Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085488

ARTICLE

Partial Multi-Label Learning with Missing Labels via Feature-Aware Label Disentanglement

Yuzhi Tao[1] **and Anhui Tan**[2,*]

[1]College of Business Administration, Huaqiao University, Quanzhou, China
[2]School of Mathematical Sciences, Huaqiao University, Quanzhou, China

*Corresponding Author: Anhui Tan. Email: 20100@hqu.edu.cn

Received: 12 May 2026; Accepted: 08 July 2026; Published: 13 August 2026

ABSTRACT: Partial multi-label learning addresses scenarios where each instance is associated with a set of candidate labels that include both relevant and irrelevant ones. In practical scenarios, such label sets are often simultaneously incomplete and noisy, which severely hampers the ability of models to extract compact and discriminative features. To address these issues, we propose an integrated learning paradigm that simultaneously enhances feature compactness and improves robustness against label noise. Our method learns an adaptive fuzzy neighborhood graph to capture the intrinsic relationships among instances. The resulting graph enables reliable label propagation, which effectively rectifies incorrect annotations and infers missing labels. In addition, we introduce a feature disentanglement mechanism that isolates reliable label-related feature representations from spurious ones introduced by noisy supervision. By integrating feature learning and label refinement into a joint optimization process, the proposed approach achieves a synergistic improvement in both representation quality and label reliability. Extensive theoretical analysis and empirical studies on multiple benchmark datasets demonstrate that our framework consistently outperforms state-of-the-art methods in terms of accuracy, stability, and robustness to annotation noise.

KEYWORDS: Multi-label learning; missing labels; noisy labels; label correlation; label-specific features

1 Introduction

Multi-label learning (MLL) [1] is a fundamental learning paradigm that aims to assign multiple semantic labels to each instance, which enables models to capture the complex correlations among categories and has been widely applied in various real-world scenarios such as image recognition, natural language processing, and recommendation systems. Traditional MLL algorithms usually assume that all training samples are annotated with complete and noise-free label sets, which provides ideal supervision for model training. However, this ideal assumption rarely holds in realistic scenarios, where label annotations are often noisy or incomplete due to ambiguous visual content, low-quality data, inconsistent human judgments, or the high cost of manual annotation. The presence of such imperfect annotations can significantly degrade the performance of conventional MLL models, as they tend to overfit to unreliable supervision signals and fail to generalize to unseen data.

To mitigate this problem, partial multi-label learning (PMLL) [2,3] has emerged as a practical paradigm for weakly supervised learning, which relaxes the strict requirement of complete and clean labels. In the PMLL setting, each instance is accompanied by a set of candidate labels that includes both relevant (true) and irrelevant (false-positive) ones, while some true labels may also be missing from the candidate set

(incomplete annotations) [4]. The core learning task is to identify the true labels from these candidates while simultaneously learning a robust predictive model that can generalize well. Although this framework alleviates the dependency on exhaustive and high-cost manual labeling, it introduces new challenges, specifically how to effectively recover missing labels and suppress the adverse effects of false-positive annotations, especially when the noise ratio in candidate labels is high.

A straightforward strategy for PMLL treats all candidate labels as correct and directly applies standard MLL techniques (e.g., ML-KNN [5] and BP-MLL [6]). Such an approach may yield acceptable results when noisy labels are scarce, but it quickly deteriorates when noise becomes dominant, as the model will inevitably learn spurious correlations from false-positive annotations. To address this limitation, recent studies have reformulated PMLL as a latent label inference problem, iteratively refining the candidate label set through knowledge representation [7], graph-based modeling [8], or self-supervised learning [9]. These methods have demonstrated notable improvements by leveraging structural dependencies among instances or features to propagate label information and reduce the uncertainty of candidate labels.

In real-world learning scenarios, obtaining a completely labeled dataset is often infeasible because manual annotation demands extensive human effort, professional domain expertise and significant time costs. Annotators may inadvertently neglect certain relevant categories or misassign irrelevant ones, resulting in label sets that are both incomplete and corrupted by noise. Such imperfect supervision has become increasingly common across diverse domains. In medical image analysis, for example, clearly discernible pathological patterns allow clinicians to make confident diagnoses, but when the manifestations are subtle or ambiguous, establishing a clear diagnosis becomes difficult, leading to uncertain or missing annotations that may require further expert validation [10]. Similar issues occur in social media content tagging, where the subjectivity of annotators and the diversity of content semantics easily lead to imperfect label annotations [11].

These practical constraints have motivated growing interest in learning paradigms that integrate the advantages of PMLL and MLL with incomplete supervision, aiming to enhance both model robustness to label noise and generalization to unseen data. The goal of such approaches is to construct models that can not only tolerate noisy and incomplete annotations but also extract discriminative features to support accurate label prediction. A central challenge in this field is that label noise is seldom purely random; instead, errors are often correlated with specific regions of the feature space, arising from ambiguous patterns, overlapping semantic categories, low-quality input data, or annotator subjectivity [12]. However, as pointed out in [13], noisy labels often arise from ambiguous content in the examples, and there exist relationships between these noisy labels and the corresponding feature representations. Moreover, existing approaches often neglect the simultaneous presence of missing and incorrect annotations, as well as the critical task of extracting reliable and informative features that can support both label refinement and prediction. Chen et al. [14] proposed a universal approach for unifying the handling of label noise and incompleteness in multi-label scenarios, where the compatibility with incomplete labels is achieved through the generalization of noise-handling mechanisms. However, the model adopts a deterministic feature extraction module that lacks the ability to quantify the uncertainty of extracted features, making it difficult to distinguish between reliable informative features and noisy irrelevant ones. This limitation hinders the model's performance when facing high noise ratios or ambiguous annotations.

To address the challenges of PMLL with missing labels, we propose a unified learning framework, termed WPML, that simultaneously enhances feature compactness and robustness against label noise. At the core of our framework is Feature-Aware Label Disentanglement, which explicitly separates reliable label-specific representations from misleading signals introduced by noisy or incomplete annotations. This ensures that the model focuses on informative, label-relevant features while mitigating the impact of corrupted supervision. By integrating feature learning and label refinement within a single optimization process, the

framework achieves a synergistic improvement in both representation quality and label reliability, forming a cohesive and interpretable methodology rather than a simple combination of techniques. Based on the above analysis, this study aims to address the following research questions:

- **RQ1:** How can missing labels be effectively recovered from incomplete multi-label annotations?
- **RQ2:** How can noisy labels be identified and separated from reliable label information?
- **RQ3:** How can label correlations be effectively exploited to improve label recovery and prediction?
- **RQ4:** How can label-specific features be learned to capture the discriminative information associated with individual labels?

To answer these research questions, we propose a unified partial multi-label learning framework that jointly performs reliable label reconstruction, noise separation, label correlation modeling, and label-specific feature learning. The main contributions of this work are summarized as follows:

- We propose a unified feature-label disentanglement framework for partial multi-label learning with both missing and noisy annotations. Unlike methods that treat label completion and noise identification as separate procedures, the proposed model jointly recovers reliable labels and separates sparse annotation noise within a single optimization framework.
- We develop a coupled label reconstruction mechanism that simultaneously integrates fuzzy neighborhood consistency and feature-induced prediction. Its novelty lies not in conventional graph regularization itself, but in using the neighborhood structure and instance features to jointly constrain the same reliable label matrix, thereby reducing error propagation from corrupted annotations.
- We introduce an orthogonality-constrained label embedding to connect shared feature projection with label-specific prediction. The orthogonal semantic factors reduce redundant label information, while each column of the embedding induces an effective label-specific predictor. This enables shared and label-specific feature information to be learned jointly rather than independently.
- Extensive experiments under different missing-label and noisy-label settings demonstrate that the proposed coupling mechanism consistently improves prediction performance.

Although graph regularization, sparse noise modeling, and feature projection have been individually investigated in previous studies, their effective integration for simultaneously handling missing and noisy labels remains insufficient. Existing methods such as PML-NI [13], DM2L [15], and Glocal [16] mainly focus on particular aspects of partial multi-label learning, whereas the interaction among reliable label recovery, noise separation, label structure, and label-specific feature learning is not explicitly modeled.

The key novelty of our method lies in a unified feature–label disentanglement mechanism rather than in the isolated use of these conventional components. Specifically, the observed annotation matrix is jointly decomposed into reliable label information and sparse annotation noise, while the recovered labels are simultaneously constrained by fuzzy neighborhood structures and instance features. Moreover, an orthogonal label embedding is introduced to learn complementary latent semantic factors, through which the shared feature projection is transformed into label-specific predictors. Consequently, label recovery, noise identification, structural correlation modeling, and label-specific feature learning mutually reinforce each other within a single optimization framework.

The remainder of this paper is organized as follows. Section 2 reviews related work on PMLL, label noise handling, and feature disentanglement. Sections 3 and 4 present the proposed framework in detail, including the similarity graph construction, feature disentanglement mechanism, and unified optimization process. Section 5 reports experimental results and comprehensive performance analysis, comparing our method with state-of-the-art approaches. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

This section reviews prior research relevant to our work, with clear distinctions among three typical weak supervision settings: (1) Partial Multi-Label Learning (PMLL), where each instance is assigned a set of candidate labels containing both ground-truth and false noisy labels; (2) Multi-Label Learning with Missing Labels (MLL-ML), where only a subset of valid labels is observed and the rest are unknown/missing rather than noisy; (3) Multi-Label Learning with simultaneous noisy and missing labels, the realistic but under-explored scenario that our work targets.

2.1 Partial Multi-Label Learning (PMLL)

Partial multi-label learning (PMLL) addresses the setting where each instance is associated with a set of candidate labels that include both correct labels and irrelevant noisy labels [17]. It differs from standard multi-label learning (MLL) [18] and partial label learning: the goal is not recovering missing values, but disambiguating true labels from a superset of noisy candidates.

A major research line focuses on label disambiguation via iterative confidence estimation. PARTICLE [3] estimates label confidence using neighborhood information. Sun et al. [19] enhance label reliability via fuzzy similarity. PML-MT [20] adopts mutual teaching and dual self-ensembling for collaborative label refinement. Xie and Huang [21] extend PMLL to semi-supervised scenarios. Li et al. [22] perform calibrated label disambiguation by combining probabilistic modeling and iterative purification. These methods aim to identify and suppress false labels within the candidate set. Another branch focuses on joint feature-label subspace modeling to capture high-order semantic dependencies. Wang et al. [23] jointly learn latent feature and label spaces. Zhong et al. [24] design a noise-aware mechanism to exploit both reliable and unreliable label signals. Wang et al. [25] use semantic embedding and label co-occurrence to strengthen feature-label alignment.

Recent advances further exploit label confidence and feature disentanglement. Han et al. [26] use label confidence to guide robust feature selection. Hang and Zhang [27] perform label-specific feature correction to reduce annotation noise interference. Jalali and Kasneci [28] focus on instance-adaptive expert selection for multi-label classification. Methods addressing noisy or incomplete annotations are also closely related. Yang et al. [29] remove noisy labels before training. Qian et al. [30] propose a noise-tolerant broad learning system. Li et al. [31] generate geometrically consistent pseudo-labels for uncertain annotations. These works highlight the importance of confidence-aware learning and noise robustness in PMLL.

2.2 Multi-Label Learning with Missing Labels (MLL-ML)

Different from PMLL, MLL-ML assumes that only a subset of true labels is observed, while unobserved labels are entirely unknown. The core task is label recovery or prediction under incomplete annotation.

A classic paradigm is based on low-rank matrix completion, exploiting inter-label correlation. Bucak et al. [32] and Yu et al. [33] formulate label inference as a matrix recovery problem via joint feature-label reconstruction. Chen et al. [34] propagate label confidence via Sylvester equation in semi-supervised settings. Jain et al. [35] and Kong et al. [36] improve scalability for large-scale missing labels. Modern methods explore nonlinear and hierarchical modeling: Wang et al. [37] use two-level nonlinear mapping fusion. Jiang et al. [38] recover missing labels via label compression and local feature correlation.

