PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

# Feature Selection via Information Decomposition and Orthogonality Constraint for Partial Multi-Label Data

**Yao Zhang , Jun Tang***
School of Electronic and Information Engineering, Anhui University, Hefei 230039, China
{ayunxiaobao, tangjunahu}@163.com

## Abstract

Partial Multi-Label Feature Selection (PMLFS) integrates partial multi-label learning and Multi-Label Feature Selection (MLFS) within a unified optimization framework. Owing to the presence of noisy labels, it is more practical and challenging compared to traditional MLFS. However, existing methods often focus solely on label disambiguation while neglecting to distinguish the contributions of features to ground-truth labels and noisy labels. To address this, we propose a novel MLFS method based on information decomposition and orthogonality constraints, named PML-IDOC. The method partitions partial label information into ground-truth label information and noisy label information, imposing a local orthogonality constraint between ground-truth and noisy labels to prevent the latter from interfering with the former. Unlike previous approaches that constrain the low-rank property and sparsity of noisy labels, our method constructs separate mapping relationships from instances to ground-truth labels and noisy labels. By enforcing global orthogonality between the two coefficient matrices, it ensures low correlation between them, thus achieving the decoupling of feature contributions to ground-truth and noisy labels. Extensive experimental results demonstrate the effectiveness of PML-IDOC. The code is available at https://github.com/yunbao520/PML-IDOC

## 1 Introduction

In the context of Multi-Label Learning (MLL) [Wang *et al.*, 2024] tasks, which are pervasive in practical applications, each instance is inherently associated with multiple class labels. However, acquiring accurate label annotations incurs substantial costs. By contrast, obtaining a set of candidate labels that encompasses the ground-truth labels is comparatively feasible. As an emerging weak-supervised machine learning framework, Partial Multi-Label Learning (PML) [Li *et al.*, 2025b; Hang and Zhang, 2023] extends MLL and partial label learning (PLL) into a unified optimization framework to address these challenges. The fundamental assumption of PML is that each instance is associated with a candidate label set containing its ground-truth labels (e.g., Fig. 1), where the number of ground-truth labels remains unknown. This characteristic renders PML inherently more complex and challenging than traditional MLL and PLL.

*Corresponding Author

**Candidate labels:**
- **House**
- **Tree**
- **Beach**
- **Cloud**
- **Mountain**
- **Animal**

Figure 1: An example of PML data. Among the candidate labels, House, Tree, and Cloud are ground-truth labels, while Beach, Mountain, and Animal are noisy labels.

Similar to traditional MLL, high-dimensional data features in PML provide abundant supervisory information but also introduce the curse of dimensionality [Komeili *et al.*, 2017; Yu *et al.*, 2022]. Although research on traditional MLFS [Li *et al.*, 2025a; Zhang *et al.*, 2025] is relatively mature with a plethora of solutions, the ambiguous labels in PML may mislead algorithms into learning incorrect feature-label relationships, thus failing to effectively select highly representative feature subsets. Therefore, it is essential to design feature selection methods specifically for PML. Formally, for a given partial multi-label training set $D = \{X, Y\}$, where $X \in \mathrm{R}^{n\times d}$ is an instance matrix composed of $n$ training instances with $d$-dimensional features, and $Y \in \{0,1\}^{n\times m}$ is a partial label matrix consisting of $m$ labels associated with the $n$ instances, $Y_{ij} = 1$ indicates that the $j$-th label is one of the candidate labels for the $i$-th instance. The objective of PMLFS is to perform label disambiguation and select the most representative $k$-dimensional feature subset from the $d$-dimensional features.

Existing mainstream PMLFS methods either decompose $Y$ into a ground-truth label matrix $T \in \mathrm{R}^{n\times m}$ and a noisy label matrix $R \in \mathrm{R}^{n\times m}$ [Sun *et al.*, 2024], or decompose the feature weight matrix into two sub-matrices corresponding to $T$ and $R$ [Wang *et al.*, 2022], respectively, based on the feature-partial label mapping relationships. Then, they usually impose low-rank and sparse constraints on either $R$ or the fea-

