaction weight $w^*(f;l_i,l_j)$ is defined as

$$w^*(f;l_i,l_j) = \begin{cases} 1 - FIS(f;l_i;l_j), & \text{if } FIS(f;l_i;l_j) \geq 0 \\ 1 + |FIS(f;l_i;l_j)|, & \text{if } FIS(f;l_i;l_j) < 0 \end{cases}, \tag{20}$$

This formulation ensures that suppressive interactions ($FIS > 0$) reduce the relevance of feature f to label l_i, while enhancing interactions ($FIS < 0$) increase it. If the interaction is neutral ($FIS = 0$), the weight remains at 1, which aligns with traditional uniform weighting. These

adaptive weights allow the feature selection method to better capture complex label dependencies. As a result, the overall feature selection becomes more accurate and effective in multi-label scenarios.

Since FIS score lies in the interval $[-1, 1]$, the interaction weight $w^*(f; l_i, l_j)$ are confined to the interval $[0, 2]$. This bounded range prevents any single label pair from dominating the relevance estimation and keeps the rescaled mutual-information terms numerically stable and comparable across different candidate features.

3.3.2. Interaction-Enhanced Relevance (IER)

To approximate $I(f; \mathcal{L})$ while retaining interaction information, we re-weight the first-order relevance term $I(f; l_i)$ according to feature-conditioned label interactions. This leads to the interaction-enhanced relevance (IER):

Definition 3 (Interaction-Enhanced Relevance, IER)**.** Based on the interaction weights of the label, we define the relevance of the feature f as follows:

$$\mathrm{IER}(f) = \sum_{i=1}^{q} I(f; l_i)\, \phi(f; l_i), \tag{21}$$

where the correction factor $\phi(f; l_i)$ aggregates the influence of other labels via FIS-based weights:

$$\phi(f; l_i) = \frac{1}{|\mathcal{L}| - 1} \sum_{\substack{l_j \in L \\ j \neq i}} w^*(f; l_i, l_j). \tag{22}$$

When strong label interactions exist between l_i and other labels, $\phi(f; l_i) > 1$, enhancing the importance of feature f. In the presence of label redundancy, $\phi(f; l_i) < 1$, thereby reducing the importance of feature f. Therefore, IER can be interpreted as a corrected first-order approximation of $I(f; \mathbf{L})$, where the contribution of each label is adaptively amplified or attenuated depending on whether interactions with other labels are enhancing or suppressive. When no interaction exists, $\phi(f; l_i) = 1$ and IER reduces to the standard relevance term.

3.3.3. Label Interaction Enhancement (LIE)

Although IER evaluates feature relevance from a global perspective, it does not differentiate the impacts of distinct types of label interactions. Such a unified global assessment is limited, as it conceals the specific roles of various interactions in shaping a feature's discriminative ability. In fact, enhancing interactions are particularly valuable. When a label pair (l_i, l_j) exhibits an enhancing interaction, namely $I(f; l_i \mid l_j) > I(f; l_i)$, the discriminative ability of a candidate feature f for label l_i improves under the conditional constraint of label l_j. This improvement reflects a positive information gain, which directly strengthens the feature's predictive contribution to the target label. In contrast, suppressive interactions reduce discriminative ability, while neutral interactions provide no additional information. Neither yields positive gain. Hence, enhancing interactions should be modeled independently to capture local positive contributions of features, thereby addressing the limitations of IER.

Furthermore, Although IER adjusts the global relevance of a feature by incorporating interaction-aware weights, it can only indicate the relative direction and intensity of interactions and cannot directly quantify the absolute information gain derived from enhancing interactions. To overcome these limitations, we introduce the Label Interaction Enhancement (LIE) metric to explicitly quantify such positive gain.

$$\mathrm{LIE}(f) = \sum_{i \neq j} \max\{0, I(f; l_i \mid l_j) - I(f; l_i)\}. \tag{23}$$

This design selectively accumulates only enhancing interactions, corresponding to the positive part of the second-order interaction information. Suppressive or neutral interactions do not contribute positively to feature relevance and are therefore excluded. Thus, LIE complements IER effectively: IER provides a global adjustment of feature relevance through dynamic weighting, whereas LIE offers a local perspective by explicitly extracting and aggregating the positive gains that are not fully represented in IER. Together, they enable a more comprehensive and accurate assessment of feature importance.

Table 1
Mutual information and conditional mutual information between candidate features and labels.

Feature	$I(f; l_1)$	$I(f; l_2)$	$I(f; l_1 \mid l_2)$	$I(f; l_2 \mid l_1)$
f_1	0.0103	0.0194	0.0441	0.0528
f_2	0.1823	0.1984	0.0626	0.0770

Table 2
Interaction gain, FIS value, interaction type, and inclusion in LIE.

Feature	Pair $(l_i \mid l_j)$	Gain	FIS	Type	In LIE
f_1	$(l_1 \mid l_2)$	+0.0338	-0.6213	Enhancing	Yes
f_1	$(l_2 \mid l_1)$	+0.0334	-0.4626	Enhancing	Yes
f_2	$(l_1 \mid l_2)$	-0.1197	+0.4888	Suppressive	No
f_2	$(l_2 \mid l_1)$	-0.1214	+0.4408	Suppressive	No

To illustrate the effectiveness of LIE, we employ the *Emotions* dataset from the Mulan (Tsoumakas et al., 2011) repository. Specifically, we randomly select a label pair $(l_1, l_2) = (L3, L6)$ and feature pair $(f_1, f_2) = (F_{22}, F_{46})$ to demonstrate the mapping relationships. The mutual information and conditional mutual information values are summarized in Table 1, while the corresponding interaction gains and FIS scores are reported in Table 2. As shown in Table 2, feature f_1 exhibits enhancing interactions with $\Delta_{l1|l2}(f_1) = +0.0338$ (FIS= -0.6213) and $\Delta_{l2|l1}(f_1) = +0.0334$ (FIS= -0.4626), whereas feature f_2 exhibits suppressive interactions with $\Delta_{l1|l2}(f_2) = -0.1197$ (FIS= $+0.4888$)and $\Delta_{l2|l1}(f_2) = -0.1214$ (FIS= $+0.4408$).These empirical results demonstrate that the same label pair can enhance some features while suppressing others, thereby confirming the desired properties of our formulation.

In summary, the proposed LIE(f) explicitly measures the incremental contribution of a candidate feature under enhancing interaction settings, thereby complementing the global assessment capacity of IER(f). By capturing fine-grained interaction dynamics, LIE(f) improves the model's sensitivity to complex enhancing relationships and strengthens its capability to identify high-quality features for multi-label tasks, ultimately enhancing the generalization performance of the feature selection algorithm.

3.4. Label interaction-aware multi-label feature selection

Combining the above analysis, we obtain a tractable approximation for the marginal mutual information $I(f; \mathcal{L}) \approx \mathrm{IER}(f) + \mathrm{LIE}(f)$, where IER corresponds to an interaction-aware first-order approximation of feature-label relevance, while LIE acts as a selective positive second-order correction term. Accordingly, LIMFS can be viewed as a greedy optimizer that approximates the marginal gain $I(f; \mathcal{L} \mid S)$ by jointly considering global relevance and local enhancing interaction effects, while penalizing feature redundancy.

The pseudo code of the proposed Label Interaction-aware Multi-label Feature Selection (LIMFS) algorithm is presented in Algorithm 1, which consists of three main stages. In stage 1 (Lines 1), the selected feature subset S and the iteration counter k are initialized. Stage 2 (Lines 2–4) computes two key metrics for each feature: the Interaction-Enhanced Relevance (IER) and the Label Interaction Enhancement (LIE). Stage 3 (Lines 3–14) performs an iterative feature selection process. In the initial iteration, the feature with the highest IER value is selected. For subsequent iterations, each candidate feature f_j is evaluated using the composite score function $J(f)$. The feature with the highest score is added to S iteratively until the desired number of features K is reached. This selection mechanism ensures that chosen features are globally relevant, locally enhanced, and minimally redundant.

Algorithm 1: LIMFS: Label interaction-aware multi-label feature selection.