Another direction focuses on label semantic dependency modeling: Yang et al. [39] propagate information among semantically correlated labels. Wu et al. [40] use mixed dependency graphs to model feature and label relations. Braytee et al. [41] alleviate class imbalance in incomplete label spaces. Many works also integrate feature selection and subspace learning: Ma and Chow [15] perform label-specific feature selection

with two-level recovery. Yin et al. [42] use multi-scale fuzzy uncertainty for robust feature selection. Dai et al. [43] fuse weak labels via fuzzy discernibility pairs. Sun et al. [44] use fuzzy neighborhood rough sets for compact and consistent features.

2.3 Multi-Label Learning with Noisy and Missing Labels

Most existing methods treat noisy labels and missing labels as two independent problems, but real-world weak supervision often involves both simultaneously. Unified frameworks are still limited. Sun et al. [45] propose a unified weak supervision framework that jointly models label incompleteness and corruption. Ding et al. [46] decompose noisy features and exploit low-rank label structures. Wei et al. [47] design a safe prediction mechanism against unreliable labels. Fang et al. [48] propose an online active learning method with label correlation modeling.

Despite these advances, existing methods still have several limitations when noisy and missing labels coexist. Most prior studies exploit feature-label correlations mainly for label confidence estimation, candidate label refinement, or unreliable label suppression. In these methods, feature information is usually used as an auxiliary cue to judge whether an observed candidate label is reliable. However, they seldom further distinguish different sources of unreliable supervision, such as feature-induced label ambiguity and sparse annotation corruption.

Different from these methods, WPML formulates label recovery as a feature-aware disentanglement problem. Instead of simply modeling the correlation between noisy labels and instance features, WPML explicitly decomposes the observed label matrix into a reliable label confidence component and a sparse noisy component, while using the feature-induced instance similarity structure to regularize the recovered labels. Therefore, the proposed framework differs from existing feature-label correlation modeling methods in three aspects. First, in terms of the disentanglement mechanism, WPML separates feature-induced ambiguity from sparse label corruption rather than treating all unreliable labels as the same type of noise. Second, in terms of the optimization structure, WPML jointly couples feature reconstruction, graph-based label refinement, and sparse noise modeling in a unified objective. Third, in terms of constraint design, WPML preserves local instance consistency through graph Laplacian regularization and isolates abnormal noisy annotations through sparsity constraints. This unified design enables WPML to handle missing and noisy labels in a more principled and interpretable manner.

3 The Proposed Methodology

The core idea of our proposed framework is to unify ambiguous feature identification, compact feature-label collaboration, and graph-based label recovery under a single optimization framework, aiming to address the challenges of partial multi-label learning with missing labels. Unlike existing methods that treat these components independently, our framework integrates them synergistically: ambiguous features (which induce label noise) are first identified and isolated; compact label-relevant features are then extracted to enhance discriminability; finally, graph-based propagation refines label confidences by leveraging instance similarity, ensuring consistency across the data manifold. Each component is indispensable and mutually reinforcing, as elaborated in the following subsections.

3.1 Problem Statement and Notations

Consider a multi-label dataset $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ with a label set $L = \{l_1, l_2, \ldots, l_q\}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ is the feature matrix, $\mathbf{Y} \in \{-1, 0, 1\}^{n \times q}$ is the observed label matrix, $\mathbf{x}_i \in \mathbb{R}^d$ is the d-dimensional feature vector of the i-th instance, n is the number of instances, d is the feature dimension, and q is the number of labels. To address the ambiguity in problem formulation, we explicitly distinguish

three label uncertainties in our weak supervision scenario. We define the entry $\mathbf{Y}_{ij}$ (for the i-th instance $\mathbf{x}_i$ and j-th label l_j) with clear distinctions among the three label types as follows:

$$\mathbf{Y}_{ij} = \begin{cases} 1, & \text{if } l_j \text{ is a candidate label for } \mathbf{x}_i \text{ (partial label: true + noisy labels)}, \\ -1, & \text{if } l_j \text{ is explicitly excluded for } \mathbf{x}_i \text{ (noise-free negative label)}, \\ 0, & \text{if the label is unobserved/missing (true label unknown)}. \end{cases}$$

Our core problem: Given $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ with $\mathbf{Y}_{ij} \in \{-1, 0, 1\}$ (partial/noisy, missing, noise-free negative labels), we learn a robust multi-label classifier to disambiguate true labels from noisy ones ($\mathbf{Y}_{ij} = 1$), infer missing labels ($\mathbf{Y}_{ij} = 0$), and retain reliable negative labels ($\mathbf{Y}_{ij} = -1$). We define the observation mask $\mathbf{J}$ and the observed-positive mask $\mathbf{J}^+$ as

$$J_{ij} = \mathbb{I}(Y_{ij} \neq 0), \qquad J^+_{ij} = \mathbb{I}(Y_{ij} = 1),$$

where $\mathbb{I}(\cdot)$ is the indicator function. The first mask distinguishes observed and missing entries, whereas the second restricts the sparse noise component to observed candidate-positive labels. The frequently used notation is summarized in Table 1.

Table 1: Frequently used notation.

Notation	Meaning
n, d, q, m	Numbers of instances, features, labels, and latent dimensions, respectively
$r > 1$	Fuzziness exponent used in adaptive neighborhood weighting
$\mathbf{x} \in \mathbb{R}^d$, $\mathbf{X} \in \mathbb{R}^{n \times d}$	Feature vector and feature matrix
$\mathbf{y} \in \mathbb{R}^q$, $\mathbf{Y} \in \{-1, 0, 1\}^{n \times q}$	Label vector and observed label matrix; 0 denotes a missing annotation
$\mathcal{N}_k(\mathbf{x}_i)$	Set of the k nearest neighbors of $\mathbf{x}_i$, excluding $\mathbf{x}_i$
$\mathbf{N} \in \mathbb{R}^{n \times n}$	Pairwise distance matrix restricted to the neighborhood graph
$\mathbf{S}, \widehat{\mathbf{S}} \in \mathbb{R}^{n \times n}$	Directed and symmetrized adaptive neighborhood matrices
$\mathbf{S}_{i\cdot} \in \mathbb{R}^{1 \times n}$, $\mathbf{S}_{\cdot j} \in \mathbb{R}^{n \times 1}$	ith row and jth column of $\mathbf{S}$
$\mathbf{D}_S, \mathbf{C}_S = \mathbf{D}_S - \widehat{\mathbf{S}}$	Degree matrix and graph Laplacian
$\mathbf{J}, \mathbf{J}^+ \in \{0, 1\}^{n \times q}$	Observation mask and observed-positive candidate mask
$\mathbf{F} \in \mathbb{R}^{n \times q}$	Reliable label-confidence matrix
$\mathbf{Y}_0 \in \mathbb{R}_+^{n \times q}$	Sparse false-positive annotation-noise matrix
$\mathbf{T} \in \mathbb{R}^{d \times m}$, $\mathbf{P} \in \mathbb{R}^{m \times q}$	Feature projection and latent label-embedding matrices
$\mathbf{W} \in \mathbb{R}^{d \times q}$	Feature-to-noise mapping matrix
$\mathbf{A} \in \mathbb{R}^{n \times q}$, μ, ρ	Lagrange multiplier, penalty parameter, and penalty-growth factor
$\mathbf{\Omega} \in \mathbb{R}^{d \times d}$	Diagonal reweighting matrix for the $\ell_{2,1}$ regularizer
$\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_F$	ℓ_1, ℓ_2, and Frobenius norms
$\circ$, $\oslash$	Hadamard product and Hadamard division
$\mathrm{tr}(\cdot)$, $(\cdot)^\top$	Trace operator and matrix transposition
$\langle \cdot, \cdot \rangle$	Hilbert–Schmidt inner product

3.2 Instance-Level Graph Adjacency Matrix

To capture the pairwise similarity among instances and provide a structural basis for label propagation, we define a graph adjacency matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$. The similarity between two instances $\mathbf{x}_i$ and $\mathbf{x}_j$ decreases with the increase of their squared Euclidean distance $||\mathbf{x}_i - \mathbf{x}_j||^2$, indicating that closer instances are more

similar. This is critical because semantically similar instances should have consistent label confidences, which forms the foundation of our label recovery mechanism. To reflect local consistency without imposing a hard neighborhood assignment, we learn adaptive fuzzy weights over the k nearest neighbors of each instance, following the adaptive-neighborhood principle in [49]. Unlike fuzzy C-means, this formulation does not introduce cluster centers or a cluster-membership matrix; instead, it directly assigns normalized nonnegative weights to neighboring instances. The exponent $r > 1$ controls the softness of these weights, leading to the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} ||\mathbf{x}_i - \mathbf{x}_j||^2 \mathbf{S}_{ij}^r \\ &s.t. \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} = 1, \forall i = 1, \cdots, n \\ &\quad \mathbf{S}_{ii} = 0, \forall i = 1, \cdots, n \\ &\quad \mathbf{S}_{ij} \geq 0, \forall i, j = 1, \cdots, n \end{aligned} \tag{1}$$

Here, $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of the k nearest neighbors of $\mathbf{x}_i$ selected from $\mathcal{X} \backslash \{\mathbf{x}_i\}$. Thus, $\mathbf{x}_i \notin \mathcal{N}_k(\mathbf{x}_i)$, and self-similarity is excluded by setting $S_{ii} = 0$. The similarity weights over the k neighbors are normalized to sum to one, while S_{ij} is set to zero for instances outside $\mathcal{N}_k(\mathbf{x}_i)$.

Let $\mathbf{N} \in \mathbb{R}^{n \times n}$ denote the instance distance matrix, which is defined as follows:

$$\mathbf{N}_{ij} = \begin{cases} ||\mathbf{x}_i - \mathbf{x}_j||^2, & \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ \infty, & \text{otherwise} \end{cases} \tag{2}$$

Eq. (1) can be equivalently rewritten by substituting the distance matrix $\mathbf{N}$, simplifying the optimization while retaining its core logic (minimizing similarity-weighted distance):

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{N}_{ij} \mathbf{S}_{ij}^r \\ &s.t. \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} = 1, \forall i = 1, \cdots, n \\ &\quad \mathbf{S}_{ii} = 0, \forall i = 1, \cdots, n \\ &\quad \mathbf{S}_{ij} \geq 0, \forall i, j = 1, \cdots, n \end{aligned} \tag{3}$$

We then derive the Lagrangian function associated with Eq. (3) to obtain the optimal solution, ensuring the similarity matrix $\mathbf{S}$ is adaptive to the data distribution:

$$\min_{\mathbf{S}} \{ \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{N}_{ij} \mathbf{S}_{ij}^r - \theta (\sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} - 1) - \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{b}_j \mathbf{S}_{ij} \} \tag{4}$$

where θ and $\mathbf{b}$ are the Lagrangian multipliers. According to the Karush-Kuhn-Tucker (KKT) conditions, the optimal solution is given by:

$$S_{ij} = \begin{cases} \dfrac{\left(\dfrac{1}{\mathbf{N}_{ij}}\right)^{\frac{1}{r-1}}}{\sum\limits_{\mathbf{x}_l \in \mathcal{N}_k(\mathbf{x}_i)} \left(\dfrac{1}{\mathbf{N}_{il}}\right)^{\frac{1}{r-1}}}, & \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Denote $\widehat{\mathbf{S}} = \frac{1}{2}(\mathbf{S} + \mathbf{S}^\top)$ as the symmetric adjacency matrix representing pairwise instance relationships. This matrix is essential for subsequent label propagation, as it encodes the intrinsic manifold structure of the data. Unlike conventional manifold learning approaches [50] that rely on predefined graphs (which may not adapt to complex data distributions), we directly infer $\widehat{\mathbf{S}}$ from the data through adaptive fuzzy neighborhood weighting—this adaptability is necessary to handle the scattered and uncertain feature distributions common in partial multi-label learning with missing labels.