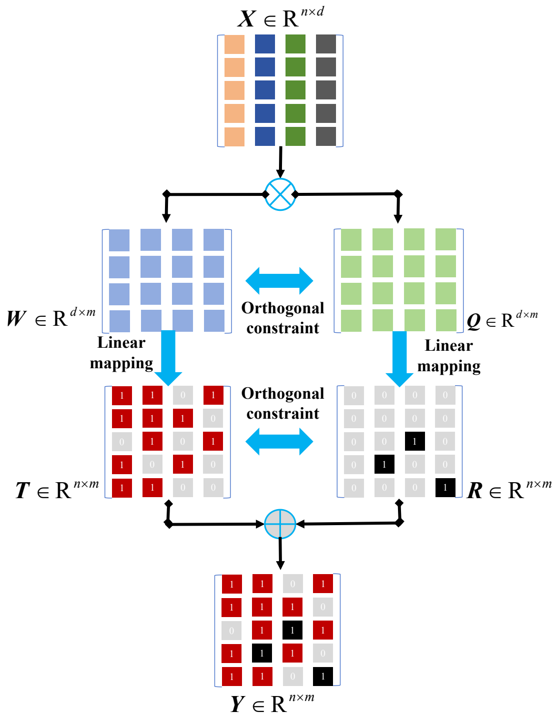


Figure 2: The overall framework of PML-IDOC. First, under the label decomposition assumption, $X$ is mapped to $T$ and $R$ respectively. Subsequently, label disambiguation is achieved via local orthogonality constraint between $T$ and $R$, facilitating the learning of feature weight matrices $W \in \mathrm{R}^{d\times m}$ and $Q \in \mathrm{R}^{d\times m}$ specific to $T$ and $R$. Additionally, a global orthogonality constraint between $W$ and $Q$ is imposed to minimize their correlation. Finally, feature selection is performed using the feature weight matrix $W$ associated with $T$.

ture weight sub-matrix associated with $R$. Additionally, there are some other methods that attempt to leverage the clean supervisory information in the feature space to constrain $T$ [Hao *et al.*, 2023]. Unfortunately, the above-mentioned approaches have overlooked a fundamental latent fact in PML: $Tr(T^T R) = 0$, where $Tr()$ denotes the trace of a matrix, and $T^T$ is the transpose of $T$. This mathematical property stems from the orthogonality between $T$ and $R$, as the label decomposition assumption [Sun *et al.*, 2024] posits that no positive label overlap exists between $T$ and $R$ for any instance.

To address the aforementioned issue, we propose a novel PMLFS method based on information decomposition and orthogonality constraint. The overall framework of PML-IDOC is illustrated in Fig. 2. First, under the label decomposition assumption, we establish linear mappings for feature-ground-truth labels and feature-noisy labels, respectively. Subsequently, label disambiguation is achieved through a local orthogonality constraint between $T$ and $R$, while a global orthogonality constraint is imposed between $W$ and $Q$ to minimize the correlation between the two mappings. Our main contributions are as follows:

- Based on the label decomposition assumption, we design a novel local orthogonality constraint term that achieves its objective by minimizing the trace of the product of ground-truth and noisy label matrices. Unlike the implicit regularization in traditional methods, the orthogonality property of this approach enables strict decoupling of semantic labels, specifically characterizing the independence of label subsets, and reducing the interference of noisy labels on ground-truth label.
- We construct mapping models for feature-ground-truth labels and feature-noisy labels, respectively, and design a global orthogonality constraint between the two feature weight matrices. By minimizing the correlation between them, we achieve the decoupling of feature contributions to the two label subsets.
- We propose a novel PMLFS method, design its optimization scheme, and theoretically prove its convergence. Extensive experiments on multiple datasets demonstrate that PML-IDOC outperforms state-of-the-art methods in most cases.

## 2 Related Work

### 2.1 Partial Multi-Label Learning

As a cutting-edge branch of MLL, PML focuses on addressing the issues of incompleteness and uncertainty in labeling in real-world scenarios. It breaks through the limitation that traditional MLL requires samples to be associated with all ground-truth labels, and allows each sample to correspond to a candidate label set containing ground-truth labels, which is more suitable for the partial labeling dilemma arising from expertise limitations or data collection costs in scenarios such as image recognition [Pham *et al.*, 2022] and ecoinformatics [Yilmaz *et al.*, 2021]. A straightforward strategy is to treat candidate labels as ground truth labels, and use state-of-the-art MLL methods to address the issue of PML. However, noisy labels often cause these approaches to learn incorrect class boundary features, which can affect the performance of classifiers.

Currently, mainstream PML methods achieve label disambiguation by identifying ground-truth labels from candidate labels using smoothing assumption, low-rank constraint, and sparse constraint. As a seminal study, Xie et al. proposed a method based on the smoothness assumption. By constraining the consistency of the underlying structure between instance similarity and label similarity, they designed an alternating iterative algorithm to achieve joint optimization of candidate label confidences and prediction models [Xie and Huang, 2018]. Subsequently, Wang et al. proposed a two-stage approach: first, they learned label confidences using feature manifold information, and then enhanced the model's capability to fit these confidences through gradient optimization [Wang *et al.*, 2019]. Sun et al. proposed an approach based on the label decomposition assumption, which deconstructs the partial label matrix into a ground-truth label matrix and a noisy label matrix, and constrains the former to satisfy the low-rank property and the latter to exhibit the sparse property [Sun *et al.*, 2019]. Similarly, Li et al. constructed a model based on the label decomposition assumption, integrating low-rank and sparse constraints. Uniquely, they introduced the Wasserstein distance incorporating label correlations as a loss function to measure the discrepancy between ground-truth confidences and predictions, thereby facilitating precise estimation of label correlations [Li *et al.*, 2024]. Unlike the aforementioned approaches, Xie et al. decomposed the feature weight matrix and imposed constraints via trace norm and $L_1$ norm regularization to achieve simultaneous optimization of ground-truth label recovery and noisy label identification [Xie and Huang, 2021].