Input: A training set D contains a full feature set $F = \{f_1, f_2, \ldots, f_n\}$, and a label set $L = \{l_1, l_2, \ldots, l_q\}$. The number of features to be selected is K.
Output: The selected feature subset S.
$S \leftarrow \emptyset$; $k \leftarrow 0$;
for $i = 1$ *to* n **do**
  Calculate $IER(f_i)$;
  Calculate $LIE(f_i)$;
while $k < K$ **do**
  if $k == 0$ **then**
    Select the feature f_j with the largest $IER(f_j)$ from the set F;
  else
    for *each candidate feature* $f_j \in F$ **do**
      Calculate $J(f_j) = IER(f_j) + LIE(f_j) - \sum_{f_m \in S} I(f_j; f_m)$;
    Select the feature f_j with the largest $J(f_j)$ from the set F;
  $S \leftarrow S \cup \{f_j\}$;
  $F \leftarrow F \setminus \{f_j\}$;
  $k \leftarrow k + 1$;
return S

4. Experiments

This section conducts an extensive empirical evaluation to validate the effectiveness and performance of the proposed LIMFS algorithm. To ensure a systematic and comprehensive assessment, the evaluation framework is structured as follows: Section 4.1 first defines the key evaluation metrics employed to quantify algorithm performance from multiple perspectives; Section 4.2 details the experimental datasets as well as the experimental parameter settings to ensure the reliability of results; Section 4.3 subsequently presents the comparative experimental results and provides an in-depth analysis and discussion, aiming to highlight the advantages of LIMFS over existing state-of-the-art methods. Section 4.4 discusses the time and space complexity of the proposed LIMFS. Section 4.5 presents an ablation study designed to analyze the contribution of key components in the LIMFS framework. Section 4.6 visualizes the FIS matrices for several representative features to intuitively demonstrate the role of label interactions.

4.1. Evaluation metrics

We evaluate the LIMFS using the same set of metrics as most baseline methods, including Label-Weighted F1, Jaccard Index (also known as Jaccard Similarity), Subset Accuracy, and the standard Hamming Loss (Hinojosa Lee et al., 2024). Formally, let the test set be defined as $T = \{(x_i, L_i) \mid i = 1, \ldots, n\}$ and $L = \{l_1, \ldots, l_q\}$ denote the label universe. For each instance x_i, $L_i \subseteq L$ is its ground-truth labels and $\hat{L}_i \subseteq L$ is its predicted labels. Alternatively, the labels of an instance can be expressed in a binary vector form: $L_i = \{l_{i1}, \ldots, l_{iq}\}$ and $\hat{L}_i = \{\hat{l}_{i1}, \ldots, \hat{l}_{iq}\}$, where $l_{ij} = \mathbf{1}[l_j \in L_i]$ and $\hat{l}_{ij} = \mathbf{1}[l_j \in \hat{L}_i]$ indicate the ground-truth and predicted label assignments of the j-th label for the i-th sample, respectively, and $\mathbf{1}[\cdot]$ is the indicator function.

(1) **Hamming Loss (HL)**: The fraction of misclassified label-instance pairs is defined as

$$\mathrm{HL} = \frac{1}{nq} \sum_{i=1}^{n} \sum_{j=1}^{q} \mathbf{1}!\left(l_{ij} \neq \hat{l}_{ij}\right), \tag{24}$$

where lower values indicate better performance.

(2) **F1 Score (label-weighted)**: For the j^{th} label l_j, the numbers of true positivesTP_j, false positives FP_j, and false negatives FP_j are given by

$$\mathrm{TP}_j = \sum_{i=1}^{n} \mathbf{1}(l_{ij} = 1 \wedge \hat{l}_{ij} = 1), \quad \mathrm{FP}_j = \sum_{i=1}^{n} \mathbf{1}(l_{ij} = 0 \wedge \hat{l}_{ij} = 1),$$
$$\mathrm{FN}_j = \sum_{i=1}^{n} \mathbf{1}(l_{ij} = 1 \wedge \hat{l}_{ij} = 0).$$

The per-label F1 score and its label-weighted average are then defined as

$$\mathrm{F1}_j = \frac{2\,\mathrm{TP}_j}{2\,\mathrm{TP}_j + \mathrm{FP}_j + \mathrm{FN}_j}, \qquad \mathrm{F1}^{\mathrm{w}} = \sum_{j=1}^{q} w_j\,\mathrm{F1}_j, \tag{25}$$

where the support weight is $w_j = \frac{\sum_{i=1}^{n} l_{ij}}{\sum_{r=1}^{q} \sum_{i=1}^{n} l_{ir}}$. Higher values correspond to better performance.

(3) **Jaccard Score/Intersection over Union (IoU, label–weighted)**: The per-label Jaccard index and its label–weighted average are given by

$$\mathrm{J}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j + \mathrm{FN}_j}, \qquad \mathrm{JS}^{\mathrm{w}} = \sum_{j=1}^{q} w_j\,\mathrm{J}_j, \tag{26}$$

where the same weights w_j are applied as in the weighted F1 score. Larger values indicate better Performance.

(4) **Subset Accuracy (Exact Match Ratio)**: A prediction is considered correct if and only if the entire predicted label set exactly matches the ground truth:

$$\mathrm{AS}_{\mathrm{subset}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(L_i = \hat{L}_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left((l_{i1}, \ldots, l_{iq}) = (\hat{l}_{i1}, \ldots, \hat{l}_{iq})\right). \tag{27}$$

Higher values correspond to better performance.

All evaluation metrics are implemented using the official `scikit-learn` library (Pedregosa et al., 2011), ensuring consistency and reproducibility.

4.2. Datasets and experimental settings

We evaluate the performance of LIMFS across eight benchmark multi-label datasets retrieved from the Mulan repository (Tsoumakas et al., 2011), specifically: *emotions*, *scene*, *medical*, *yeast*, *enron*, *birds*, *genbase*, and *tmc2007-500*. Detailed statistics of these datasets are summarized in Table 3. For dataset partitioning, we adopted the train/test splits officially recommended by the Mulan repository to ensure consistency with standard experimental protocols.

Feature selection is conducted only on the *training* subset. The feature ranking derived from this process is then applied to both the training and testing subsets. Specifically, each feature selection method extracts the top-m (where $m \in \{1, \ldots, 50\}$) features from its generated rankings. For the classification task, we employed the ML-kNN algorithm with the number of neighbors set to $k = 10$. The ML-kNN implementation is sourced from the `scikit-multilearn` library, while all evaluation metrics were computed using functions from the `scikit-learn` library to ensure result standardization.

For each feature selection method, we calculated its evaluation metric on the test subset for m values ($m \in \{1, \ldots, 50\}$). The entries presented in the subsequent tables represent the mean ± standard deviation of these 50 metric scores (i.e., the average performance and its variability across all tested values of $m \in \{1, \ldots, 50\}$).

4.3. Results and discussion

Tables 4–7 report the performance of LIMFS against night representative MLFS methods, including ATR (Eskandari & Ghassabi, 2023),

Table 3
multi-label datasets.

Name	Instances	Features	Labels	Train	Test	Cardinality
emotions	593	72	6	391	202	1.869
scene	2407	294	6	1211	1196	1.074
medical	978	1449	45	333	645	1.245
yeast	2417	103	14	1500	917	4.237
enron	1702	1001	53	1123	579	3.378
birds	645	260	19	322	323	1.014
genbase	662	1186	27	463	199	1.252
tmc2007-500	28596	500	22	21519	7077	2.220

SCLS (Lee & Kim, 2017), LRFS (Zhang et al., 2019), LSMFS (Zhang et al., 2021), MLSMFS (Zhang et al., 2021) , PPT-MI (Doquire & Verleysen, 2011),VMFS (Yu et al., 2024a), SRLG-LMA (Dai et al., 2024) and MFS-MFR (Zhou et al., 2024b). Best results are in bold. All methods are evaluated using ML-kNN ($k=10$) on the train/test splits recommended by Mulan. For each method, we report average results over the top-m feature subsets with $m \in \{1, \ldots, 50\}$. The bottom rows of each table show the overall average performance across eight datasets as well as the average rank (Rank.Avg).

Table 4 presents the Hamming Loss across all datasets, where lower values indicate better performance. LIMFS achieves the lowest Hamming Loss on four out of the eight datasets (*emotions*, *enron*, *medical*, and *tmc2007-500*), which demonstrates its effectiveness at reducing label-wise misclassification. On *birds*, MFS-MFR attains the best Hamming Loss, while LIMFS still delivers the second-best value. The remaining datasets where other methods obtain the best Hamming Loss are *scene* (best: SCLS), *genbase* (best: ATR and PPT-MI, tied), and *yeast* (best: SRLG-LMA). As shown in Table 5, LIMFS attains the highest Jaccard score on five datasets (*emotions*, *birds*, *medical*, *yeast*, *tmc2007-500*). SCLS is strongest on *scene*, while MLSMFS leads on *enron* and *genbase*. In terms of Accuracy, LIMFS ranks first on three datasets (*emotions*, *yeast*, and *tmc2007-500*), with MFS-MFR performs best on *birds*. Regarding F1-score, LIMFS remains top-ranked on four datasets (*emotions*, *birds*, *yeast*, and *tmc2007-500*) and maintains competitive performance on *medical* and *genbase*. Higher values for Jaccard score, accuracy, and F1-score indicate superior performance.