3.3 Feature-Aware Label Disentanglement

To robustly handle unreliable annotations in multi-label learning, we propose a feature-aware label disentanglement module as the core of our framework. Specifically, the observed annotation matrix $\mathbf{Y}$ is decomposed into a reliable label confidence matrix $\mathbf{F} \in \mathbb{R}^{n\times q}$ and a nonnegative sparse noise matrix $\mathbf{Y}_0 \in \mathbb{R}^{n\times q}$:

$$\mathbf{Y} = \mathbf{F} \circ \mathbf{J} + \mathbf{Y}_0, \tag{6}$$

where $\circ$ denotes the element-wise Hadamard product, and $\mathbf{J} \in \{0,1\}^{n\times q}$ indicates the observed label entries.

The nonnegativity constraint on $\mathbf{Y}_0$ follows from the one-sided annotation noise considered in this work. Specifically, $\mathbf{Y}_0$ captures spurious positive annotations among the observed labels, whereas potentially missing positive labels are treated as unobserved entries with $J_{ij} = 0$ and are inferred through the reliable label matrix $\mathbf{F}$. Accordingly, the support constraint $\mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}^+$ restricts the noise component to observed candidate-positive entries. False-positive annotations are modeled by $\mathbf{Y}_0$, while missing labels are handled through the observation mask $\mathbf{J}$ and the subsequent label reconstruction process.

Feature-aware label reconstruction. Inspired by the shared feature projection and label-specific prediction mechanism in [51], we construct a feature-aware label reconstruction model to exploit the discriminative information contained in the instance features $\mathbf{X}$. Specifically, given $\mathbf{X} \in \mathbb{R}^{n\times d}$, we jointly learn a feature projection matrix $\mathbf{T} \in \mathbb{R}^{d\times q}$ and a label embedding matrix $\mathbf{P} \in \mathbb{R}^{q\times q}$ to reconstruct the reliable label matrix $\mathbf{F} \in \mathbb{R}^{n\times q}$:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{T},\mathbf{F}} \quad & \frac{1}{2}\|\mathbf{XTP} - \mathbf{F}\|_F^2 + \frac{\alpha}{2}\|\mathbf{T}\|_F^2 \\ \text{s.t.} \quad & \mathbf{PP}^\top = \mathbf{I}. \end{aligned} \tag{7}$$

Here, the rows of $\mathbf{P}$ correspond to the latent semantic bases over the label space. The constraint $\mathbf{PP}^\top = \mathbf{I}$ enforces these latent bases to be mutually orthogonal, thereby reducing redundant semantic correlations. Note that this constraint imposes row orthogonality on $\mathbf{P}$ rather than orthogonality among its columns.

Although inspired by [51], our formulation differs from it in both objective and structure. Rather than directly constructing label-specific features or prediction functions, we jointly learn $\mathbf{T}$ and $\mathbf{P}$ to reconstruct the reliable label matrix $\mathbf{F}$. Furthermore, the orthogonality constraint $\mathbf{PP}^\top = \mathbf{I}$ is introduced to encourage complementary latent semantic factors and suppress redundant label correlations.

Preserving label manifold structure. High-dimensional input data often contain redundant or irrelevant features, which may degrade model performance by introducing noise and increasing computational complexity. Therefore, extracting compact label-relevant features is crucial for improving both interpretability and generalization. Such compact features are typically decorrelated, encoding distinct information aligned with label semantics. To further preserve the local similarity structure among instances in the feature space, we apply a graph Laplacian regularization on the reliable label matrix $\mathbf{F}$:

$$\operatorname{tr}(\mathbf{F}^{\top}\mathbf{C}_S\mathbf{F}) = \frac{1}{2}\sum_{i,j}\widehat{\mathbf{S}}_{ij}\|\mathbf{F}_{i\cdot} - \mathbf{F}_{j\cdot}\|_2^2, \tag{8}$$

where $\mathbf{C}_S = \mathbf{D}_S - \widehat{\mathbf{S}}$ is the graph Laplacian and $\mathbf{D}_S$ is the degree matrix satisfying $[\mathbf{D}_S]_{ii} = \sum_{j=1}^{n}\widehat{\mathbf{S}}_{ij}$. This term ensures that instances with similar features have consistent reliable labels.

Sparse noise modeling. Prior studies [13] have demonstrated that noisy labels in real-world datasets are often correlated with ambiguous or confounding features, rather than being purely random. Existing approaches typically do not address the underlying causes of label noise, which may result in suboptimal label correction. To explicitly model and isolate such noisy annotations, we formulate a sparse regression problem:

$$\min_{\mathbf{W},\mathbf{Y}_0}\frac{1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2 + \beta\|\mathbf{W}\|_{2,1}, \tag{9}$$

where $\mathbf{W}$ maps features to the noise component $\mathbf{Y}_0$, and the $\ell_{2,1}$-norm encourages row-wise sparsity, effectively isolating spurious annotations.

Together, these components form a unified module that integrates feature-guided reconstruction, manifold regularization, and sparse noise modeling. This disentanglement not only reduces the impact of noisy labels but also produces clean label embeddings $\mathbf{F}$ for downstream multi-label prediction, serving as the core building block of our framework.

By integrating Eqs. (6)–(9), we obtain a unified framework that simultaneously performs feature mapping, label consistency refinement, instance manifold learning, and ambiguous feature identification. The resulting joint optimization problem is expressed as:

$$\begin{aligned}
\min_{\mathbf{T},\mathbf{P},\mathbf{W},\mathbf{F},\mathbf{Y}_0}\ & \frac{1}{2}\|\mathbf{X}\mathbf{T}\mathbf{P} - \mathbf{F}\|_F^2 + \frac{1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2 + \frac{\lambda}{2}\operatorname{tr}(\mathbf{F}^{\top}\mathbf{C}_S\mathbf{F}) \\
& + \frac{\alpha}{2}\|\mathbf{T}\|_F^2 + \beta\|\mathbf{W}\|_{2,1} + \gamma\|\mathbf{Y}_0\|_1 \\
\text{s.t.}\quad & \mathbf{Y} = \mathbf{F}\circ\mathbf{J} + \mathbf{Y}_0, \\
& \mathbf{Y}_0 = \mathbf{Y}_0\circ\mathbf{J}^{+}, \\
& \mathbf{P}\mathbf{P}^{\top} = \mathbf{I}_m, \\
& \mathbf{Y}_0 \geq \mathbf{0}.
\end{aligned} \tag{10}$$

By adopting a block coordinate descent strategy, each variable is iteratively updated while holding the others fixed, yielding an efficient alternating optimization procedure. The detailed algorithmic procedure is provided in the subsequent section.

4 Solutions to the Optimization Problem

Following the Linearized Alternating Direction Method with Adaptive Penalty (LADMAP) framework, Eq. (10) can be equivalently reformulated as:

$$\begin{aligned}\mathcal{L}(\mathbf{T},\mathbf{P},\mathbf{W},\mathbf{F},\mathbf{Y}_0,\mathbf{A}) =& \frac{1}{2}\|\mathbf{XTP}-\mathbf{F}\|_F^2+\frac{1}{2}\|\mathbf{XW}-\mathbf{Y}_0\|_F^2+\frac{\lambda}{2}\operatorname{tr}(\mathbf{F}^\top\mathbf{C}_S\mathbf{F})\\ &+\frac{\alpha}{2}\|\mathbf{T}\|_F^2+\beta\|\mathbf{W}\|_{2,1}+\gamma\|\mathbf{Y}_0\|_1\\ &+\frac{\mu}{2}\left\|\mathbf{Y}-\mathbf{F}\circ\mathbf{J}-\mathbf{Y}_0+\frac{\mathbf{A}}{\mu}\right\|_F^2\\ \text{s.t.}\quad &\mathbf{Y}_0=\mathbf{Y}_0\circ\mathbf{J}^+,\\ &\mathbf{PP}^\top=\mathbf{I}_m,\\ &\mathbf{Y}_0\geq\mathbf{0}.\end{aligned} \tag{11}$$

where $\mathbf{A}\in\mathbb{R}^{n\times q}$ is the Lagrange multiplier matrix and $\mu>0$ is a trade-off parameter controlling the penalty for constraint violation.

This reformulation linearizes the augmented term, allowing each variable to be updated efficiently in a block coordinate descent manner while adaptively adjusting the penalty parameter μ. This LADMAP-style reformulation enables efficient alternating updates while respecting the orthogonality and sparsity constraints. Because the overall problem is nonconvex, the convergence discussion below is limited to blockwise descent and numerical stability rather than global optimality.

4.1 Updating $\mathbf{P}$ *by Fixing Other Variables*

The subproblem for $\mathbf{P}$ is formulated as:

$$\begin{aligned}\min_{\mathbf{P}}\quad &\frac{1}{2}\|\mathbf{XTP}-\mathbf{F}\|_F^2\\ \text{s.t.}\quad &\mathbf{PP}^\top=\mathbf{I}_m.\end{aligned} \tag{12}$$

Let $\mathbf{F}^\top\mathbf{XT}=\mathbf{B\Sigma R}^\top$ be a singular value decomposition. The orthogonal Procrustes solution is

$$\mathbf{P}=\mathbf{RB}^\top. \tag{13}$$

4.2 Updating $\mathbf{T}$ *by Fixing Other Variables*

Optimizing $\mathbf{T}$ corresponds to solving:

$$\min_{\mathbf{T}}\mathcal{L}=\frac{1}{2}\|\mathbf{XTP}-\mathbf{F}\|_F^2+\frac{\alpha}{2}\|\mathbf{T}\|_F^2. \tag{14}$$

The closed-form solution can be written as:

$$\mathbf{T}=\mathbf{U}\left[(\mathbf{U}^\top\mathbf{X}^\top\mathbf{FP}^\top\mathbf{V})\oslash(\Lambda_1\mathbf{1}_d\mathbf{1}_m^\top\Lambda_2+\alpha\mathbf{1}_d\mathbf{1}_m^\top)\right]\mathbf{V}^\top, \tag{15}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the eigenvector matrices of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{PP}^\top$, respectively, and Λ_1, Λ_2 are the corresponding diagonal eigenvalue matrices.

4.3 Updating $\mathbf{W}$ *by Fixing Other Variables*

With $\mathbf{T}$, $\mathbf{P}$, $\mathbf{F}$, $\mathbf{Y}_0$ fixed, the subproblem for $\mathbf{W}$ is:

$$\min_{\mathbf{W}}\mathcal{L}=\frac{1}{2}\|\mathbf{XW}-\mathbf{Y}_0\|_F^2+\beta\|\mathbf{W}\|_{2,1}. \tag{16}$$

Its gradient is

$$\nabla_{\mathbf{W}}\mathcal{L} = \mathbf{X}^{\top}(\mathbf{X}\mathbf{W} - \mathbf{Y}_0) + \beta\Omega\mathbf{W}, \tag{17}$$

where $\Omega \in \mathbb{R}^{d\times d}$ is diagonal:

$$\Omega = \begin{bmatrix} \frac{1}{\|\mathbf{W}_{1\cdot}\|_2+\varepsilon} & & \\ & \ddots & \\ & & \frac{1}{\|\mathbf{W}_{d\cdot}\|_2+\varepsilon} \end{bmatrix} \tag{18}$$

where $\varepsilon > 0$ prevents division by zero. Treating Ω as fixed at iteration t (denoted $\Omega^{(t)}$), the closed-form update for $\mathbf{W}$ is:

$$\mathbf{W}^{(t+1)} = (\mathbf{X}^{\top}\mathbf{X} + \beta\Omega^{(t)})^{-1}\mathbf{X}^{\top}\mathbf{Y}_0. \tag{19}$$

4.4 Updating $\mathbf{F}$ *by Fixing Other Variables*

The subproblem for $\mathbf{F}$ is:

$$\min_{\mathbf{F}} \mathcal{L} = \frac{1}{2}\|\mathbf{X}\mathbf{T}\mathbf{P} - \mathbf{F}\|_F^2 + \frac{\lambda}{2}\operatorname{tr}(\mathbf{F}^{\top}\mathbf{C}_S\mathbf{F}) + \frac{\mu}{2}\|\mathbf{Y} - \mathbf{F}\circ\mathbf{J} - \mathbf{Y}_0 + \frac{\mathbf{A}}{\mu}\|_F^2. \tag{20}$$