## 2.2 Embedded Approaches for Feature Selection

The embedded approach, one of the mainstream techniques in MLFS, essentially aims to establish a deep coupling between feature selection and model learning, enabling automatic optimization and screening of feature weights during training. Traditional embedded methods mainly construct models based on linear regression frameworks or information theory, and then constrain the model by introducing regularization terms with different semantics (such as manifold learning and sparse constraints). For example, Hu et al. introduced labeled manifolds, feature manifolds, and $L_{2,1}$ norm sparsity constraints to impose full constraints on the feature weight matrix [Hu *et al.*, 2020]. Li et al. designed a flexible sparse paradigm to address feature redundancy and combined it with feature manifolds to constrain the feature selection process [Li *et al.*, 2023]. Hashemi et al. modeled the feature selection process as a multi-criteria decision-making framework, calculated the weight of each label by information entropy theory, and embedded them into the TOPSIS algorithm to achieve feature selection [Hashemi *et al.*, 2020].

In PML scenarios, traditional methods are prone to misclassify spurious correlations between noisy labels and features as genuine associations, which in turn interferes with the feature selection process. To this end, Wang et al. proposed the seminal method in the field of PMLFS, which first decomposes the feature weight matrix learned by linear mapping and then realizes label disambiguation using label correlation and the sparsity of noisy labels [Wang *et al.*, 2022]. Hao et al. constructed a joint framework for label subspace and feature subspace based on linear weighting theory. Through the collaborative optimization of label matrix decomposition and feature manifold learning, they effectively eliminated the impact of noisy labels [Hao *et al.*, 2023]. Similarly, Pan et al. proposed a latent structure alignment constraint between feature subspace and label subspace for label disambiguation. Additionally, they introduced the OPTICS clustering technique to adaptively determine the optimal dimension of the latent space [Pan *et al.*, 2025]. Sun et al. decomposed the partial label matrix into a low-rank ground-truth label matrix and a sparse noisy label matrix based on the label decomposition assumption. They then introduced an instance manifold to constrain the latent manifold structure of the ground-truth label [Sun *et al.*, 2024]. Han et al. proposed a joint optimization mapping model for features and labels, which achieves precise evaluation of label confidences by fusing the average distance between instances with the same label and the distance from instances to cluster centers [Han *et al.*, 2025].

Notably, while the above PMLFS methods have achieved certain effects by imposing low-rank or sparse constraints on the matrices representing ground-truth and noisy labels, they overlook a critical fact: the orthogonality between ground-truth and noisy labels. This orthogonality is rooted in the label decomposition assumption, which posits that there is no semantic overlap between ground-truth and noisy labels, within the positive label space. However, existing methods fail to explicitly embed this potential geometric constraint into the model optimization objective, leading to the retention of potentially pseudo-correlated features induced by noisy labels during feature selection. This, in turn, affects the thoroughness of label disambiguation and the precision of feature screening.

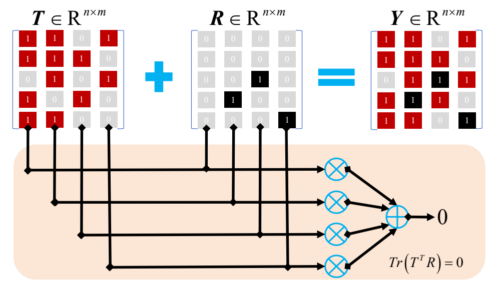


Figure 3: In PML, an intrinsic property is that $Tr(T^T R) = 0$. The local orthogonality constraint can detect and eliminate the interference of $R$ with $T$, because there is no positive label overlap between $T$ and $R$ under the label decomposition assumption.