Experimental results indicate that although the LIMFS algorithm performs slightly less competitively on several datasets, it consistently outperforms all competing methods in terms of overall effectiveness. This conclusion is strongly supported by two key indicators: average performance and average rank. The datasets where LIMFS shows relatively weaker performance mainly include *genbase*, *scene*, *medical*, and *enron*. Specifically, on the *medical* dataset, both accuracy and F1-score of LIMFS are inferior to those of MLSMFS. On the *genbase* and *enron* datasets, LIMFS underperforms MLSMFS in terms of Jaccard score, accuracy, and F1-score. In contrast, on the *scene* dataset, its performance across all metrics is slightly lower than that of SCLS. Further analysis suggests that these performance gaps primarily stem from intrinsic characteristics of the aforementioned datasets, such as label distribution, sample size, and feature–label association patterns, etc.

genbase, *enron*, and *medical* all suffer from severe label sparsity or label distribution imbalance. Under extreme label sparsity, although the total number of labels in the dataset is relatively large, each instance is associated with only a few labels, resulting in label matrices dominated by zeros. For example, in the *genbase* dataset, different functional labels are largely mutually exclusive or only weakly correlated, and most protein samples are annotated with only one primary functional label. In contrast, label distribution imbalance is reflected in the presence of a large number of 'head' frequent labels and 'long-tail' infrequent labels. For instance, the *medical* dataset exhibits extreme imbalance, where common disease classes correspond to 'head' labels with abundant samples, while many rare diseases are associated with only a handful of instances, forming 'tail' labels. Similarly, in *enron*, common day-to-day

Table 4
Hamming Loss (↓) comparison using ML-kNN.

Dataset	LIMFS	ATR	SCLS	LSMFS	MLSMFS	LRFS	PPT-MI	VMFS	SRLG-LMA	MFS-MFR
emotions	**0.2250 ± 0.0160**	0.2428 ± 0.0145	0.2442 ± 0.0141	0.2608 ± 0.0087	0.2544 ± 0.0107	0.2447 ± 0.0140	0.2476 ± 0.0174	0.2408 ± 0.0199	0.2613 ± 0.0082	0.2607 ± 0.0344
birds	0.0489 ± 0.0010	0.0494 ± 0.0015	0.0508 ± 0.0011	0.0512 ± 0.0015	0.0502 ± 0.0017	0.0493 ± 0.0008	0.0514 ± 0.0015	0.0544 ± 0.0027	0.0523 ± 0.0026	**0.0471 ± 0.0012**
enron	**0.0481 ± 0.0032**	0.0505 ± 0.0020	0.0514 ± 0.0041	0.0507 ± 0.0033	0.0492 ± 0.0026	0.0573 ± 0.0007	0.0522 ± 0.0016	0.0561 ± 0.0021	0.0517 ± 0.0016	0.0540 ± 0.0017
medical	**0.0150 ± 0.0019**	0.0165 ± 0.0021	0.0164 ± 0.0020	0.0154 ± 0.0024	0.0153 ± 0.0020	0.0171 ± 0.0016	0.0164 ± 0.0021	0.0187 ± 0.0031	0.0152 ± 0.0019	0.0171 ± 0.0030
scene	0.1488 ± 0.0069	0.1469 ± 0.0087	**0.1354 ± 0.0127**	0.1489 ± 0.0081	0.1544 ± 0.0058	0.1634 ± 0.0063	0.1686 ± 0.0054	0.1595 ± 0.0094	0.1508 ± 0.0063	0.1461 ± 0.0117
yeast	0.2187 ± 0.0073	0.2207 ± 0.0044	0.2227 ± 0.0047	0.2215 ± 0.0041	0.2218 ± 0.0046	0.2216 ± 0.0039	0.2212 ± 0.0052	0.2282 ± 0.0111	**0.2076 ± 0.0069**	0.2166 ± 0.0091
genbase	0.0149 ± 0.0083	**0.0072 ± 0.0079**	0.0174 ± 0.0063	0.0160 ± 0.0084	0.0166 ± 0.0075	0.0156 ± 0.0077	0.0072 ± 0.0081	0.0143 ± 0.0094	0.0074 ± 0.0074	0.0080 ± 0.0071
tmc2007-500	**0.0713 ± 0.0117**	0.0727 ± 0.0101	0.0719 ± 0.0091	0.0732 ± 0.0079	0.0738 ± 0.0070	0.0728 ± 0.0104	0.0733 ± 0.0105	0.0748 ± 0.0085	0.0742 ± 0.0088	0.0748 ± 0.0093
Average	**0.0988**	0.1008	0.1013	0.1047	0.1045	0.1052	0.1047	0.1059	0.1021	0.1031
Avg. rank	**2.3750**	3.5625	5.3125	6.0000	6.0000	6.6875	6.1250	7.9375	5.6250	5.3750

Table 5
Jaccard Score (↑) comparison using ML-*k*NN.

Dataset	LIMFS	ATR	SCLS	LSMFS	MLSMFS	LRFS	PPT-MI	VMFS	SRLG-LMA	MFS-MFR
emotions	**0.4597 ± 0.0366**	0.4158 ± 0.0432	0.4123 ± 0.0398	0.3846 ± 0.0352	0.3914 ± 0.0379	0.4161 ± 0.0417	0.4113 ± 0.0487	0.4232 ± 0.0416	0.3394 ± 0.0287	0.3561 ± 0.0968
birds	**0.1401 ± 0.0286**	0.0910 ± 0.0211	0.0950 ± 0.0277	0.0899 ± 0.0318	0.0850 ± 0.0271	0.1171 ± 0.0303	0.0748 ± 0.0234	0.0929 ± 0.0341	0.1032 ± 0.0340	0.0930 ± 0.0258
enron	0.2540 ± 0.0404	0.2824 ± 0.0411	0.2743 ± 0.0729	0.2846 ± 0.0619	**0.3095 ± 0.0419**	0.1864 ± 0.0128	0.2472 ± 0.0333	0.2578 ± 0.0417	0.2897 ± 0.0339	0.2406 ± 0.0270
medical	**0.5049 ± 0.0885**	0.4646 ± 0.0798	0.4528 ± 0.0762	0.4969 ± 0.0895	0.5001 ± 0.0761	0.4394 ± 0.0685	0.4586 ± 0.0828	0.4042 ± 0.1198	0.4937 ± 0.0740	0.4237 ± 0.1199
scene	0.3545 ± 0.0708	0.3461 ± 0.0695	**0.3965 ± 0.0882**	0.3359 ± 0.0650	0.3108 ± 0.0544	0.2160 ± 0.0697	0.1708 ± 0.0654	0.3031 ± 0.0522	0.2675 ± 0.0459	0.3050 ± 0.0741
yeast	**0.4355 ± 0.0174**	0.4108 ± 0.0290	0.4052 ± 0.0321	0.4061 ± 0.0328	0.4048 ± 0.0263	0.4066 ± 0.0248	0.4077 ± 0.0288	0.3928 ± 0.0244	0.4111 ± 0.0267	0.3760 ± 0.0430
genbase	0.8813 ± 0.1772	0.8566 ± 0.1725	0.8775 ± 0.1701	0.8369 ± 0.2072	**0.8857 ± 0.1811**	0.8660 ± 0.1765	0.8570 ± 0.1758	0.7226 ± 0.1909	0.8422 ± 0.1633	0.8293 ± 0.1584
tmc2007-500	**0.3920 ± 0.0700**	0.3797 ± 0.0917	0.3892 ± 0.0925	0.3774 ± 0.0840	0.3845 ± 0.0791	0.3722 ± 0.1020	0.3692 ± 0.1003	0.3834 ± 0.0654	0.3885 ± 0.0685	0.3487 ± 0.0895
Average	**0.4278**	0.4059	0.4129	0.4015	0.4090	0.3775	0.3746	0.3725	0.3919	0.3716
Avg. rank	**2.0000**	4.7500	4.2500	5.8750	4.6250	6.1250	7.2500	6.8750	4.8750	8.3750

Table 6
Accuracy Score (↑) comparison using ML-*k*NN.