Setting the derivative with respect to $\mathbf{F}$ to zero gives

$$\mathbf{F} - \mathbf{X}\mathbf{T}\mathbf{P} + \lambda\mathbf{C}_S\mathbf{F} + \mu\left[\left(\mathbf{F} - \mathbf{Y} + \mathbf{Y}_0 - \frac{\mathbf{A}}{\mu}\right)\circ\mathbf{J}\right] = \mathbf{0}. \tag{21}$$

Since $\mathbf{J}$ is a binary matrix, the linear system for each column $\mathbf{F}_{\cdot,j}$ is: $1 \le j \le q$,

$$\left[\mathbf{I} + \lambda\mathbf{C}_S + \mu\operatorname{diag}(\mathbf{J}_{\cdot,j})\right]\mathbf{F}_{\cdot,j} = (\mathbf{X}\mathbf{T}\mathbf{P})_{\cdot,j} + \operatorname{diag}(\mathbf{J}_{\cdot,j})\left[\mu(\mathbf{Y}_{\cdot,j} - \mathbf{Y}_{0,\cdot,j}) + \mathbf{A}_{\cdot,j}\right]. \tag{22}$$

4.5 Updating $\mathbf{Y_0}$ *by Fixing Other Variables*

The subproblem over $\mathbf{Y}_0$ is:

$$\begin{aligned} \min_{\mathbf{Y}_0}\ \mathcal{L} =& \frac{1}{2}||\mathbf{X}\mathbf{W} - \mathbf{Y}_0||_F^2 + \gamma||\mathbf{Y}_0||_1 + \frac{\mu}{2}||\mathbf{Y} - \mathbf{F}\circ\mathbf{J} - \mathbf{Y}_0 + \frac{\mathbf{A}}{\mu}||_F^2 \\ \text{s.t.}\quad & \mathbf{Y}_0 = \mathbf{Y}_0\circ\mathbf{J}^{+} \\ & \mathbf{Y}_0 \ge \mathbf{0} \end{aligned} \tag{23}$$

This admits the closed-form solution:

$$\mathbf{Y}_0 = \mathcal{S}^{+}_{\frac{\gamma}{1+\mu}}\left[\frac{\mathbf{X}\mathbf{W} + \mu\mathbf{Y} - \mu\mathbf{F}\circ\mathbf{J} + \mathbf{A}}{1+\mu}\right]\circ\mathbf{J}^{+}, \tag{24}$$

where $\mathcal{S}^{+}_{w}(a) = \max(a - w, 0)$ is the element-wise positive soft-thresholding operator.

Finally, the Lagrange multiplier $\mathbf{A}$ and penalty parameter μ are updated as:

$$\begin{aligned} \mathbf{A}^{(t+1)} =& \mathbf{A}^{(t)} + \mu^{(t)}(\mathbf{Y} - \mathbf{F}\circ\mathbf{J} - \mathbf{Y}_0) \\ \mu^{(t+1)} =& \min(\mu_{\max}, \rho\mu^{(t)}) \end{aligned} \tag{25}$$

where $\rho > 1$ controls the adaptive increment of μ and $\mu_{\max}$ is an upper bound.

4.6 Convergence Analysis

To theoretically verify the convergence of the proposed optimization framework, we first introduce a useful lemma.

Lemma 1 [49]: *Given any two real numbers $x, y \geq 0$, the following inequality holds:*

$$x - \frac{x^2}{2y} \leq y - \frac{y^2}{2y}.$$

Theorem 1: *Algorithm 1 monotonically decreases the objective function with respect to $\mathbf{W}$ in each iteration.*

Proof: Define the surrogate function for $\mathbf{W}$ at iteration t as

$$\mathcal{F}(\mathbf{W}) = \frac{1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2 + \beta \operatorname{tr}(\mathbf{W}^\top \Omega^{(t)} \mathbf{W}). \tag{26}$$

The gradient of $\mathcal{F}(\mathbf{W})$ with respect to $\mathbf{W}$ is

$$\nabla_{\mathbf{W}} \mathcal{F}(\mathbf{W}) = \mathbf{X}^\top (\mathbf{X}\mathbf{W} - \mathbf{Y}_0) + 2\beta \Omega^{(t)} \mathbf{W}. \tag{27}$$

Let $\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{W}} \mathcal{F}(\mathbf{W})$. Then, by definition,

$$\mathcal{F}(\mathbf{W}^{(t+1)}) \leq \mathcal{F}(\mathbf{W}^{(t)}). \tag{28}$$

By Lemma 1, for each row $\mathbf{W}_{i\cdot}$, we have

$$\|\mathbf{W}_{i\cdot}^{(t+1)}\|_2 - \frac{\|\mathbf{W}_{i\cdot}^{(t+1)}\|_2^2}{2\|\mathbf{W}_{i\cdot}^{(t)}\|_2} \leq \|\mathbf{W}_{i\cdot}^{(t)}\|_2 - \frac{\|\mathbf{W}_{i\cdot}^{(t)}\|_2^2}{2\|\mathbf{W}_{i\cdot}^{(t)}\|_2}. \tag{29}$$

Summing over all rows yields

$$\|\mathbf{W}^{(t+1)}\|_{2,1} - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{(t+1)\top} \Omega^{(t)} \mathbf{W}^{(t+1)}\right) \leq \|\mathbf{W}^{(t)}\|_{2,1} - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{(t)\top} \Omega^{(t)} \mathbf{W}^{(t)}\right). \tag{30}$$

Combining Eqs. (28) and (30), we obtain

$$\mathcal{F}(\mathbf{W}^{(t+1)}) + \beta\|\mathbf{W}^{(t+1)}\|_{2,1} \leq \mathcal{F}(\mathbf{W}^{(t)}) + \beta\|\mathbf{W}^{(t)}\|_{2,1}. \tag{31}$$

Eq. (31) demonstrates that the objective function $\mathcal{L}(\mathbf{W})$ decreases monotonically at each iteration, completing the proof. □

The remaining blocks are either solved exactly or updated by closed-form minimizers when the other variables are fixed: $\mathbf{P}$ is an orthogonal Procrustes problem, $\mathbf{T}$ and $\mathbf{F}$ are convex quadratic subproblems, and $\mathbf{Y}_0$ is obtained by positive soft-thresholding. The primal objective is bounded below because all of its terms are nonnegative for $\lambda, \alpha, \beta, \gamma \geq 0$. These properties establish blockwise descent and boundedness of the primal objective. Because the complete problem contains coupled variables, an orthogonality constraint, and adaptive multiplier updates, we do not claim global optimality or provide a general stationary-point guarantee. Instead, the convergence curves empirically verify that the objective and variable changes stabilize in practice.

Algorithm 1: The proposed WPML algorithm

Require: Feature matrix $\mathbf{X}$, observed label matrix $\mathbf{Y}$, latent dimension m, neighborhood size k, fuzziness exponent $r > 1$, parameters $\lambda, \alpha, \beta, \gamma, \mu_0, \rho, \mu_{\max}$, tolerance τ, and maximum iteration number $T_{\max}$.
Ensure: Prediction matrix $\mathbf{B} = \mathbf{TP}$ and reliable label-confidence matrix $\mathbf{F}$.

Construct $\mathbf{J}$ and $\mathbf{J}^+$ from $\mathbf{Y}$, and compute the distance matrix $\mathbf{N}$.
Compute $\mathbf{S}$ according to Eq. (5), symmetrize it as $\widehat{\mathbf{S}} = (\mathbf{S} + \mathbf{S}^\top)/2$, and form $\mathbf{C}_S$.
Initialize $t \leftarrow 0$, $\mathbf{F}^{(0)} \leftarrow \mathbf{Y}$, $\mathbf{Y}_0^{(0)} \leftarrow \mathbf{0}$, $\mathbf{W}^{(0)} \leftarrow \mathbf{0}$, $\mathbf{A}^{(0)} \leftarrow \mathbf{0}$, $\mu^{(0)} \leftarrow \mu_0$, and initialize $\mathbf{T}^{(0)} \in \mathbb{R}^{d\times m}$ and $\mathbf{P}^{(0)} \in \mathbb{R}^{m\times q}$ with compatible dimensions.
repeat
  Update $\mathbf{P}^{(t+1)}$ according to Eq. (13).
  Update $\mathbf{T}^{(t+1)}$ according to Eq. (15).
  Form $\Omega^{(t)}$ according to Eq. (18) and update $\mathbf{W}^{(t+1)}$ according to Eq. (19).
  Update $\mathbf{F}^{(t+1)}$ according to Eq. (22).
  Update $\mathbf{Y}_0^{(t+1)}$ according to Eq. (24).
  Update $\mathbf{A}^{(t+1)}$ and $\mu^{(t+1)}$ according to Eq. (25).
  $t \leftarrow t + 1$.
until $t = T_{\max}$ or the primal residual and the maximum relative variable change are below τ
return $\mathbf{B} = \mathbf{TP}$ and $\mathbf{F}$.

For a test feature matrix $\mathbf{X}_{\text{te}}$, the prediction scores are computed as $\widehat{\mathbf{Y}}_{\text{te}} = \mathbf{X}_{\text{te}}\mathbf{TP}$.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets and Preprocessing

We evaluate WPML on 15 publicly available multi-label benchmark datasets, including Bibtex, Birds, Business, CAL500, Computers, Education, Emotions, Enron, Health, Medical, Reference, Science, Social, Scene, and Yeast. These datasets cover text, image, music, biology, and web-related applications and exhibit considerable differences in sample size, feature dimensionality, label cardinality, and label density. Their detailed statistics are summarized in Table 2, where n, d, and q denote the numbers of instances, features, and labels, respectively, while ℓ_c and ℓ_d denote label cardinality and label density.

Table 2: Statistics of the benchmark datasets used in the experiments.

Data Set	n	d	q	ℓ_c	ℓ_d
Bibtex	7395	1836	159	2.402	0.015
Birds	645	260	19	1.014	0.053
Business	5000	438	30	1.588	0.053
CAL500	502	68	174	26.044	0.150
Computers	5000	681	33	1.509	0.048
Education	5000	550	33	1.461	0.044
Emotions	593	72	6	1.868	0.311
Enron	1702	1001	53	3.378	0.064
Health	5000	612	32	1.662	0.052
Medical	978	1449	45	1.245	0.028

(Continued)

Table 2 (continued)

Data Set	n	d	q	ℓ_c	ℓ_d
Reference	5000	793	33	1.169	0.035
Scene	2407	294	6	1.074	0.179
Science	5000	743	40	1.451	0.036
Social	5000	743	39	1.283	0.033
Yeast	2417	103	14	4.237	0.303

To generate feature-dependent corruption without using test-label information, a ridge-regression mapping $\mathbf{B}_c$ is learned independently on each training split by minimizing.

$$\|\mathbf{X}_{\mathrm{tr}}\mathbf{B}_c - \mathbf{Y}_{\mathrm{tr}}\|_F^2 + \eta\|\mathbf{B}_c\|_F^2.$$

The resulting training scores are used only to rank annotation entries. Among originally positive entries, those with the lowest scores are preferentially masked until the prescribed missing-label rate is reached; the rate is measured relative to the number of positive training annotations. Among originally negative entries, those with the highest scores are flipped to candidate-positive labels until the prescribed noise rate is reached; the rate is measured relative to the number of negative training annotations. Both rates are set to 20%, 40%, and 60%. Corruption is regenerated independently for each training split, whereas the test labels remain unchanged. The regularization parameter η and all model hyperparameters are selected using training data only.

5.2 Compared Methods

We compare WPML with six representative multi-label and partial multi-label learning methods. Publicly available implementations are used whenever possible, and their parameters are configured according to the original papers or released codes.