# 3 The Proposed Method

In the PML scenario, each instance is associated with a set of candidate labels that include the ground-truth labels. As a core theoretical foundation, the label decomposition assumption posits that the candidate label set can be deconstructed into a ground-truth label subset and a noisy label subset, i.e., $Y = T + R$ [Sun *et al.*, 2024; Liu *et al.*, 2025]. This assumption enables the construction of more precise feature screening mechanisms by explicitly distinguishing the associative patterns of ground-truth labels and noisy labels. Existing approaches assume that $T$ is a low-rank matrix and $R$ is a sparse matrix [Sun *et al.*, 2019]. However, this assumption characterizes the label structure only at the level of statistical properties, overlooking the inherent local geometric orthogonality between them. As shown in Fig. 3, by constraining the local orthogonality between $T$ and $R$, we can achieve the thorough decomposition of the partial label matrix, effectively mitigating the interference of noisy labels on ground-truth labels. In addition, studies have shown that continuous labels can break through the semantic limitations of traditional binary labels, and compared with binary labels, continuous labels can provide richer label information [Hou *et al.*, 2016]. Therefore, we use continuous labels to define the local orthogonality constraint in Fig. 3 as follows:

$$\min_{T,R} Tr(T^T R), \quad s.t. \ Y = T + R, \{T, R\} \geq 0. \tag{1}$$

The term $\min_{T,R} Tr(T^T R)$ in Eq. (1) formalizes the local orthogonality between $T$ and $R$, enforcing their semantic disjointness by minimizing their inner product directly. Moreover, slight overlap between positive labels of $T$ and $R$ in practice does not violate the core logic of label decomposition, since continuous $T$ and $R$ in Eq. (1) characterize label correlation via non-negative values and avoid excessive penalty. However, Eq. (1) optimizes only label information without explicit feature guidance. Thus, we extend it by introducing linear weighting-based feature learning to:

$$\begin{aligned}&\min_{T,R,W,Q} \|XW-T\|_F^2+\|XQ-R\|_F^2+2\gamma Tr(T^TR),\\&s.t.\ Y=T+R,\{T,R\}\geq 0,\end{aligned} \tag{2}$$

where $\gamma$ is the regularization parameter. Eq. (2) maps the instances to the ground-truth and noisy label spaces by introducing the feature weight matrices $W$ and $Q$, respectively. It focuses on balancing two main objectives: minimizing the reconstruction error of the feature representation for $R$ on $T$, and penalizing their non-orthogonality. This joint optimization mechanism ensures that the feature learning for ground-truth labels is not only discriminative but also geometrically decoupled from features induced by noisy labels, thereby significantly enhancing the precision of feature screening.

To further strengthen the decoupling between ground-truth and noisy labels in the feature space, we introduce a global orthogonality constraint on the feature weight matrices, extending the original framework to Eq. (3).

$$\begin{aligned}&\min_{T,R,W,Q} \|XW-T\|_F^2+\|XQ-R\|_F^2+\\&\beta\|W^TQ\|_F^2+2\gamma Tr(T^TR),\\&s.t.\ Y=T+R,\{T,R\}\geq 0,\end{aligned} \tag{3}$$

where $\beta$ is the regularization parameter. Specifically, while retaining the feature reconstruction error term and label local orthogonality penalty term, Eq. (3) introduces a novel feature weight global orthogonality term $\|W^TQ\|_F^2$. This term enforces geometric orthogonality of the mapping spaces for the two types of labels at the fundamental level of the feature transformation mechanism by directly constraining the inner product norm of $W$ and $Q$, imposing dual constraints on the label decomposition process: ensuring semantic mutual exclusivity through the local orthogonality constraint, and achieving explicit decoupling of feature mapping spaces through the global orthogonality constraint. Notably, this global orthogonality constraint effectively prevents semantic information mixing between the two label types during feature learning, thereby constructing a more robust PMLFS model.

To reduce the risk of local optima caused by the strict equality constraint $Y = T + R$, we adopt a flexible constraint strategy: by introducing the decomposition error term in Frobenius norm form, the model can adaptively balance the accuracy of label decomposition and the orthogonality of feature mappings while maintaining semantic decoupling. Additionally, considering the physical non-negativity of label semantics, we impose non-negativity constraints on all variables. This constraint not only aligns with the semantic characteristics of labels in real-world scenarios but also prevents parameter oscillations during optimization.

$$\begin{aligned}&\min_{T,R,W,Q} \|XW-T\|_F^2+\|XQ-R\|_F^2+\\&\alpha\|Y-T-R\|_F^2+\beta\|W^TQ\|_F^2+2\gamma Tr(T^TR),\\&s.t.\ \{T,R,W,Q\}\geq 0.\end{aligned} \tag{4}$$

where $\alpha$ is the regularization parameter.