Dataset	LIMFS	ATR	SCLS	LSMFS	MLSMFS	LRFS	PPT-MI	VMFS	SRLG-LMA	MFS-MFR
emotions	**0.2352 ± 0.0360**	0.2236 ± 0.0328	0.2080 ± 0.0268	0.1739 ± 0.0262	0.1777 ± 0.0260	0.2207 ± 0.0282	0.2178 ± 0.0346	0.2176 ± 0.0319	0.1446 ± 0.0191	0.1797 ± 0.0580
birds	0.4684 ± 0.0114	0.4661 ± 0.0126	0.4553 ± 0.0084	0.4544 ± 0.0125	0.4615 ± 0.0152	0.4620 ± 0.0119	0.4466 ± 0.0175	0.4543 ± 0.0256	0.4536 ± 0.0161	**0.4953 ± 0.0193**
enron	0.0396 ± 0.0275	0.0811 ± 0.0469	0.0631 ± 0.0407	0.0908 ± 0.0415	**0.1060 ± 0.0292**	0.0149 ± 0.0026	0.0230 ± 0.0248	0.0710 ± 0.0265	0.0661 ± 0.0359	0.0132 ± 0.0043
medical	0.5029 ± 0.0949	0.4685 ± 0.0822	0.4527 ± 0.0757	0.5053 ± 0.0945	**0.5122 ± 0.0802**	0.4406 ± 0.0694	0.4591 ± 0.0838	0.4328 ± 0.1327	0.4954 ± 0.0786	0.4237 ± 0.1202
scene	0.3887 ± 0.0834	0.3892 ± 0.0769	**0.4410 ± 0.0941**	0.3816 ± 0.0718	0.3486 ± 0.0585	0.2389 ± 0.0692	0.2007 ± 0.0769	0.3422 ± 0.0562	0.2934 ± 0.0526	0.3176 ± 0.0799
yeast	**0.1458 ± 0.0191**	0.1366 ± 0.0308	0.1332 ± 0.0291	0.1258 ± 0.0248	0.1202 ± 0.0245	0.1320 ± 0.0262	0.1349 ± 0.0289	0.1199 ± 0.0319	0.1406 ± 0.0307	0.1054 ± 0.0400
genbase	0.8743 ± 0.1566	0.8680 ± 0.1487	0.8764 ± 0.1644	0.8216 ± 0.2255	**0.8801 ± 0.1848**	0.8647 ± 0.1646	0.8650 ± 0.1481	0.7366 ± 0.2391	0.8614 ± 0.1466	0.8467 ± 0.1373
tmc2007-500	**0.2335 ± 0.0978**	0.2055 ± 0.0666	0.2081 ± 0.0588	0.2048 ± 0.0539	0.1973 ± 0.0448	0.2010 ± 0.0695	0.1984 ± 0.0700	0.1847 ± 0.0538	0.1879 ± 0.0568	0.1839 ± 0.0601
Average	**0.3610**	0.3548	0.3547	0.3448	0.3504	0.3219	0.3182	0.3199	0.3304	0.3207
Avg. rank	**2.6250**	3.1250	4.3750	5.5000	4.5000	6.2500	6.6250	7.5000	6.6250	7.8750

Table 7
F1 Score (↑) comparison using ML-*k*NN.

Dataset	LIMFS	ATR	SCLS	LSMFS	MLSMFS	LRFS	PPT-MI	VMFS	SRLG-LMA	MFS-MFR
emotions	**0.5852 ± 0.0388**	0.5761 ± 0.0556	0.5714 ± 0.0514	0.5405 ± 0.0453	0.5487 ± 0.0485	0.5755 ± 0.0520	0.5713 ± 0.0582	0.5796 ± 0.0420	0.4680 ± 0.0350	0.4950 ± 0.1245
birds	**0.1974 ± 0.0439**	0.1518 ± 0.0282	0.1582 ± 0.0380	0.1522 ± 0.0498	0.1428 ± 0.0420	0.1922 ± 0.0431	0.1298 ± 0.0347	0.1525 ± 0.0527	0.1599 ± 0.0492	0.1472 ± 0.0388
enron	0.4034 ± 0.0537	0.3765 ± 0.0545	0.3762 ± 0.0903	0.3896 ± 0.0826	**0.4148 ± 0.0642**	0.2675 ± 0.0218	0.3297 ± 0.0408	0.3625 ± 0.0538	0.3814 ± 0.0415	0.3218 ± 0.0303
medical	0.5726 ± 0.1027	0.5547 ± 0.1009	0.5409 ± 0.0958	0.5882 ± 0.1094	**0.5903 ± 0.0926**	0.5287 ± 0.0889	0.5471 ± 0.1040	0.4788 ± 0.1460	0.5787 ± 0.0935	0.4891 ± 0.1386
scene	0.4979 ± 0.1010	0.4915 ± 0.0931	**0.5466 ± 0.1139**	0.4875 ± 0.0888	0.4569 ± 0.0741	0.3220 ± 0.0953	0.2666 ± 0.0983	0.4542 ± 0.0710	0.3781 ± 0.0614	0.4270 ± 0.0992
yeast	**0.5787 ± 0.0182**	0.5331 ± 0.0437	0.5287 ± 0.0492	0.5290 ± 0.0478	0.5247 ± 0.0394	0.5297 ± 0.0386	0.5293 ± 0.0429	0.5110 ± 0.0311	0.5259 ± 0.0360	0.4767 ± 0.0614
genbase	0.8729 ± 0.1965	0.8664 ± 0.1744	0.8825 ± 0.1725	0.8430 ± 0.2067	**0.8853 ± 0.1791**	0.8712 ± 0.1764	0.8659 ± 0.1751	0.7556 ± 0.1886	0.8570 ± 0.1683	0.8431 ± 0.1643
tmc2007-500	**0.5286 ± 0.0734**	0.5036 ± 0.1181	0.5186 ± 0.1206	0.5072 ± 0.1108	0.5172 ± 0.1043	0.4924 ± 0.1312	0.4885 ± 0.1300	0.5213 ± 0.0497	0.5263 ± 0.0720	0.4609 ± 0.1120
Average	**0.5296**	0.5067	0.5154	0.5047	0.5101	0.4724	0.4660	0.4769	0.4844	0.4576
Avg. rank	**1.8750**	4.6250	4.3750	5.3750	4.6250	6.0000	7.3750	6.5000	5.5000	8.7500

topics dominate the label space, whereas sensitive or specialized topics appear rarely. Such label sparsity and distribution imbalance reduce the label co-occurrence probabilities, resulting in simplified, static label dependencies rather than complex, dynamic interactions.

Under these conditions, the IER component of LIMFS, which estimates global label relevance by integrating dynamic interaction weights over all label pairs, becomes less effective. When label interactions are predominantly single-type (e.g., purely enhancing or neutral), dynamic weighting adjustment provides little benefit. For instance, in the *Enron* dataset, certain labels show strong but narrow complementary relationships (e.g., label "C.C1" consistently supplements "A.A1"), while in Medical, features are often domain-specific and directly tied to disease types, leaving little room for cross-label mediation. Consequently, the dynamic weighting mechanism in IER offers no practical advantage. In contrast, MLSMFS directly identifies the most relevant complementary labels for each feature, avoiding neutral or suppressive interactions. This focused mechanism allows MLSMFS to capture key dependencies more effectively, which explains why LIMFS achieves slightly lower feature selection accuracy on these datasets.

The *scene* dataset, which contains 294 features primarily describing low-level image attributes (e.g., color, texture) and covers six labels (Beach, Sunset, Fall Foliage, Field, Mountain, Urban), presents a different challenge. Here, label correlations are static and semantically determined rather than dynamically induced by features. For example, Beach frequently co-occurs with Sunset due to inherent semantic overlap, independent of feature variation. In such cases, the FIS component of LIMFS, designed to capture dynamic label interactions, offers no real utility. Likewise, the IER module's adaptive weighting collapses into uniform weighting, yielding redundant computation without performance gains. By contrast, SCLS, which evaluates features using a simple Rel–Red (relevance vs. redundancy) criterion without modeling label interactions, is well aligned with the static label structure of Scene, resulting in better performance.

In summary, LIMFS is designed for datasets characterized by complex, feature-sensitive, and heterogeneous label interactions. On datasets where label relationships are simple, highly imbalanced, or directly determined by features, LIMFS is prone to information dilution, estimation errors, and noise introduction, which in turn undermines its effectiveness.

To provide deeper insight into the performance dynamics, we examine the Hamming-loss curves of each dataset as a function of the number of selected features m (Figs. 1 and 2). A clear pattern of diminishing marginal gains can be observed: when m is small, adding features yields a sharp decline in Hamming loss; as m increases further, the improvement brought by each additional feature gradually diminishes, causing the curves to plateau or even rebound slightly due to redundant or noisy features. The observation empirically demonstrates the inherent property of the greedy forward-selection scheme, whereby the most informative features are chosen first, making subsequent features naturally less contributive. Even under diminishing marginal gains, the curves of LIMFS remain competitive or leading over a wide range of m values. This indicates the robustness of LIMFS, as the feature subsets it selects are of high quality and it is not overly sensitive to the specific number of features K chosen. Formal statistical comparisons (Friedman + Nemenyi) are reported in Section 5 to assess whether the observed numeric differences are statistically significant across datasets and metrics.

Overall, these experimental results demonstrate the effectiveness of the proposed LIMFS method, particularly its strengths in explicitly modeling complex dynamic label interactions, thus offering improved multi-label classification performance compared to existing state-of-the-art methods.