- **WPML (Algorithm 1):** Parameters λ and α are selected from $\{10^{-2}, 10^{-1}, \ldots, 10^{1}\}$, β from $\{10^{-3}, 10^{-2}, \ldots, 10^{1}\}$, and γ from $\{0.1, 0.2, \ldots, 2.5\}$. The neighborhood size is fixed at $k = 10$.
- **NLR**[1] [29]: Jointly learns a predictive classifier and a competing classifier to identify and remove noisy candidate labels. We set $\beta = 100$, $\gamma = 10$, and the prediction threshold to 0.9, while λ is selected for each dataset.
- **PML-MA**[2] [52]: Recovers reliable pseudo-labels through low-rank decomposition and aligns the feature and pseudo-label modalities by preserving their global and local structures. The parameters follow the recommended settings in the released code, with $k = 5$ and a maximum of 50 iterations.
- **ML-KNN**[3] [5]: Predicts labels using nearest-neighbor label statistics and maximum a posteriori estimation. The neighborhood size is set to $k = 10$, and the smoothing parameter is fixed at 1.
- **PML-LD**[4] [53]: Recovers numerical label distributions by exploiting feature topology and label correlations. We set $\lambda_1 = \lambda_2 = 0.01$ and $K = 10$ and use the recommended regression parameters.

[1] https://github.com/Yangfc-ML/NLR
[2] https://github.com/CcAmbiguous/PML-MA
[3] https://www.lamda.nju.edu.cn/code_MLkNN.ashx
[4] https://github.com/palm-ml/PML_LD

- **PML-NI**[5] [13]: Jointly learns a low-rank multi-label classifier and a sparse noisy-label identification model. Following the original setting, $\lambda = 1$, $\beta = 1$, and $\gamma = 0.5$.
- **PML-LENF**[6] [54]: Enhances candidate labels using near and far neighbors and learns a nonlinear label predictor. We set $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.01$, $\lambda_4 = 10$, and $k = q$.

The parameters of WPML are selected using the training data only. No test-label information is used for either parameter selection or feature-dependent label generation.

5.3 Evaluation Metrics

We adopt Macro-AUC, Ranking Loss, Average Precision, and Coverage to evaluate different aspects of multi-label prediction. Let $\mathcal{L}'$ denote the labels containing both positive and negative test instances. Macro-AUC is calculated by first evaluating the binary AUC for each valid label and then averaging over labels:

$$\text{Macro-AUC} = \frac{1}{|\mathcal{L}'|} \sum_{c \in \mathcal{L}'} \text{AUC}_c. \tag{32}$$

It assigns equal importance to each label and is therefore suitable for datasets with imbalanced label frequencies. Let $\mathcal{Y}_i^+$ and $\mathcal{Y}_i^-$ denote the relevant and irrelevant label sets of the ith test instance, let $f_i(c)$ be the prediction score of label c, and let $\pi_i(c)$ be its descending rank. The remaining metrics are computed as

$$\text{RankingLoss} = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \frac{|\{(a,b) \in \mathcal{Y}_i^+ \times \mathcal{Y}_i^- : f_i(a) \le f_i(b)\}|}{|\mathcal{Y}_i^+||\mathcal{Y}_i^-|}, \tag{33}$$

$$\text{AveragePrecision} = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \frac{1}{|\mathcal{Y}_i^+|} \sum_{a \in \mathcal{Y}_i^+} \frac{|\{b \in \mathcal{Y}_i^+ : \pi_i(b) \le \pi_i(a)\}|}{\pi_i(a)}, \tag{34}$$

$$\text{Coverage} = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \frac{\max_{a \in \mathcal{Y}_i^+} \pi_i(a) - 1}{q - 1}. \tag{35}$$

Macro-AUC and Average Precision are higher-is-better metrics, whereas Ranking Loss and normalized Coverage are lower-is-better metrics.

5.4 Comparison Results

Tables 3–6 report the detailed results under the most challenging setting, in which both the feature-dependent missing-label rate and label-noise rate are 60%. Additional critical-difference diagrams under the 20% and 40% settings provide complementary comparisons across weaker corruption levels.

As shown in Tables 3–6, WPML achieves the best overall average rank for all four metrics. Its average ranks are 1.400 for AUC, 1.600 for Ranking Loss, 1.733 for Average Precision, and 1.533 for Coverage. NLR is the strongest competing method overall, with corresponding average ranks of 2.667, 2.667, 1.867, and 3.733.

[5] https://xiemk.github.io/code/PMLNIcode.zip
[6] https://github.com/CcAmbiguous/PML-LENFN

Table 3: Comparison of different algorithms in terms of AUC under 60% missing-label and label-noise rates, with the best results highlighted in bold (higher is better).

Datasets	NLR	PML-MA	ML-KNN	PML-LD	PML-NI	PML-LENF	WPML
Bibtex	0.663	0.657	0.527	0.386	0.386	0.386	**0.681**
Birds	0.343	0.328	0.478	0.337	0.337	0.337	**0.795**
Business	0.453	0.211	**0.534**	0.471	0.472	0.472	0.506
CAL500	0.465	0.470	0.505	0.473	0.473	0.473	**0.742**
Computers	**0.613**	0.316	0.454	0.581	0.581	0.581	0.605
Education	0.652	0.645	0.492	0.651	0.651	0.651	**0.771**
Emotions	0.719	0.669	0.669	0.645	0.645	0.645	**0.736**
Enron	0.446	**0.592**	0.512	0.365	0.365	0.365	0.466
Health	**0.640**	0.366	0.361	0.614	0.614	0.615	0.618
Medical	0.384	0.317	0.480	0.297	0.297	0.297	**0.872**
Reference	0.630	0.448	0.415	0.617	0.617	0.617	**0.674**
Science	**0.808**	0.758	0.758	0.706	0.706	0.706	0.780
Social	0.695	0.667	0.532	0.678	0.678	0.678	**0.756**
Scene	0.586	0.544	0.573	0.554	0.554	0.554	**0.808**
Yeast	0.667	0.625	0.576	0.653	0.653	0.653	**0.765**
Average rank	2.667	5.267	4.267	5.200	4.433	4.767	**1.400**

Table 4: Comparison of different algorithms in terms of Ranking Loss under 60% missing-label and label-noise rates, with the best results highlighted in bold (lower is better).

Datasets	NLR	PML-MA	ML-KNN	PML-LD	PML-NI	PML-LENF	WPML
Bibtex	0.337	0.344	0.476	0.637	0.637	0.637	**0.317**
Birds	0.712	0.717	0.515	0.722	0.722	0.722	**0.173**
Business	0.586	0.838	**0.480**	0.565	0.564	0.564	0.516
CAL500	0.541	0.530	0.497	0.530	0.530	0.530	**0.255**
Computers	**0.380**	0.742	0.563	0.422	0.422	0.422	0.384
Education	0.329	0.339	0.513	0.333	0.333	0.333	**0.210**
Emotions	0.261	0.318	0.318	0.341	0.341	0.341	**0.235**
Enron	0.552	**0.427**	0.500	0.637	0.637	0.637	0.534
Health	**0.344**	0.631	0.703	0.373	0.372	0.372	0.381
Medical	0.614	0.680	0.527	0.704	0.703	0.703	**0.115**
Reference	0.349	0.554	0.605	0.365	0.365	0.365	**0.311**
Science	**0.163**	0.214	0.214	0.268	0.268	0.268	0.191
Social	0.284	0.319	0.466	0.305	0.305	0.305	**0.226**
Scene	0.409	0.459	0.413	0.448	0.448	0.448	**0.160**
Yeast	0.321	0.363	0.427	0.337	0.337	0.337	**0.223**
Average rank	2.667	4.867	4.267	5.367	4.433	4.800	**1.600**

Table 5: Comparison of different algorithms in terms of average precision under 60% missing-label and label-noise rates, with the best results highlighted in bold (higher is better).

Datasets	NLR	PML-MA	ML-KNN	PML-LD	PML-NI	PML-LENF	WPML
Bibtex	0.161	0.099	0.052	0.030	0.030	0.030	**0.169**
Birds	0.129	0.124	0.171	0.129	0.129	0.129	**0.649**
Business	**0.200**	0.110	0.141	0.183	0.183	0.184	0.164
CAL500	0.175	0.165	0.172	0.164	0.164	0.164	**0.422**
Computers	**0.283**	0.105	0.117	0.244	0.244	0.244	0.258
Education	0.317	0.292	0.127	0.290	0.290	0.290	**0.416**
Emotions	**0.724**	0.669	0.669	0.645	0.645	0.645	0.718
Enron	0.112	**0.211**	0.132	0.092	0.091	0.091	0.129
Health	**0.353**	0.159	0.083	0.299	0.299	0.299	0.337
Medical	0.079	0.059	0.072	0.069	0.069	0.069	**0.642**
Reference	**0.344**	0.193	0.149	0.298	0.298	0.298	0.325
Science	**0.759**	0.666	0.666	0.604	0.604	0.604	0.689
Social	0.326	0.291	0.121	0.292	0.291	0.292	**0.435**
Scene	0.198	0.174	0.161	0.171	0.171	0.171	**0.567**
Yeast	0.588	0.539	0.483	0.570	0.570	0.570	**0.703**
Average rank	1.867	4.867	5.000	4.800	4.967	4.767	**1.733**

Table 6: Comparison of different algorithms in terms of coverage under 60% missing-label and label-noise rates, with the best results highlighted in bold (lower is better).

Datasets	NLR	PML-MA	ML-KNN	PML-LD	PML-NI	PML-LENF	WPML
Bibtex	0.521	0.493	0.629	0.765	0.765	0.765	**0.466**
Birds	0.741	0.736	0.539	0.744	0.744	0.744	**0.229**
Business	0.677	0.877	**0.542**	0.646	0.644	0.644	0.586
CAL500	0.973	0.959	0.958	0.959	0.959	0.959	**0.844**
Computers	0.455	0.806	0.610	0.483	0.483	0.482	**0.449**
Education	0.389	0.392	0.565	0.387	0.387	0.387	**0.261**
Emotions	0.399	0.432	0.432	0.445	0.445	0.445	**0.371**
Enron	0.776	**0.668**	0.731	0.830	0.831	0.830	0.759
Health	**0.433**	0.711	0.763	0.455	0.455	0.454	0.457
Medical	0.638	0.700	0.547	0.725	0.724	0.724	**0.145**
Reference	0.393	0.570	0.613	0.391	0.391	0.391	**0.161**
Science	**0.157**	0.196	0.196	0.241	0.241	0.241	0.176
Social	0.361	0.367	0.512	0.354	0.354	0.354	**0.275**
Scene	0.474	0.491	0.447	0.479	0.479	0.479	**0.196**
Yeast	0.608	0.616	0.723	0.607	0.607	0.607	**0.532**
Average rank	3.733	4.733	4.200	5.000	4.300	4.500	**1.533**

At the dataset level, WPML obtains the best AUC and Ranking Loss on 10 of the 15 datasets, the best Average Precision on 8 datasets, and the best Coverage on 11 datasets. The improvements are particularly evident on Birds, Medical, and Scene. For example, on Medical, WPML obtains an AUC of 0.872, a Ranking Loss of 0.115, an Average Precision of 0.642, and a Coverage of 0.145. The best competing results on this dataset are 0.480, 0.527, 0.079, and 0.547, respectively. This substantial difference indicates that WPML remains effective when feature-dependent corruption severely reduces the reliability of the observed labels.

On Birds, WPML improves AUC from the best competing result of 0.478 to 0.795 and reduces Ranking Loss from 0.515 to 0.173. Similar improvements are observed on Scene, where WPML achieves an AUC of 0.808 and an Average Precision of 0.567. These results suggest that jointly identifying ambiguous features and propagating reliable label information is particularly useful when annotation errors are concentrated around feature-ambiguous instances.

WPML does not achieve the best result on every individual dataset. ML-KNN performs better on Business for AUC, Ranking Loss, and Coverage; PML-MA produces the best results on Enron; and NLR performs strongly on Computers, Health, and Science. NLR also obtains slightly higher Average Precision on Emotions and Reference. These dataset-specific advantages demonstrate that different data structures may favor different learning assumptions. Nevertheless, WPML achieves the most favorable average rank for every metric, indicating more stable performance across datasets rather than universal superiority on every individual task.

The critical-difference diagrams under the 20% and 40% settings exhibit a similar overall tendency. When the corruption rate increases, all methods are exposed to progressively weaker supervision. However, WPML maintains a favorable overall ranking, showing that its performance advantage is not restricted to one particular corruption level. The feature-dependent setting is more challenging than independent random flipping because annotation errors are concentrated on instances that are difficult to distinguish from their features.