# 4 Optimization

To solve the optimization problem in Eq. (4), we construct the Lagrangian function $\mathcal{L}(T,R,W,Q)$ and design an iterative update algorithm by combining the Karush–Kuhn–Tucker (KKT) [Huang *et al.*, 2015] conditions. First, we introduce the Lagrange multiplier matrices $\Upsilon\geq 0$, $\Phi\geq 0$, $\Psi\geq 0$, and $\Omega\geq 0$ to transform the constrained optimization problem into an unconstrained form:

$$\begin{aligned}&\mathcal{L}(T,R,W,Q)=\|XW-T\|_F^2+\|XQ-R\|_F^2+\\&\alpha\|Y-T-R\|_F^2+\beta\|W^TQ\|_F^2+2\gamma Tr(T^TR)-\\&Tr(\Upsilon T^T)-Tr(\Phi R^T)-Tr(\Psi W^T)-Tr(\Omega Q^T),\end{aligned} \tag{5}$$

where $\Upsilon\in \mathrm{R}^{n\times m}$, $\Phi\in \mathrm{R}^{n\times m}$, $\Psi\in \mathrm{R}^{d\times m}$, and $\Omega\in \mathrm{R}^{d\times m}$.

Then, the partial derivatives of $\mathcal{L}(T,R,W,Q)$ w.r.t variables $W$, $R$, $Q$ and $T$ are:

$$\frac{\partial\mathcal{L}}{\partial W}=2(X^TXW-X^TT+\beta QQ^TW)-\Psi. \tag{6}$$

$$\frac{\partial\mathcal{L}}{\partial R}=2((\alpha+1)R-XQ-\alpha Y+(\alpha+\gamma)T)-\Phi. \tag{7}$$

$$\frac{\partial\mathcal{L}}{\partial Q}=2(X^TXQ-X^TR+\beta WW^TQ)-\Omega. \tag{8}$$

$$\frac{\partial\mathcal{L}}{\partial T}=2((\alpha+1)T-XW-\alpha Y+(\alpha+\gamma)R)-\Upsilon. \tag{9}$$

According to the KKT conditions, the optimal solution must satisfy the complementary slackness conditions, i.e., $\Upsilon\circ T=0$, $\Phi\circ R=0$, $\Psi\circ W=0$, and $\Omega\circ Q=0$, here $\circ$ denotes the Hadamard product. Consequently, we derive the following:

$$(X^TXW-X^TT+\beta QQ^TW)\circ W=0. \tag{10}$$

$$((\alpha+1)R-XQ-\alpha Y+(\alpha+\gamma)T)\circ R=0. \tag{11}$$

$$(X^TXQ-X^TR+\beta WW^TQ)\circ Q=0. \tag{12}$$

$$((\alpha+1)T-XW-\alpha Y+(\alpha+\gamma)R)\circ T=0. \tag{13}$$

Finally, leveraging non-negative matrix factorization [Lee and Seung, 1999] techniques, we derive the update rules for $W$, $R$, $Q$, and $T$ in Formula (4):

$$W_{ij}^{t+1}\leftarrow W_{ij}^t\frac{(X^TT)_{ij}}{(UW+\beta QQ^TW)_{ij}}, \tag{14}$$

$$R_{ij}^{t+1}\leftarrow R_{ij}^t\frac{(XQ+\alpha Y)_{ij}}{((\alpha+1)R+(\alpha+\gamma)T)_{ij}}, \tag{15}$$

$$Q_{ij}^{t+1}\leftarrow Q_{ij}^t\frac{(X^TR)_{ij}}{(UQ+\beta WW^TQ)_{ij}}, \tag{16}$$

$$T_{ij}^{t+1}\leftarrow T_{ij}^t\frac{(XW+\alpha Y)_{ij}}{((\alpha+1)T+(\alpha+\gamma)R)_{ij}}, \tag{17}$$

| Dataset | # Instances | # Features | # Labels | Type |
|---|---|---|---|---|
| Yeast | 2417 | 103 | 14 | Biology |
| Emotion | 593 | 72 | 6 | Music |
| Birds | 645 | 260 | 19 | Audio |
| Scene | 2407 | 294 | 6 | Image |
| CAL500 | 502 | 68 | 174 | Audio |
| Music_style | 6839 | 98 | 10 | Music |
| Mirflickr | 10433 | 100 | 7 | Image |
| Music_emotion | 978 | 1445 | 45 | Music |

Table 1: Characteristics of the eight experimental datasets.

where $U = X^T X$, $t$ denotes the number of iterations.

The above update rules can exhibit good numerical stability, ensuring that parameters always satisfy the non-negativity constraints in each iteration. Additionally, the convergence of the algorithm can be proven by constructing auxiliary functions, showing that each iteration reduces the original objective function value until converging to a local optimal solution. Due to space constraints, the pseudo code of PML-IDOC and its convergence proof are provided in Appendices A and B, respectively.

## 5 Experiments

This section will verify the effectiveness and performance advantages of PML-IDOC through extensive experiments. Due to space limitations, part of the experimental analysis is presented in Appendix C.