4.4. Computational complexity and scalability analysis

To address the practical applicability of LIMFS, we analyze its computational complexity and compare it with representative baseline

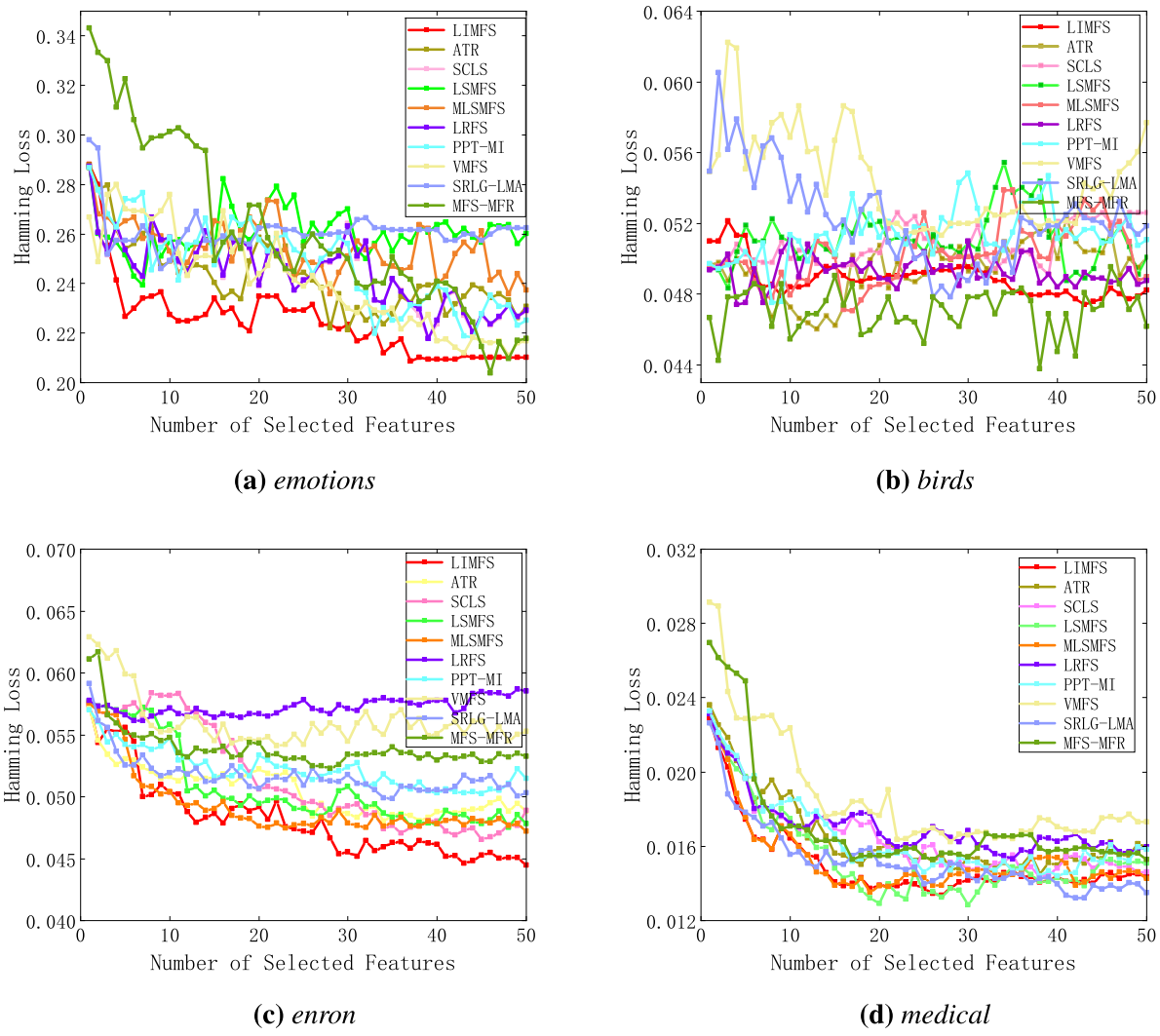


Fig. 1. Hamming Loss comparison on *emotions*, *birds*, *enron*, and *medical*.

methods. This analysis focuses on the time complexity of the feature scoring phase, which dominates the overall cost in filter-based feature selection.

The computational cost of LIMFS primarily stems from estimating information-theoretic measures. Let n denote the number of training instances, $|F|$ the total number of features, $|L|$ the number of labels, and $|S|$ the size of the selected feature subset during the greedy selection process. The key operations and their costs are as follows:

1. **Mutual Information (MI) Estimation:** Computing $I(f;l_i)$ for all feature-label pairs requires $O(|F|\cdot|L|\cdot n)$ operations (assuming histogram-based estimation with fixed bins).
2. **Conditional Mutual Information (CMI) Estimation:** Computing $I(f;l_i\,|\,l_j)$ for all triples (f,l_i,l_j) is the most expensive step, with a cost of $O(|F|\cdot|L|^2\cdot n)$.
3. **FIS Calculation:** Given the precomputed MI and CMI, computing $FIS(f;l_i;l_j)$ for all triples is $O(|F|\cdot|L|^2)$, which is negligible compared to the estimation cost.
4. **IER and LIE Computation:** Aggregating FIS into IER and LIE scores per feature costs $O(|F|\cdot|L|^2)$.
5. **Redundancy Penalty:** During greedy selection, computing $\sum_{f_j\in S} I(f;f_j)$ for each candidate costs $O(|F|\cdot|S|\cdot n)$ per iteration.

Therefore, the overall *dominant* time complexity of LIMFS is $O(|F|\cdot|L|^2\cdot n)$, driven by the CMI estimation across all feature-label-label triplets.

To visually demonstrate its computational cost advantages, we compare the theoretical complexity of LIMFS with several key baselines:

- **SCLS** (Lee & Kim, 2017): It is dominated by MI estimation $I(f;l_i)$ and feature-feature MI $I(f;f_j)$, yielding $O(|F|\cdot(|L|+|S|)\cdot n)$. It avoids pairwise label computations.
- **MLSMFS** (Zhang et al., 2021): it involves label-label and feature-label interactions, leading to roughly $O(|F|\cdot|L|^2\cdot n)$ for supplementary term calculation, similar to LIMFS in label-order complexity.
- **ATR** (Eskandari & Ghassabi, 2023): The ATR algorithm necessitates an initial preprocessing step for calculating the pairwise MI for all feature-feature and feature-label pairs, and the time complexity for this pre-computation is $O(|F|^2\cdot n+|F|\cdot|L|)\cdot n$. The feature selection process itself has complexity $O(|F|^3+|F|^2|L|)$.Therefore, the complete time complexity of the ATR algorithm, including both phases, is $O(|F|^2\cdot n+|F|\cdot|L|\cdot n+|F|^3+|F|^2\cdot|L|)$.

Since SCLS only considers first-order label correlations while LIMFS takes second-order label correlations into account, the time complexity of LIMFS is higher than that of SCLS. However, by incorporating conditional mutual information to account for redundancy between label pairs, LIMFS can better handle correlations among labels, especially when $|L|$ is large and there are strong correlations between labels. Under such conditions, LIMFS typically achieves better performance and is thus able to select more effective features.