5.5 Statistical Significance Analysis

We employ the Friedman test [55] to examine whether the differences among the seven algorithms are statistically significant. Let $\bar{r}_j$ denote the average rank of the j-th algorithm, K the number of algorithms, and N the number of datasets. The Friedman statistic and its Iman–Davenport correction [56] are calculated as

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}, \qquad \chi_F^2 = \frac{12N}{K(K+1)}\left[\sum_{j=1}^{K}\bar{r}_j^2 - \frac{K(K+1)^2}{4}\right]. \tag{36}$$

Ties are assigned their average ranks, and all ranks are computed from the unrounded experimental results. Table 7 reports the corrected statistics and their numerical p-values under the 60% corruption setting.

Table 7: Iman–Davenport corrected Friedman test results under 60% feature-dependent missing and noisy labels.

Metric	F_F	p-Value
AUC	11.188	3.831×10^{-9}
Ranking Loss	9.146	1.071×10^{-7}
Average Precision	13.245	1.683×10^{-10}
Coverage	5.666	5.577×10^{-5}

To report the post-hoc results numerically rather than only through critical-difference diagrams, Table 8 lists the Bonferroni-adjusted two-sided p-values for comparisons between WPML and each competing method.

Table 8: Bonferroni-adjusted p-values for post-hoc comparisons with WPML under 60% feature-dependent missing and noisy labels.

Comparison	AUC	Ranking Loss	Average Precision	Coverage
WPML vs. NLR	0.649916	1.000000	1.000000	0.031722
WPML vs. PML-MA	5.6950×10^{-6}	2.0725×10^{-4}	4.2723×10^{-4}	2.9857×10^{-4}
WPML vs. ML-KNN	0.001673	0.004339	2.0725×10^{-4}	0.004339
WPML vs. PML-LD	8.7274×10^{-6}	1.0776×10^{-5}	6.0717×10^{-4}	6.6524×10^{-5}
WPML vs. PML-NI	0.000722	0.001970	2.4897×10^{-4}	0.002715
WPML vs. PML-LENF	0.000118	2.9857×10^{-4}	7.2198×10^{-4}	0.001016

All four omnibus p-values in Table 7 are smaller than 0.05, rejecting the null hypothesis that the compared algorithms have equivalent performance. We further employ the Bonferroni–Dunn test with WPML as the control method. With $K = 7$ algorithms and $N = 15$ datasets, the critical difference is $CD = 2.0809$ at the 0.05 significance level. As shown in Figs. 1–3, under the 60%, 40%, and 20% settings, WPML significantly outperforms ML-KNN, PML-LD, PML-NI, PML-MA, and PML-LENF in terms of AUC, Ranking Loss, and Average Precision. Although WPML achieves a better average rank than NLR, their rank difference does not exceed the critical difference for these three metrics. For Coverage, WPML is significantly better than all competing methods.

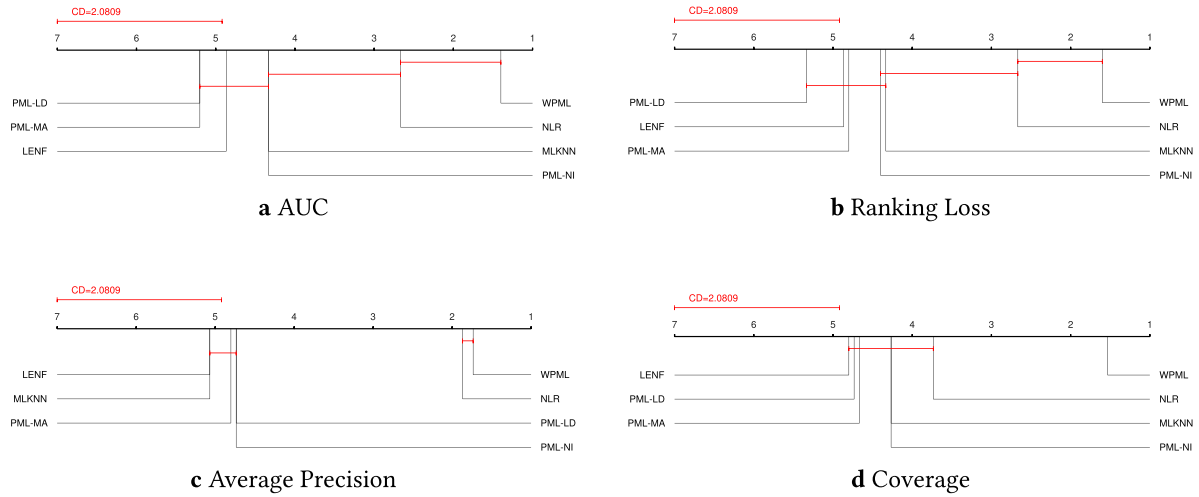
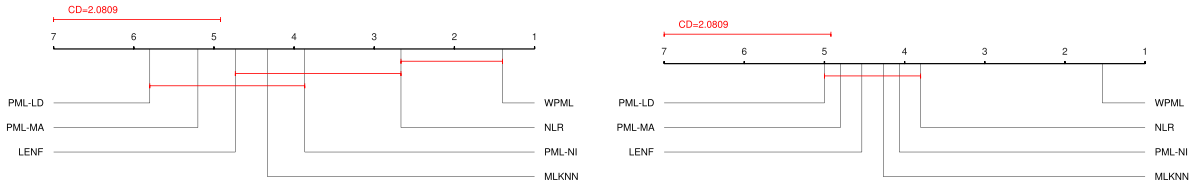
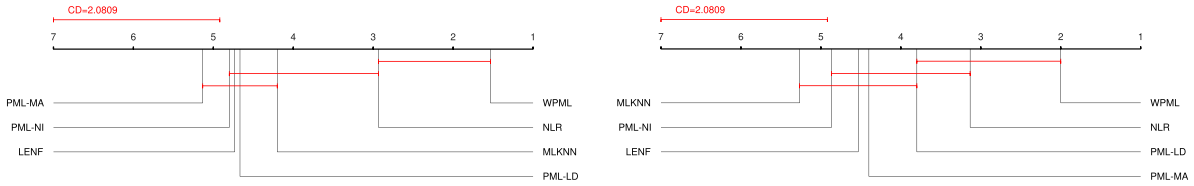


Figure 1: Comparison of WPML with the competing algorithms under 60% missing-label and label-noise rates using the Bonferroni–Dunn test ($CD = 2.081$ at the 0.05 significance level). Algorithms not connected to WPML in the CD diagrams are considered to perform significantly differently from the control method.



a AUC **b** Coverage

Figure 2: Comparison of WPML with the competing algorithms under 40% missing-label and label-noise rates using the Bonferroni–Dunn test (CD = 2.081 at the 0.05 significance level).

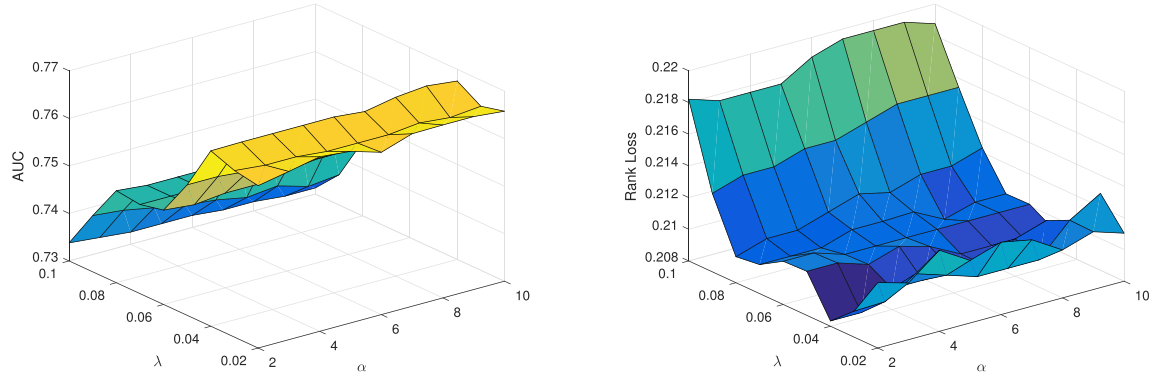


a Ranking Loss **b** Average Precision

Figure 3: Comparison of WPML with the competing algorithms under 20% missing-label and label-noise rates using the Bonferroni–Dunn test (CD = 2.081 at the 0.05 significance level).

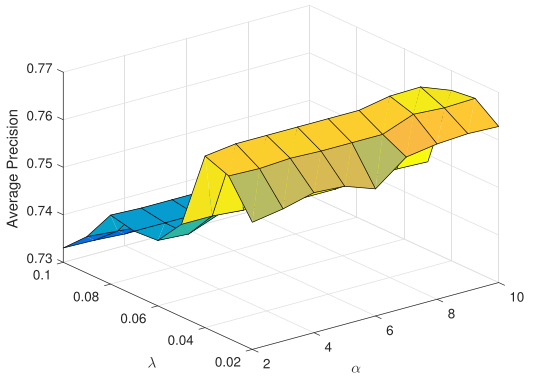
5.6 Sensitivity Analysis

We investigate the effects of λ, α, β, and γ by varying one pair of parameters while keeping the remaining parameters fixed. The enlarged sensitivity plots are presented in Figs. 4 and 5. For visualization, the plotted grids are $\lambda \in \{0.02, 0.04, 0.06, 0.08, 0.10\}$, $\alpha \in \{2, 4, 6, 8, 10\}$, $\beta \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and $\gamma \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$. The Coverage panels display the original unnormalized coverage values, whereas the comparison tables report normalized Coverage; this rescaling does not change the parameter trends.

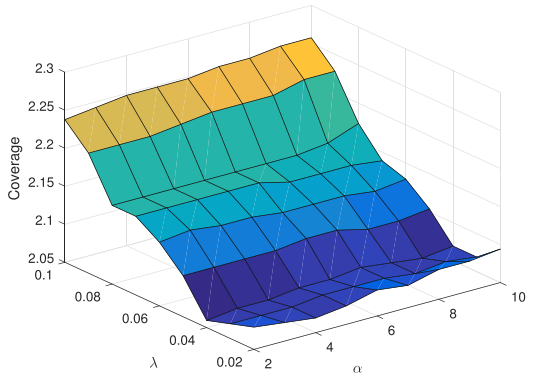


a AUC **b** Ranking Loss

Figure 4: (Continued)

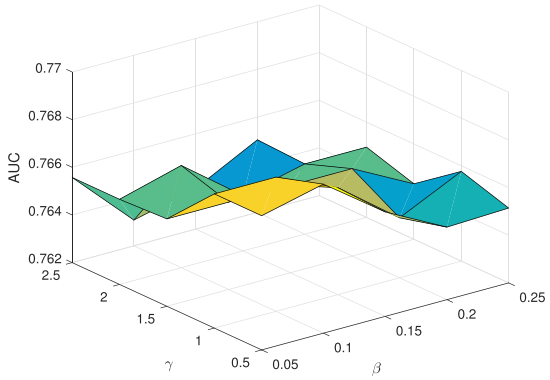


c Average Precision

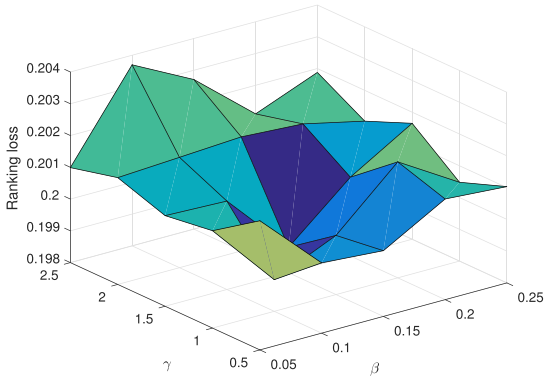


d Coverage

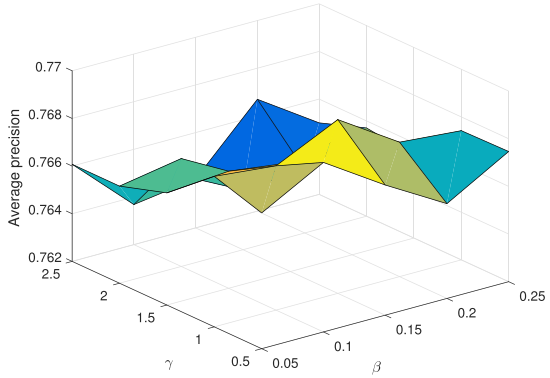
Figure 4: Sensitivity analysis of WPML with respect to parameters λ and α on the emotions dataset.