### 5.1 Datasets

To comprehensively validate the effectiveness of the PML-IDOC model, we conducted systematic experiments on five synthetic datasets and three real-world datasets. These datasets cover four typical domains: Biology domain dataset Yeast, Music domain datasets Emotion, Music_emotion and Music_style, Audio domain datasets Birds and CAL500, and Image domain datasets Scene and Mirflickr. Notably, Music_emotion, Music_style, and Mirflickr are real-world partial multi-label datasets. The detailed characteristics of the eight experimental datasets are shown in Tab. 1.

### 5.2 Experimental Setup

In the experimental performance comparison, we conduct a systematic comparison of PML-IDOC with seven state-of-the-art PMLFS and MLFS algorithms, specifically covering three advanced MLFS methods (RFSFS [Li *et al.*, 2023], MFS-MCDM [Hashemi *et al.*, 2020], and DRMFS [Hu *et al.*, 2020]), and the four PMLFS methods integrating different constraint mechanisms. These include the PMLFS [Wang *et al.*, 2022] method fusing sparse and low-rank constraints, the label decomposition model PML-LRS [Sun *et al.*, 2024], the latent subspace alignment algorithm PML-FSLA [Pan *et al.*, 2025], and the shared subspace method PML-FSSO [Hao *et al.*, 2023]. All algorithms are trained using five-fold cross-validation. We then evaluate the generalization performance of feature subsets selected by each method using an ML-KNN [Zhang and Zhou, 2007] classifier. Five common multi-label metrics are used: Hamming Loss (HL), Ranking Loss (RL), Average Precision (AP), Macro-$F_1$ (MAF), and Micro-$F_1$ (MIF) [Zhang and Ma, 2022]. Detailed experimental setups are provided in the appendix due to space limitations.

### 5.3 Discussion

Tabs. 2–6 present detailed experimental results for each metric. During the experiments, we select 2% to 20% (step size 2%) of the features from each dataset based on feature ranking results. Then, we conduct five repeated experiments for each algorithm to screen the optimal results. Finally, the mean and standard deviation under different feature percentages are recorded in Tabs. 2–6. Additionally, for the five synthetic datasets, we set their partial labeling rates from 10% to 50% (step size 10%). To visually compare the performance differences of each algorithm across varying feature percentages, we plot the corresponding performance curves in Fig. 4. From the overall comparative experimental results, the following observations can be drawn:

- In all cases, the top-ranked results are from the models PML-LRS and PML-IDOC, which are based on the label decomposition assumption. Notably, PML-IDOC ranks first in 82.5% of cases. This phenomenon not only validates the effectiveness of the label decomposition hypothesis but also demonstrates that the local and global orthogonality constraints in PML-IDOC are more conducive to achieving label disambiguation.
- In terms of specific metrics, PML-IDOC demonstrates significant advantages in HL, AP, and MIF. For example, in Tab. 2, PML-IDOC achieves the lowest HL across all datasets. In Tab. 4, PML-IDOC shows performance improvements ranging from 2.72% to 8.13% compared to baseline methods. Additionally, PML-IDOC remains among the top two performers in all cases for RL and MAF. These findings indicate that global and local orthogonality constraints effectively mitigate the interference of noisy labels on ground-truth labels, thereby providing more accurate supervision for feature selection.
- In terms of specific datasets, PML-IDOC ranks second in some metrics for Scene, Music_style, and Music_emotion. This is primarily because the orthogonality constraint mechanism of PML-IDOC places greater emphasis on positive label identification, optimizing feature selection by suppressing the interference of noisy labels on ground-truth labels, while assigning relatively lower priority to negative label identification.
- Fig. 4 further reveals the advantages of PML-IDOC across different feature selection ratios on the CAL500 dataset. As the selected feature ratio increases from 2% to 20%, PML-IDOC consistently achieves lower Hamming Loss and Ranking Loss, while maintaining higher AP and MAF compared to other methods. This suggests that the global orthogonality constraint between feature weight matrices $W$ and $Q$ effectively decouples feature contributions to ground-truth and noisy labels.