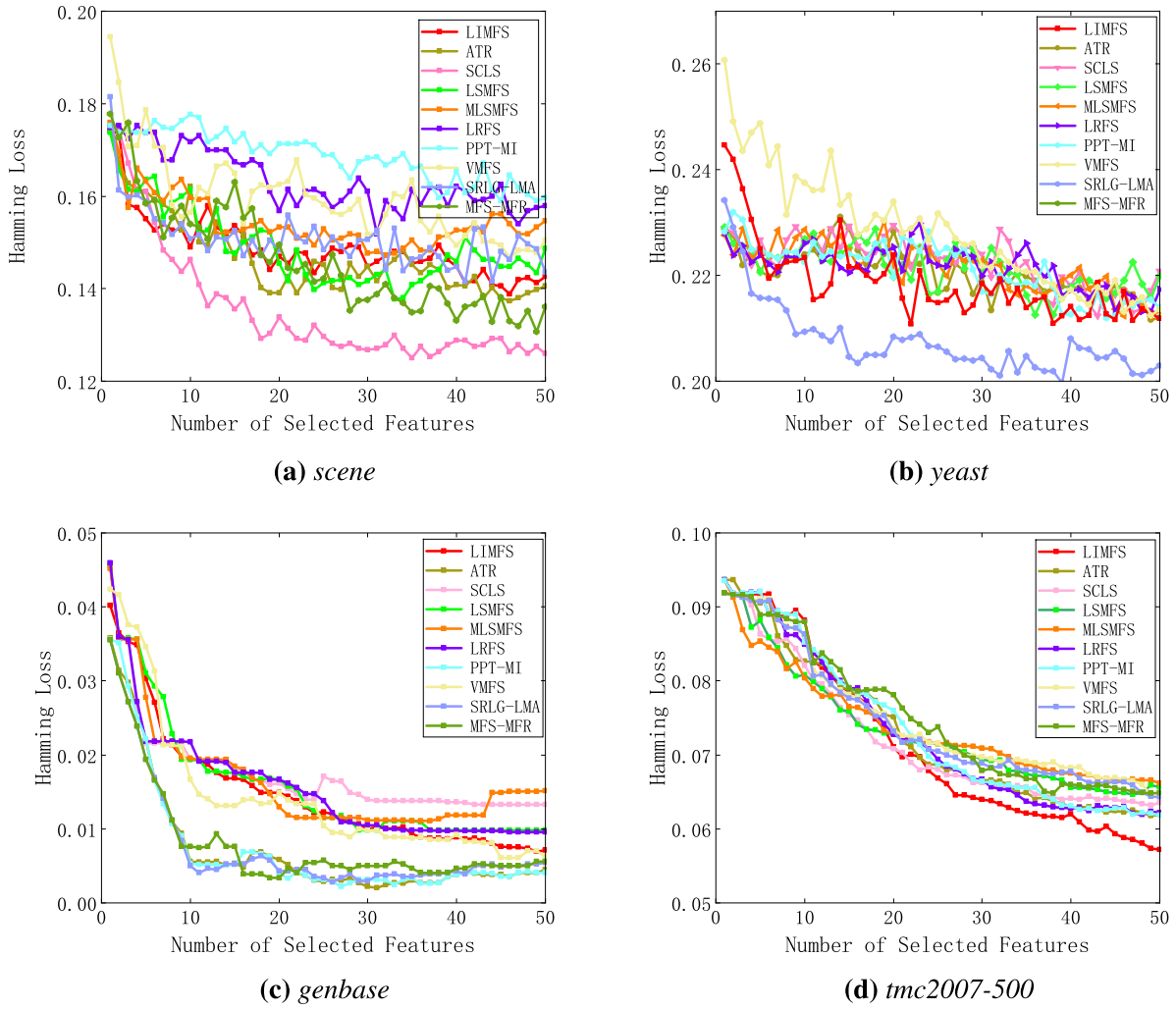


Fig. 2. Hamming Loss comparison on *scene*, *yeast*, *genbase*, and *tmc2007-500*.

Compared to LSMFS and MLSMFS, which have similar time complexities, LIMFS may achieve comparable or better accuracy with a simpler heuristic. The LIMFS directly measures feature relevance to individual labels while penalizing redundancy with already selected features. It avoids the additional label supplementation terms used in LSMFS and MLSMFS, which can sometimes introduce noise.

Regarding sensitivity to the label count $|L|$, ATR scales linearly with $|L|$ during iteration, whereas LIMFS exhibits quadratic scaling with $|L|$ in both its precomputation and iteration phases. Consequently, LIMFS becomes substantially more computationally expensive as the number of labels increases. As for sensitivity to the feature count $|F|$, ATR demonstrates cubic, quadratic, and linear dependencies on $|F|$, while LIMFS shows only linear dependence in the corresponding terms. Therefore, when $|L|$ is moderate, LIMFS scales better with increasing feature count. Furthermore, despite its higher theoretical complexity, LIMFS achieves better ranking in empirical evaluations on benchmark datasets.

In a word, the $|L|^2$ factor implies that LIMFS is best suited for problems with a moderate number of labels. For *extreme multi-label learning* (XMLL) scenarios with thousands of labels, the direct computation of all pairwise label CMI becomes prohibitive. Future work will focus on scalable approximations, such as leveraging label hierarchy or clustering to reduce effective $|L|$.

Despite its quadratic growth in complexity with respect to the number of labels, the proposed dynamic interaction modeling provides significant performance improvements on datasets where label correlations are complex and feature-sensitive, as demonstrated in Sections 4.3 and 5. For many practical multi-label tasks with a moderate label space, the additional computational cost incurred by LIMFS is justified by its improved selection accuracy.

4.5. Ablation study

To systematically evaluate the contributions of FIS, IER, and LIE components, we designed three ablated variants of LIMFS:

- **LIMFS-noFIS**: FIS is replaced by a static label correlation matrix. This variant uses the same IER and LIE but with static weights. Specifically, the FIS-induced weighting factor $\phi(f;l_i)$ in $\mathrm{IER}(f) = \sum_{l_i \in L} I(f;l_i) \cdot \phi(f;l_i))$ is fixed to 1 for all labels, so that IER is computed as the sum of mutual information $I(f;l_i)$ without interaction weighting.
- **LIMFS-noIER**: The IER component is removed. Only LIE and the redundancy suppression term are retained.
- **LIMFS-noLIE**: The LIE component is removed, leaving only IER (with dynamic weights) and the redundancy supression term.

All variants, along with the full LIMFS model, were evaluated on the same four benchmark datasets using identical experimental protocols (ML-kNN classifier with $k = 10$, top-m features with $m \in \{1, \ldots, 50\}$, and the four evaluation metrics). The detailed ablation study results on the *emotions*, *birds*, *scene*, and *yeast* datasets are provided in Tables 8–11, respectively. They report the average performance across four datasets for each metric. Higher values are better for JS, AS, F1; Lower is better

Table 8
Ablation analysis of LIMFS on the *emotions* dataset.

Variant	Hamming Loss (HL) ↓	Jaccard Score (JS) ↑	Accuracy (AS) ↑	F1 Score ↑
LIMFS (full)	**0.2250**	**0.4597**	**0.2352**	**0.5852**
LIMFS-noFIS	0.2608	0.3846	0.1739	0.5405
LIMFS-noIER	0.2354	0.4329	0.2172	0.5344
LIMFS-noLIE	0.2393	0.4227	0.1981	0.5132

Table 9
Ablation analysis of LIMFS on the *birds* dataset.

Variant	Hamming Loss (HL) ↓	Jaccard Score (JS) ↑	Accuracy (AS) ↑	F1 Score ↑
LIMFS (full)	**0.0489**	**0.1401**	**0.4684**	**0.1974**
LIMFS-noFIS	0.0512	0.0899	0.4544	0.1522
LIMFS-noIER	0.0535	0.1049	0.4678	0.1890
LIMFS-noLIE	0.0532	0.1056	0.4666	0.1904

Table 10
Ablation analysis of LIMFS on the *scene* dataset.

Variant	Hamming Loss (HL) ↓	Jaccard Score (JS) ↑	Accuracy (AS) ↑	F1 Score ↑
LIMFS (full)	**0.1488**	**0.3545**	**0.3887**	**0.4979**
LIMFS-noFIS	0.1489	0.3359	0.3816	0.4875
LIMFS-noIER	0.1833	0.1563	0.1766	0.1965
LIMFS-noLIE	0.1832	0.1548	0.1747	0.1946

Table 11
Ablation analysis of LIMFS on the *yeast* dataset.

Variant	Hamming Loss (HL) ↓	Jaccard Score (JS) ↑	Accuracy (AS) ↑	F1 Score ↑
LIMFS (full)	**0.2187**	**0.4355**	**0.1458**	**0.5787**
LIMFS-noFIS	0.2215	0.4061	0.1258	0.5290
LIMFS-noIER	0.2303	0.4069	0.1164	0.5506
LIMFS-noLIE	0.2306	0.4039	0.1133	0.5473

for HL. Best results are in **bold**. Standard deviations are omitted for clarity.

Tables 8 through 11 demonstrate that three components are all essential for LIMFS. FIS is crucial for dynamic interaction modeling. Removing FIS (LIMFS-noFIS) leads to the most significant performance drop across all metrics, which confirms that static label correlation assumptions are inadequate for capturing context-sensitive dependencies. IER provides global interaction-aware relevance adjustment. Without IER (LIMFS-noIER), performance deteriorates noticeably, which demonstrates that adaptive weighting based on FIS is essential for accurate global feature relevance estimation. LIE captures valuable local enhancement gains. Removing LIE (LIMFS-noLIE) results in a moderate but consistent performance decline, which validates that local positive information gain complements global relevance assessment. When the FIS, IER, and LIE work synergistically, the full LIMFS model consistently outperforms all ablated variants.

4.6. Visualization of feature-conditioned interaction strength

To intuitively demonstrate the role of label interactions, this subsection presents FIS matrices for several representative features on the *emotions* dataset. For each selected feature f, we construct a $q \times q$ heatmap where rows correspond to target labels l_i and columns correspond to conditioning labels l_j. The cell at position (i, j) encodes the directed interaction strength $\mathrm{FIS}(f; l_i; l_j) \in [-1, 1]$ defined in Section 3.2.1, which quantifies how the presence of label l_j influences the relevance of feature f to label l_i. Diagonal entries (i, i) are fixed to zero, reflecting the fact that a label is not conditioned on itself.

Fig. 3 illustrates FIS heatmaps for three features randomly selected from the top-ranked subset obtained by LIMFS on the *emotions* dataset ($q = 6$). A common color scale is used across all subplots, where blue (FIS < 0) indicates an enhancing interactions and red (FIS > 0) indicates a suppressive interactions; white (FIS ≈ 0) corresponds to a neutral interactions. Numeric annotations are overlaid in each cell to improve readability.

As shown in the Fig. 3, the three features yield markedly different FIS patterns despite sharing the same label set. Notably, some label pairs are strongly enhanced (dark blue) for one feature but mildly suppressed (light red) for another, while other pairs remain almost neutral across all features. This evidence indicates that label interactions in LIMFS are not treated as a fixed global property but are dynamically shaped by each candidate feature. Consequently, the visualization concretely illustrates the feature-conditioned nature of label interactions, offering intuitive support that complements the quantitative experiments.