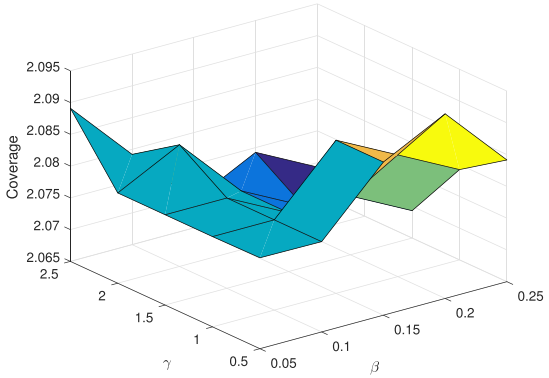
a AUC



b Ranking Loss



c Average Precision



d Coverage

Figure 5: Sensitivity analysis of WPML with respect to parameters β and γ on the emotions dataset.

The parameter λ controls the contribution of graph-based label consistency. A very small value cannot adequately preserve local relationships among instances, whereas an excessively large value may overemphasize graph smoothness. The parameter α controls the complexity of the compact feature projection. Parameters β and γ regulate ambiguous-feature selection and sparse noisy-label estimation, respectively.

The results show that WPML remains relatively stable within broad intermediate parameter ranges, while extreme values may degrade performance. In practical applications, these parameters can be selected through cross-validation on the available training data.

5.7 Ablation Study

To investigate the contribution of each component, four variants are evaluated under the same settings as the full model. In **w/o ISG**, the learned adaptive weights are replaced by a fixed unweighted symmetric k-NN graph, while graph propagation is retained. In **w/o CFLC**, the factorized predictor **TP** and its orthogonality constraint are replaced by a direct ridge-regression predictor. In **w/o GLP**, the graph regularizer is removed by setting $\lambda = 0$. In **w/o AFI**, $\mathbf{Y}_0$ and $\mathbf{W}$ are removed, so the model no longer separates feature-dependent false-positive noise. Due to space limitations, only AUC and Ranking Loss are reported in Tables 9 and 10.

Table 9: Ablation results of WPML in terms of AUC on all datasets under 60% missing-label and label-noise rates. The best results are highlighted in bold (higher is better).

Datasets	w/o ISG	w/o CFLC	w/o GLP	w/o AFI	WPML
Bibtex	0.669	0.656	0.640	0.635	**0.681**
Birds	0.774	0.770	0.756	0.752	**0.795**
Business	0.484	0.478	0.474	0.462	**0.506**
CAL500	0.730	0.717	0.705	0.695	**0.742**
Computers	0.583	0.579	0.572	0.547	**0.605**
Education	0.756	0.752	0.727	0.712	**0.771**
Emotions	0.719	0.713	0.707	0.688	**0.736**
Enron	0.451	0.444	0.429	0.413	**0.466**
Health	0.595	0.591	0.587	0.560	**0.618**
Medical	0.850	0.845	0.832	0.825	**0.872**
Reference	0.660	0.647	0.643	0.614	**0.674**
Science	0.765	0.760	0.743	0.730	**0.780**
Social	0.741	0.722	0.718	0.711	**0.756**
Scene	0.796	0.779	0.775	0.751	**0.808**
Yeast	0.747	0.732	0.728	0.722	**0.765**
Average	0.688	0.679	0.669	0.655	**0.705**

Table 10: Ablation results of WPML in terms of Ranking Loss on all datasets under 60% missing-label and label-noise rates. The best results are highlighted in bold (lower is better).

Datasets	w/o ISG	w/o CFLC	w/o GLP	w/o AFI	WPML
Bibtex	0.334	0.347	0.351	0.370	**0.317**
Birds	0.185	0.200	0.215	0.219	**0.173**
Business	0.534	0.547	0.558	0.562	**0.516**

(Continued)

Table 10 (continued)

Datasets	w/o ISG	w/o CFLC	w/o GLP	w/o AFI	WPML
CAL500	0.269	0.275	0.296	0.312	**0.255**
Computers	0.403	0.415	0.422	0.434	**0.384**
Education	0.226	0.239	0.243	0.259	**0.210**
Emotions	0.254	0.259	0.265	0.285	**0.235**
Enron	0.553	0.557	0.573	0.590	**0.534**
Health	0.395	0.411	0.422	0.426	**0.381**
Medical	0.128	0.145	0.149	0.164	**0.115**
Reference	0.324	0.340	0.349	0.362	**0.311**
Science	0.209	0.213	0.227	0.241	**0.191**
Social	0.244	0.257	0.261	0.269	**0.226**
Scene	0.177	0.184	0.196	0.214	**0.160**
Yeast	0.234	0.250	0.258	0.278	**0.223**
Average	0.298	0.309	0.319	0.332	**0.282**

The complete WPML achieves the best average AUC of 0.705 and the lowest average Ranking Loss of 0.282. Removing any component causes consistent performance degradation, demonstrating that all components contribute to the proposed framework. Specifically, ISG preserves the local manifold structure by modeling similarities among instances; therefore, removing it weakens the exploitation of neighborhood information. CFLC establishes an interaction between feature learning and label recovery, allowing the two processes to reinforce each other. Its removal results in less discriminative feature representations and less reliable label estimates. GLP propagates reliable supervision through the constructed graph to recover missing labels while alleviating the influence of noisy labels. Without GLP, the model cannot sufficiently exploit the structural relationships among instances. AFI identifies features that may introduce ambiguity into label prediction and reduces their negative influence. Consequently, w/o AFI produces the largest performance degradation, decreasing the average AUC from 0.705 to 0.655 and increasing the average Ranking Loss from 0.282 to 0.332. Overall, these results indicate that the four components perform complementary roles and jointly improve the robustness of WPML against missing and noisy labels.

5.8 Empirical Convergence Analysis

To empirically verify the convergence behavior of the proposed WPML algorithm, we report the objective function values and the relative changes of several key variables across iterations in Fig. 6. The relative change of a variable $\mathbf{Z}$ is defined as

$$\mathrm{RelChange}(\mathbf{Z}^{(t)}) = \frac{\|\mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}\|_F}{\|\mathbf{Z}^{(t)}\|_F + \varepsilon}, \tag{37}$$

where $\mathbf{Z} \in \{\mathbf{T}, \mathbf{W}, \mathbf{F}, \mathbf{Y}_0\}$ and ε is a small positive constant for numerical stability.

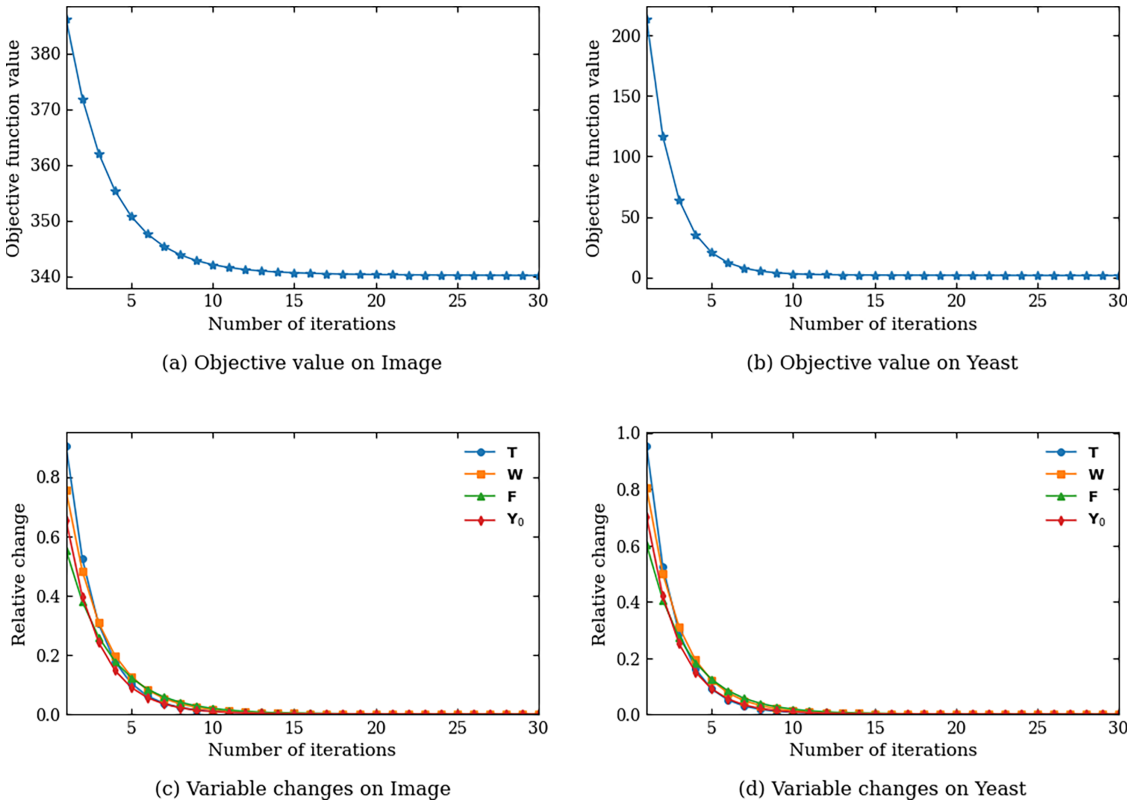


Figure 6: Convergence analysis of WPML on the scene and yeast datasets. (**a**,**b**) show the objective function values. (**c**,**d**) show the relative changes of **T**, **W**, **F**, and $\mathbf{Y}_0$ across iterations.

As shown in Fig. 6, the objective function values on both the Scene and Yeast datasets decrease rapidly during the first few iterations and then gradually become stable. This indicates that the proposed optimization procedure can effectively reduce the objective value. Meanwhile, the relative changes of **T**, **W**, **F**, and $\mathbf{Y}_0$ consistently decrease and approach zero as the iteration proceeds, suggesting that all key variables gradually reach stable states.

6 Conclusion

In this work, we introduced a novel framework for partial multi-label learning that simultaneously addresses missing and noisy labels. By combining graph-based similarity propagation with collaborative compact feature learning, our approach improves label prediction accuracy while mitigating the adverse effects of weak annotations. Specifically, an adaptive fuzzy neighborhood graph learned directly from pairwise distances enables effective label consistency propagation, facilitating the detection of noisy labels and recovery of missing ground-truth labels. Additionally, the model separates refined, label-specific features from ambiguous, noise-related ones, ensuring that only informative signals guide classifier training. Through joint optimization of feature representations and label predictions, the proposed method exhibits enhanced robustness and predictive performance. Both theoretical insights and extensive empirical evaluations demonstrate its clear advantage over existing approaches in weakly annotated multi-label learning scenarios.

Future work may focus on several directions: (1) exploring alternative structural clustering techniques to potentially improve similarity-based label propagation; (2) enhancing the scalability of the framework for large-scale, high-dimensional datasets through computational optimizations; and (3) extending the approach

to real-world applications such as healthcare, where data is often noisy and incomplete, for tasks including medical diagnosis, patient monitoring, and predictive modeling of clinical outcomes.

Acknowledgement: None.

Funding Statement: This work was supported by the Natural Science Foundation of Fujian Province under Grant 2026J010045, the Research Initiation Project of Huaqiao University under Grant 24BS114, and the Open Project Foundation of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education and Key Laboratory of Data Intelligence and Cognitive Computing of Shanxi Province under Grant CICIP2024006.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Yuzhi Tao and Anhui Tan; methodology, Yuzhi Tao; software, Yuzhi Tao; validation, Yuzhi Tao and Anhui Tan; formal analysis, Yuzhi Tao; investigation, Yuzhi Tao; resources, Yuzhi Tao; data curation, Yuzhi Tao; writing—original draft preparation, Yuzhi Tao; writing—review and editing, Anhui Tan; visualization, Yuzhi Tao; supervision, Anhui Tan; project administration, Anhui Tan; funding acquisition, Anhui Tan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The benchmark datasets used in this study are publicly available from their original repositories. The source links for the compared methods are provided in Section 5. The feature-dependent corruption-generation scripts and the implementation of WPML will be made available upon publication.