### 5.4 Ablation Study

To verify whether the local and global orthogonality constraints in PML-IDOC can effectively achieve label disam-

| Algorithms | RFSFS | MFS-MCDM | DRMFS | PMLFS | PML-FSSO | PML-LRS | PML-FSLA | PML-IDOC |
|---|---|---|---|---|---|---|---|---|
| Yeast | 22.28±0.59 | 22.2±0.77 | 22.26±0.82 | 21.82±0.97 | 21.86±0.61 | 21.40±0.87 | 21.56±0.75 | **21.31±0.74** |
| Emotion | 27.12±2.94 | 25.92±1.53 | 24.84±2.30 | 27.59±3.67 | 25.88±1.79 | 26.82±2.66 | 26.77±1.43 | **24.32±2.16** |
| Birds | 4.99±0.03 | 5.13±0.08 | 4.86±0.06 | 5.28±0.35 | 5.04±0.03 | 4.92±0.09 | 5.26±0.10 | **4.82±0.05** |
| Scene | 17.12±0.51 | 15.27±0.80 | 15.59±0.65 | 14.96±1.36 | 15.67±1.26 | 14.47±0.64 | 15.16±0.59 | **14.28±0.65** |
| CAL500 | 86.82±0.12 | 86.96±0.18 | 86.92±0.14 | 87.13±0.15 | 86.68±0.16 | 86.82±0.17 | 87.08±0.19 | **86.62±0.15** |
| Music_style | 82.19±0.60 | 83.25±0.14 | 81.86±0.65 | 83.72±1.38 | 81.99±0.19 | 81.57±0.45 | 84.45±0.77 | **80.98±0.58** |
| Mirflickr | 19.62±1.12 | 20.76±1.29 | 19.66±1.27 | 22.37±0.87 | 19.31±0.91 | 22.11±0.93 | 22.23±0.99 | **19.06±1.03** |
| Music_emotion | 35.29±0.84 | 36.50±0.48 | 34.49±0.74 | 37.20±0.93 | 35.75±0.45 | 35.2±0.50 | 36.78±1.39 | **34.11±0.43** |

Table 2: Experimental results (mean ± std) (%) of each method under the HL metric. (The best results are marked in bold.)

| Algorithms | RFSFS | MFS-MCDM | DRMFS | PMLFS | PML-FSSO | PML-LRS | PML-FSLA | PML-IDOC |
|---|---|---|---|---|---|---|---|---|
| Yeast | 19.73±0.50 | 18.95±0.52 | 18.74±0.70 | 18.97±0.82 | 19.07±0.79 | 18.76±0.84 | 18.83±0.64 | **18.57±0.79** |
| Emotion | 25.42±4.04 | 24.25±2.44 | 22.56±2.75 | 26.97±6.90 | 24.06±2.30 | 24.76±4.84 | 27.28±2.01 | **21.51±3.53** |
| Birds | 27.24±0.86 | 29.30±1.41 | 25.41±1.33 | 34.41±1.88 | 29.11±0.67 | 27.84±1.42 | 32.01±1.63 | **24.19±1.22** |
| Scene | 25.81±4.68 | 15.17±3.66 | 17.29±3.26 | 17.92±4.74 | 18.76±2.96 | **13.86±1.97** | 15.98±2.44 | 14.56±2.83 |
| CAL500 | 18.40±0.12 | 18.60±0.21 | 18.45±0.15 | 18.77±0.18 | 18.52±0.18 | 18.64±0.17 | 18.94±0.25 | **18.35±0.17** |
| Music_style | 22.69±0.52 | 24.26±0.07 | 22.34±0.55 | 22.45±1.01 | 22.64±0.33 | **20.91±0.62** | 21.84±0.48 | 21.27±0.26 |
| Mirflickr | 14.03±3.05 | 17.13±3.80 | 15.10±1.62 | 22.17±2.43 | 13.83±2.90 | 21.88±2.36 | 21.99±2.05 | **12.87±3.13** |
| Music_emotion | 34.46±0.16 | 36.62±0.78 | 34.28±0.57 | 33.25±0.54 | 34.71±0.21 | 33.06±0.25 | 32.38±0.42 | **32.15±0.33** |

Table 3: Experimental results (mean ± std) (%) of each method under the RL metric. (The best results are marked in bold.)

| Algorithms | RFSFS | MFS-MCDM | DRMFS | PMLFS | PML-FSSO | PML-LRS | PML-FSLA | PML-IDOC |
|---|---|---|---|---|---|---|---|---|
| Yeast | 72.49±0.82 | 73.45±0.82 | 73.89±0.97 | 73.56±1.09 | 73.32±1.00 | 73.77±1.07 | 73.81±0.97 | **74.12±1.08** |
| Emotion | 71.66±3.84 | 72.64±2.15 | 74.36±2.92 | 71.47±5.45 | 73.34±2.20 | 72.17±3.53 | 72.76±2.01 | **75.46±3.41** |
| Birds | 44.67±0.67 | 43.45±1.38 | 47.88±1.30 | 35.50±2.50 | 43.75±0.86 | 45.86±2.11 | 40.38±0.51 | **50.30±1.57** |
| Scene | 64.82±5.41 | 76.97±4.66 | 74.06±4.18 | 73.12±5.64 | 72.53±3.95 | **78.73±2.65** | 75.70±3.03 | 77.67±4.04 |
| CAL500 | 49.45±0.34 | 49.07±0.51 | 49.44±0.25 | 48.78±0.44 | 49.31±0.33 | 49.05±0.41 | 48.31±0.61 | **49.62±0.40** |
| Music_style | 66.03±0.54 | 65.25±0.11 | 66.13±0.37 | 66.07±0.71 | 65.92±0.20 | 66.24±0.32 | 66.54±0.42 | **67.00±0.13** |
| Mirflickr | 80.61±4.47 | 74.67±6.12 | 79.10±6.89 | 69.61±4.31 | 81.32±4.20 | 69.95±4.08 | 69.13±3.75 | **82.94±4.31** |
| Music_emotion | 51.52±0.21 | 50.26±0.53 | 51.71±0.44 | 52.64±0.65 | 51.43±0.21 | 52.62±0.45 | 53.24±0.39 | **53.53±0.36** |