5. Statistical analysis

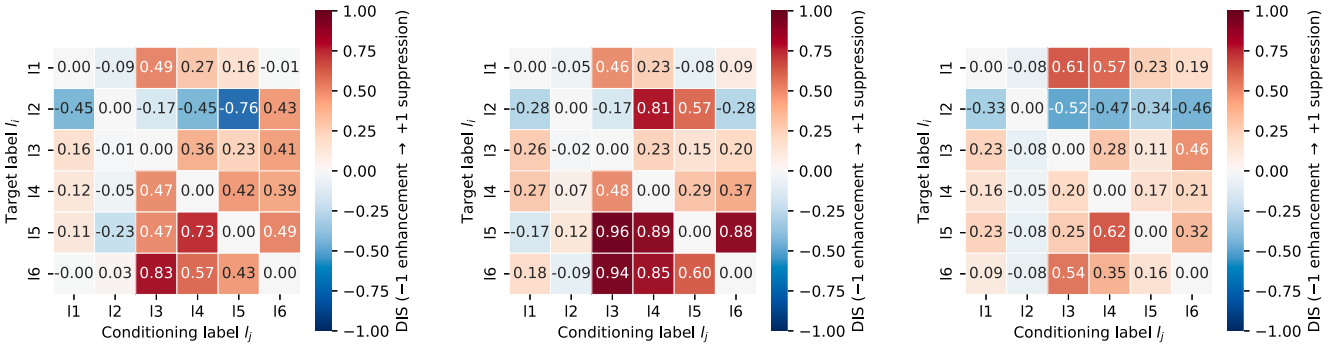
We adopted the standard nonparametric protocol for comparing different methods across multiple datasets (Demšar, 2006). For each dataset, the ten competing methods ($k = 10$) were ranked (with rank 1 indicating the best performance; tied ranks were assigned the average value). Given eight datasets ($N = 8$), we first conducted the Friedman test with the Iman-Davenport correction. If the test result was significant, we further applied the Nemenyi post-hoc test and visualized the results using critical-difference (CD) diagrams.

5.1. Overall test

Let R_j denote the average rank of the j-th method . The Friedman statistic (Friedman, 1940) and the Iman-Davenport F statistic are calculated as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{j=1}^{k} R_j^2\right) - 3N(k+1), \qquad F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}.$$

The Iman-Davenport F statistic F_F follows an F-distribution with $(k-1)$ degrees of freedom in the numerator and $(k-1)(N-1)$ degrees



(a) Feature $f_{(1)}$ **(b)** Feature $f_{(2)}$ **(c)** Feature $f_{(3)}$

Fig. 3. FIS heatmaps of three representative features on the *emotions* dataset. Each heatmap is a 6×6 matrix with target labels l_i on the rows and conditioning labels l_j on the columns. Colors encode $\mathrm{FIS}(f; l_i; l_j) \in [-1, 1]$, where blue indicates enhancement (negative values) and red indicates suppression (positive values); diagonal entries are set to zero.

Table 12
Friedman and Iman-Davenport omnibus tests ($k = 10$, $N = 8$).

Metric	χ_F^2	p-value (χ^2)	F_F	p-value (F)
HL	19.1079	0.0243	2.5288	0.0152
JS	25.5545	0.0024	3.8514	0.00063
F1	27.3000	0.0012	4.2752	0.00023
AS	25.2273	0.0027	3.7755	0.00075

of freedom in the denominator. For our experimental setting with $k = 10$ methods and $N = 8$ datasets, F_F therefore follows $F_{9,63}$.

Table 12 reports the results of the omnibus tests (using the Iman-Davenport correction) for the four evaluation metrics: Hamming Loss (HL), Jaccard Score (JS), F1 Score (F1), and Accuracy Score (AS). The p-values associated with F_F for all four metrics are well below the significance level of $\alpha = 0.05$. This allows us to confidently reject the null hypothesis that all ten methods perform equivalently across the eight datasets.

5.2. Average ranks

Table 13 lists the average ranks of the ten methods across the eight datasets. Lower values indicate better performance. Notably, LIMFS achieved the best average rank for every evaluation metric.

5.3. Post-hoc comparisons

When the Friedman test yielded a significant result, the Nemenyi test was used to determine whether two methods differed significantly. Specifically, two methods were considered significantly different if the gap between their average ranks exceeded the critical difference (CD) value, which is calculated as:

$$\mathrm{CD} = q_{0.05} \sqrt{\frac{k(k+1)}{6N}}$$

where $k = 10$ (number of methods) and $N = 8$ (number of datasets), and the critical value $q_{0.05} = 3.103$ for a significance level of $\alpha = 0.05$ (Demšar, 2006). Substituting these values into the formula, the CD value is approximately 4.72. The results of pairwise comparisons for each evaluation metric are as follows:

- **Hamming Loss (HL):** Only the rank between LIMFS and VMFS exceeds the CD (≈ 4.72), indicating that LIMFS is significantly better than VMFS; no other pair involving LIMFS reaches significance, indicating no statistically significant performance differences detected under this metric.

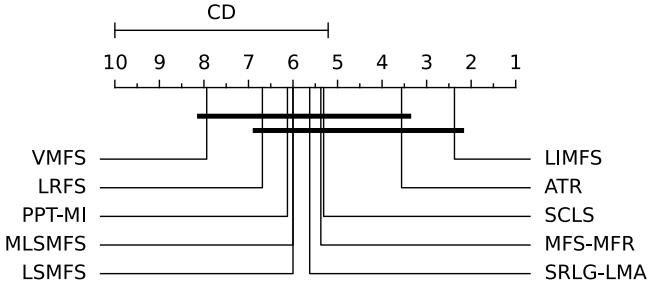


Fig. 4. Critical-difference (CD) diagram for Hamming Loss.

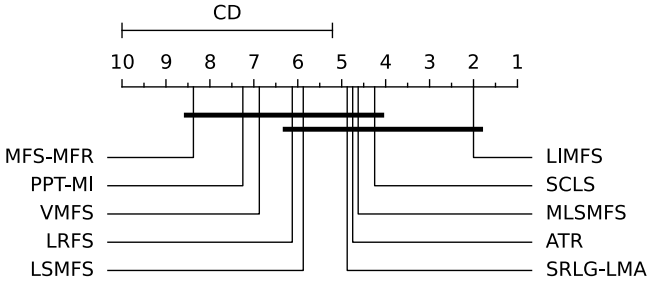


Fig. 5. Critical-difference (CD) diagram for Jaccard Score.

- **Jaccard Score (JS):** LIMFS significantly outperforms PPT-MI, VMFS, and MFS-MFR; other pairwise gaps with LIMFS do not exceed the CD.
- **Accuracy Score (AS):** Only the rank differences between LIMFS and MFS-MFR, as well as between LIMFS and VMFS, exceed the CD (≈ 4.72), indicating that LIMFS significantly outperforms both MFS-MFR and VMFS on this metrics. No significant differences are observed between LIMFS and the other methods.
- **F1 Score (F1):** LIMFS significantly outperforms PPT-MI and MFS-MFR, indicating that LIMFS significantly outperforms PPT-MI and MFS-MFR in terms of F1; other pairwise gaps with LIMFS do not exceed the CD.

Figs. 4–7 respectively present the CD diagrams for the HL, JS, F1, and AS metrics, which visually illustrate the ranking of each method and the significance of differences. LIMFS is positioned on the far right in all CD diagrams (indicating the optimal ranking) and shows no significant disadvantage compared to any competing method, further verifying the stability of its performance.

Table 13
Average ranks across 8 datasets (lower is better). Best in bold.

Metric	LIMFS	ATR	SCLS	LSMFS	MLSMFS	LRFS	PPT-MI	VMFS	SRLG-LMA	MFS-MFR
HL	**2.3750**	3.5625	5.3125	6.0000	6.0000	6.6875	6.1250	7.9375	5.6250	5.3750
JS	**2.0000**	4.7500	4.2500	5.8750	4.6250	6.1250	7.2500	6.8750	4.8750	8.3750
AS	**2.6250**	3.1250	4.3750	5.5000	4.5000	6.2500	6.6250	7.5000	6.6250	7.8750
F1	**1.8750**	4.6250	4.3750	5.3750	4.6250	6.0000	7.3750	6.5000	5.5000	8.7500

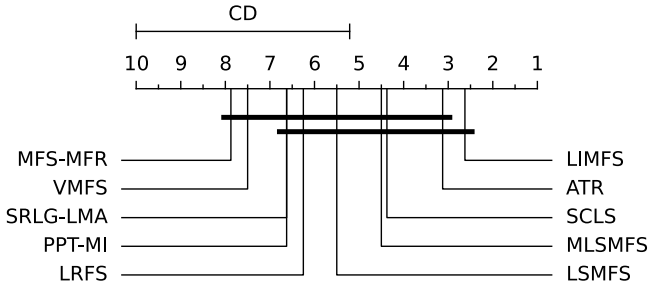


Fig. 6. Critical-difference (CD) diagram for Accuracy Score.