Ethics Approval: Not applicable.

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

References

1. Liu W, Wang H, Shen X, Tsang IW. The emerging trends of multi-label learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(11):7955–74. doi:10.1109/tpami.2021.3119334.
2. Xie M, Huang S. Partial multi-label learning. Proc AAAI Conf Artif Intell. 2018;32(1):4302–9. doi:10.1609/aaai.v32i1.11644.
3. Zhang M, Fang J. Partial multi-label learning via credible label elicitation. IEEE Trans Pattern Anal Mach Intell. 2021;43(10):3587–99. doi:10.1109/tpami.2020.2985210.
4. Li C, Ni P, Zhao S, Chen H. Partial label learning via conditional-label-aware disambiguation. J Comput Sci Technol. 2021;36(3):502–14. doi:10.1007/s11390-021-0992-x.
5. Zhang M, Zhou Z. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognit. 2007;40(7):2038–48.
6. Zhang M, Zhou Z. Multi-label neural networks with applications to functional genomics and text categorization. IEEE Trans Knowl Data Eng. 2006;18(10):1338–51. doi:10.1109/tkde.2006.162.
7. Lu X, Long J, Zhang H, Xie W, Zhao L, Ye Y, et al. Partial multi-view incomplete multi-label learning network with quality-aware representation fusion. IEEE Trans Circuits Syst Video Technol. 2025;35(11):11186–99. doi:10.1109/tcsvt.2025.3570702.
8. Tan A, Xu J, Wu W, Ding W, Liang J. Partial multilabel learning via dynamic fuzzy aggregations of multigranularity features. IEEE Trans Fuzzy Syst. 2025;33(9):3156–67. doi:10.1109/tfuzz.2025.3584340.
9. Song T, Bai S, Yang F, Gao C, Chen H, Li J. Exploring hybrid contrastive learning and scene-to-label information for multilabel remote sensing image classification. IEEE Trans Geosci Remote Sens. 2024;62:1–14. doi:10.1109/tgrs.2024.3422031.
10. Liu C, Ge Z, He M, Han X. A label uncertainty-guided multi-stream model for disease screening. IEEE J Biomed Health Inform. 2022;26:1245–56. doi:10.1109/isbi52829.2022.9761483.
11. Wu F, Wang Z, Zhang Z, Yang Y, Luo J, Zhu W, et al. Weakly semi-supervised deep learning for multi-label image annotation. IEEE Trans Big Data. 2015;1(3):109–22. doi:10.1109/tbdata.2015.2497270.

12. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. 2023;34(11):8135–53. doi:10.1109/tnnls.2022.3152527.
13. Xie MK, Huang SJ. Partial multi-label learning with noisy label identification. IEEE Trans Pattern Anal Mach Intell. 2022;44(7):3676–87. doi:10.1109/tpami.2021.3059290.
14. Chen JY, Li SY, Huang SJ, Chen S, Wang L, Xie MK. UNM: a universal approach for noisy multi-label learning. IEEE Trans Knowl Data Eng. 2024;36(9):4968–80.
15. Ma J, Chow TW. Label-specific feature selection and two-level label recovery for multi-label classification with missing labels. Neural Netw. 2019;118(7):110–26. doi:10.1016/j.neunet.2019.04.011.
16. Zhu Y, Kwok J, Zhou Z. Multilabel learning with global and local label correlation. IEEE Trans Knowl Data Eng. 2018;30(6):1081–94. doi:10.1109/tkde.2017.2785795.
17. Zhang M, Yu F, Tang C. Disambiguation-free partial label learning. IEEE Trans Knowl Data Eng. 2017;29(10):2155–67. doi:10.1109/tkde.2017.2721942.
18. Gibaja E, Ventura S. A tutorial on multilabel learning. ACM Comput Surv. 2015;47(3):1–38. doi:10.1145/2716262.
19. Sun L, Du W, Ding W, Long Q, Xu J. Granular ball-based fuzzy multineighborhood rough set for feature selection via label enhancement. Eng Appl Artif Intell. 2025;145(10):110191. doi:10.1016/j.engappai.2025.110191.
20. Yan Y, Li S, Feng L. Partial multi-label learning with mutual teaching. Knowl-Based Syst. 2021;212(8):106624. doi:10.1016/j.knosys.2020.106624.
21. Xie M, Huang S. Semi-supervised partial multi-label learning. Proc IEEE Int Conf Data Min. 2020:691–700.
22. Li Z, Jia Y, Yu M, Miao Z. Calibrated disambiguation for partial multi-label learning. Proc AAAI Conf Artif Intell. 2025;39(17):18620–8. doi:10.1609/aaai.v39i17.34049.
23. Wang C, Zhang Y, An S, Deng T. Adaptive feature selection based on fuzzy rough set fusion model with class variance. Pattern Recognit. 2026;170(12):112014. doi:10.1016/j.patcog.2025.112014.
24. Zhong J, Shang R, Zhao F, Zhang W, Xu S. Negative label and noise information guided disambiguation for partial multi-label learning. IEEE Trans Multim. 2024;26(1):9920–35. doi:10.1109/tmm.2024.3402534.
25. Wang C, Wang Y, Deng T, Huang Y. A nonlinear multi-label learning model based on tanh mapping. Eng Appl Artif Intell. 2023;126(2):106837. doi:10.1016/j.engappai.2023.106837.
26. Han Q, Hu L, Gao W. Integrating label confidence-based feature selection for partial multi-label learning. Pattern Recognit. 2025;161:111281. doi:10.1016/j.patcog.2024.111281.
27. Hang JY, Zhang ML. Partial multi-label learning via label-specific feature corrections. Sci China Inf Sci. 2025;68(3):132104. doi:10.1007/s11432-023-4230-2.
28. Jalali H, Kasneci G. Multilabel classification for entry-dependent expert selection in distributed Gaussian processes. Entropy. 2025;27(3):307. doi:10.3390/e27030307.
29. Yang F, Jia Y, Liu H, Dong Y, Hou J. Noisy label removal for partial multi-label learning. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2024:3724–35.
30. Qian W, Tu Y, Huang J, Shu W, Cheung YM. Partial multilabel learning using noise-tolerant broad learning system with label enhancement and dimensionality reduction. IEEE Trans Neural Netw Learn Syst. 2025;36(2):3758–72. doi:10.1109/tnnls.2024.3352285.
31. Li RJ, Ma YC, Chen H, Yang XF, Xing ZW. Coordinate descent for top-*k* multi-label feature selection with pseudo-label learning and manifold learning. Neurocomputing. 2025;658:131640. doi:10.2139/ssrn.5242690.
32. Bucak SS, Jin R, Jain AK. Multi-label learning with incomplete class assignments. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition; 2011 Jun 20–25; Colorado Springs, CO, USA. p. 2801–8.
33. Yu HF, Jain P, Kar P, Dhillon I. Large-scale multi-label learning with missing labels. Proc Int Conf Mach Learn. 2014;32(1):593–601. doi:10.1007/s13042-026-03136-y.
34. Chen G, Song Y, Wang F, Zhang C. Semi-supervised multi-label learning by solving a Sylvester equation. In: Proceedings of the 2008 SIAM International Conference on Data Mining; 2008 Apr 24–26; Atlanta, GA, USA. p. 410–9.
35. Jain V, Modhe N, Rai P. Scalable generative models for multi-label learning with missing labels. Proc Int Conf Mach Learn. 2017;70(6):1636–44. doi:10.1007/s13042-026-03136-y.

36. Kong X, Wu Z, Li LJ, Zhang R, Yu PS, Wu H, et al. Large-scale multi-label learning with incomplete label assignments. In: Proceedings of the 2014 SIAM International Conference on Data Mining; 2014 Apr 24–26; Philadelphia, PA, USA. p. 920–8.
37. Wang C, Wang Y, Deng T, Ding W. Missing multi-label learning based on the fusion of two-level nonlinear mappings. Inf Fusion. 2024;103(9):102105. doi:10.1016/j.inffus.2023.102105.
38. Jiang L, Yu GX, Guo MZ, Wang J. Feature selection with missing labels based on label compression and local feature correlation. Neurocomputing. 2020;395(8):95–106. doi:10.1016/j.neucom.2019.12.059.
39. Yang H, Zhou JT, Cai J. Improving multi-label learning with missing labels by structured semantic correlations. Lect Notes Comput Sci. 2016;9905(3):835–51. doi:10.1007/978-3-319-46448-0_50.
40. Wu B, Jia F, Liu W, Ghanem B, Lyu S. Multi-label learning with missing labels using mixed dependency graphs. Int J Comput Vis. 2018;126(8):875–96. doi:10.1007/s11263-018-1085-3.
41. Braytee A, Liu W, Anaissi A, Kennedy PJ. Correlated multi-label classification with incomplete label space and class imbalance. ACM Trans Intell Syst Technol. 2019;10(5):1–26. doi:10.1145/3342512.
42. Yin T, Chen H, Wang Z, Liu K, Yuan Z, Horng SJ, et al. Feature selection for multilabel classification with missing labels via multi-scale fusion fuzzy uncertainty measures. Pattern Recognit. 2024;154(1):110580. doi:10.1016/j.patcog.2024.110580.
43. Dai J, Li M, Zhang C. Multi-label feature selection with missing labels by weak-label fusion fuzzy discernibility pair. Inf Fusion. 2025;117(3):102921. doi:10.1016/j.inffus.2024.102921.
44. Sun L, Yin T, Ding W, Qian Y, Xu J. Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy. IEEE Trans Fuzzy Syst. 2021;30(5):1197–211. doi:10.1109/tfuzz.2021.3053844.
45. Sun L, Lyu G, Feng S, Huang X. Beyond missing: weakly-supervised multi-label learning with incomplete and noisy labels. Appl Intell. 2021;51(3):1552–64.
46. Ding J, Zhang Y, Jia L, Fu X, Jiang Y. Noisy feature decomposition-based multi-label learning with missing labels. Inf Sci. 2024;662:120228. doi:10.1016/j.ins.2024.120228.
47. Wei T, Guo L, Li Y, Gao W. Learning safe multi-label prediction for weakly labeled data. Mach Learn. 2018;107(4):703–25. doi:10.1007/s10994-017-5675-z.
48. Fang Q, Xiang C, Duan J, Soufiyan B, Shao C, Yang X, et al. OMAL: a multi-label active learning approach from data streams. Entropy. 2025;27:363.
49. Nie F, Xu D, Tsang I, Zhang C. Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. IEEE Trans Image Process. 2010;19(7):1921–32.
50. Liu G, Li Q, Yang X, Xing Z, Ma Y. Partial multi-label feature selection based on label matrix decomposition. Neural Comput Appl. 2025;37(6):4207–27. doi:10.1007/s00521-024-10822-x.
51. Yu Z, Zhang M. Multi-label classification with label-specific feature generation: a wrapped approach. IEEE Trans Pattern Anal Mach Intell. 2022;44:5199–210. doi:10.1109/tpami.2021.3070215.
52. Chen Y, Lv W, Huang Y, Fang X, Wen J, Xu Y, et al. Feature-label modal alignment for robust partial multi-label learning. arXiv:2604.09064. 2026.
53. Xu N, Liu YP, Geng X. Partial multi-label learning with label distribution. Proc AAAI Conf Artif Intell. 2020;34(4):6510–7. doi:10.1609/aaai.v34i04.6124.
54. Chen Y, Wu Y, Han N, Fang X, Chen B, Wen J. Partial multi-label learning based on near-far neighborhood label enhancement and nonlinear guidance. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, Australia. p. 3722–31.
55. Friedman M. A comparison of alternative tests of significance for the problem of *m* rankings. Ann Math Stat. 1940;11(1):86–92. doi:10.1214/aoms/1177731944.
56. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.