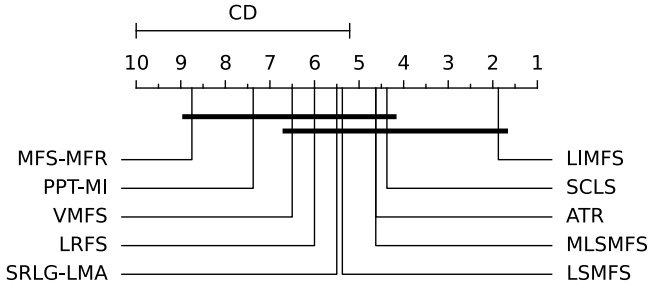


Fig. 7. Critical-difference (CD) diagram for F1 Score.

6. Conclusion

This paper proposes LIMFS (Label Interaction-aware Multi-label Feature Selection), a filter-style multi-label feature selection framework that explicitly models dynamic, feature-conditioned label interactions and integrates such interactions into the feature evaluation, addressing the limitation of traditional methods that often overlook the dependency between labels when selecting features. The core contributions of this work are threefold:

- We introduce the Feature-conditioned Interaction Strength (FIS), a bounded and directional metric that quantifies how the dependency between a pair of labels changes when conditioned on a candidate feature.
- We propose two complementary relevance measures: Interaction-Enhanced Relevance (IER), which adaptively adjusts feature relevance using interaction-aware weights, and Label Interaction Enhancement (LIE), which captures local positive information gains resulting from synergistic label interactions.
- We design a unified greedy selection algorithm that integrates IER, LIE, and a redundancy suppression mechanism to efficiently generate compact and discriminative feature subsets for multi-label classification.

Empirical results on eight multi-label benchmarks demonstrate that LIMFS consistently outperforms or matches several state-of-the-art multi-label feature selection methods across multiple evaluation metrics. These findings indicate that explicitly accounting for feature-conditioned label interactions can substantially improve filter-based multi-label feature selection.

We also note several limitations that motivate future work. First, the method relies on accurate estimates of mutual information and conditional mutual information, which can be sensitive to the choice of estimators and discretization strategies for continuous features. Second, the computational cost of the greedy selection procedure increases with the number of features and labels, which may limit scalability in very high-dimensional or extreme multi-label scenarios. Third, although our approach models how label interactions vary with different features, it does not update these interactions dynamically as the feature subset expands.

For future work, we plan to focus on the following three directions:

- **Richer interaction models:** Extend FIS to graph-based or representation-based models, such as graph neural networks or conditional graphical models, to capture higher-order and structured label dependencies.
- **Scalability:** Design scalable estimators and approximation techniques, such as sampling-based MI estimators, feature/label clustering or sketching, to enable LIMFS for large-scale and extreme multi-label problems.
- **Dynamic interaction updating:** Investigate mechanisms to dynamically update label interaction estimates as the feature subset evolves during selection, moving toward an adaptive and iterative interaction modeling framework.

CRediT authorship contribution statement

Ying Yu: Conceptualization, Methodology, Formal analysis, Project administration; **Bowen Li:** Data curation, Writing – original draft, Investigation; **Jin Qian:** Validation, Visualization; **Wenhao Shu:** Writing – review & editing; **Duoqian Miao:** Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 62462033, 62163016, 62466017), the Natural Science Foundation of Jiangxi Province (Nos. 20242BAB25092, 20232ACB202013), the Open Project of the State Key Laboratory (No. HJGZ2023203), and the Ganpo Talent Support Plan (No. 20242BCE50030).

References

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory. (2nd ed.). Wiley.
Dai, J., Huang, W., Zhang, C., & Liu, J. (2024). Multi-label feature selection by strongly relevant label gain and label mutual aid. *Pattern Recognition*, *145*, 109945.
Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.
Dong, H., Sun, J., Sun, X., & Ding, R. (2020). A many-objective feature selection for multi-label classification. *Knowledge-Based Systems*, *208*, 106456.

Doquire, G., & Verleysen, M. (2011). Feature selection for multi-label classification problems. In J. Cabestany, I. Rojas, & G. Joya (Eds.), *Advances in computational intelligence* (pp. 9–16). Berlin, Heidelberg: Springer (*vol. 6691*). https://doi.org/10.1007/978-3-642-21501-8_2

Eskandari, S., & Ghassabi, S. (2023). Multi-label feature selection using adaptive and transformed relevance. arXiv preprint arXiv:2309.14768

Feng, C.M., Yu, K., Xu, X., Khan, S., Goh, R.S.M., Zuo, W., & Liu, Y. (2025). Text to image for multi-label image recognition with joint prompt-adapter learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(9), 7660–7674.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, *11*(1), 86–92.

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

Hancer, E., Xue, B., & Zhang, M. (2024). A multimodal multiobjective evolutionary algorithm for filter feature selection in multilabel classification. *IEEE Transactions on Artificial Intelligence*, *5*(9), 4428–4442.

Hao, P., Gao, W., & Hu, L. (2025). Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, *157*, 110888.

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted f1-scores. *Applied Sciences*, *14*(21), 9863–9863.

Lee, J., & Kim, D. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, *66*, 342–352.

Lee, J., & Kim, D. W. (2016). Efficient multi-label feature selection using entropy-based label selection. *Entropy*, *18*(11), 405.

Lee, J. S., & Kim, D. W. (2015). Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, *42*(4), 2013–2025.

Li, L., Liu, H., Ma, Z., Mo, Y., Duan, Z., Zhou, J., & Zhao, J. (2014). Multi-label feature selection via information gain. In *Proceedings of the 10th international conference on advanced data mining and applications (ADMA)* (pp. 345–355).

Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*, *53*(3), 551–577.

Lin, Y., Hu, Q., Liu, J., & Duan, J. (2015). Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, *168*, 92–103.

Liu, W., Wang, H., Shen, X., & Tsang, I. W. (2022). The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7955–7974.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research : JMLR*, *12*, 2825–2830.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. C. (2018). Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, *49*(1), 57–78.

Qian, W., Huang, J., Xu, F., Shu, W., & Ding, W. (2023). A survey on multi-label feature selection from perspectives of label fusion. *Information Fusion*, *100*, 101948.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Singh, I. P., Ghorbel, E., Oyedotun, O., & Aouada, D. (2024). Multi-label image classification using adaptive graph convolutional networks: From a single domain to multiple domains. *Computer Vision and Image Understanding : CVIU*, *247*, 104062.

Soykök, I.T., & Güvenir, H. A. (2025). Multi-label multi-modal classification of movie scenes. *Knowledge-Based System*, *318*, 113459.

Spolaôr, N., Monard, M. C., Tsoumakas, G., & Lee, H. D. (2016). A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*, *180*, 3–15.

Tsoumakas, G., Xioufis, E. S., Vilcek, J., & Vlahavas, I. P. (2011). MULAN: A java library for multi-label learning. *Journal of Machine Learning Research*, *12*, 2411–2414.

Wang, W., Li, Y., Liu, J., & Zhou, H. (2024). A filter-based improved multi-objective equilibrium optimizer for single-label and multi-label feature selection problem. *International Journal of Computational Intelligence and Applications*, *23*(1), 2350028.

Yu, Y., Wan, M., Qian, J., Miao, D., Zhang, Z., & Zhao, P. (2024a). Feature selection for multi-label learning based on variable-degree multi-granulation decision-theoretic rough sets. *International Journal of Approximate Reasoning*, *169*, 109181.

Zhang, N., Wang, A., Lu, P., Feng, T., Xu, Y., & Du, G. (2025a). Multi-label feature selection with feature reconstruction and label correlations. *Expert Systems with Applications*, *285*, 127993.

Zhang, P., & Gao, W. (2021). Feature relevance term variation for multi-label feature selection. *Applied Intelligence*, *51*(7), 5095–5110.

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

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

Zhang, P., Sheng, J., Gao, W., Hu, J., & Li, Y. (2022). Multi-label feature selection method based on dynamic weight. *Soft Computing*, *26*(6), 2793–2805.

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

Zheng, L., Shi, S., Lu, M., Fang, P., Pan, Z., Zhang, H., Zhou, Z., Zhang, H., Mou, M., Huang, S., Tao, L., Xia, W., Li, H., Zeng, Z., Zhang, S., Chen, Y., Li, Z., & Zhu, F. (2024). AnnoPRO: A strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biology*, *25*, 41.

Zhou, C., Dong, J., Huang, X., Liu, Z., Zhou, K., & Xu, Z. (2024a). Quest: Efficient extreme multi-label text classification with large language models on commodity hardware. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 3929–3940).

Zhou, G., Li, R., Shang, Z., Li, X., & Jia, L. (2024b). Multi-label feature selection based on minimizing feature redundancy of mutual information. *Neurocomputing*, *607*, 128392. https://doi.org/10.1016/j.neucom.2024.128392

Zou, Y., Hu, X., & Li, P. (2024). Gradient-based multi-label feature selection considering three-way variable interaction. *Pattern Recognition*, *145*, 109900.