Table 4: Experimental results (mean ± std) (%) of each method under the AP metric. (The best results are marked in bold.)

| Algorithms | RFSFS | MFS-MCDM | DRMFS | PMLFS | PML-FSSO | PML-LRS | PML-FSLA | PML-IDOC |
|---|---|---|---|---|---|---|---|---|
| Yeast | 42.98±0.27 | 42.65±0.07 | 43.10±0.27 | 42.73±0.18 | 42.96±0.19 | 42.77±0.22 | 43.21±0.16 | **43.67±0.17** |
| Emotion | 55.33±3.97 | 55.02±2.50 | 56.99±2.42 | 54.59±3.38 | 55.75±2.95 | 54.52±2.74 | 52.85±1.31 | **59.74±2.58** |
| Birds | 12.18±0.96 | 10.23±0.83 | 13.27±0.72 | 9.89±0.82 | 11.99±0.94 | 11.52±1.09 | 10.00±0.33 | **14.18±0.80** |
| Scene | 40.75±6.25 | 56.39±6.76 | 51.66±6.35 | 51.14±8.98 | 50.98±7.64 | **59.51±3.35** | 54.81±5.09 | 57.40±5.09 |
| CAL500 | 98.08±0.13 | 97.95±0.18 | 98.06±0.16 | 97.76±0.11 | 98.23±0.17 | 98.04±0.19 | 97.77±0.21 | **98.28±0.18** |
| Music_style | 22.69±0.07 | 22.55±0.02 | 22.72±0.04 | 22.63±0.08 | 22.73±0.03 | **22.87±0.10** | 22.61±0.05 | 22.86±0.10 |
| Mirflickr | 39.52±4.15 | 37.62±3.79 | 45.06±4.10 | 43.08±3.80 | 45.73±2.88 | 44.00±3.83 | 45.25±3.56 | **45.83±2.59** |
| Music_emotion | 29.25±0.36 | 23.12±1.58 | 29.60±1.12 | 30.97±0.88 | 28.99±0.95 | **32.81±1.04** | 30.72±0.68 | 32.66±0.67 |

Table 5: Experimental results (mean ± std) (%) of each method under the MAF metric. (The best results are marked in bold.)



biguation and enhance the ability of selected feature subsets to identify positive labels, we systematically designed an ablation study on the local and global orthogonality constraints. The experimental results are shown in Tabs. 7-8, where $\beta = 0$ and $\gamma = 0$ indicate the absence of global and local orthogonality constraint terms, respectively.

In all cases, PML-IDOC demonstrates optimal performance. Specifically, its performance improves by an average of 47.16% compared to PML-IDOC ($\beta = 0$, $\gamma = 0$) without orthogonality constraints. Additionally, when using local or global orthogonality constraints alone, the model shows significant performance degradation. These findings further validate that latent orthogonality constraints under the label decomposition assumption provide effective supervision for label disambiguation in PML. Unlike heuristic sparse and low-rank constraints in existing methods, the orthogonality constraints explicitly can leverage the geometric properties of the label space to achieve strict decoupling of ground-truth and noisy labels. This mechanism not only suppresses noisy label interference in feature selection but also enhances the

| Algorithms | RFSFS | MFS-MCDM | DRMFS | PMLFS | PML-FSSO | PML-LRS | PML-FSLA | PML-IDOC |
|---|---|---|---|---|---|---|---|---|
| Yeast | 60.00±2.09 | 61.29±1.98 | 62.53±1.75 | 61.71±2.13 | 61.47±1.90 | 62.36±1.86 | 62.47±1.75 | **63.13±1.81** |
| Emotion | 57.40±4.72 | 59.10±3.36 | 60.28±3.09 | 56.92±4.46 | 59.27±3.25 | 57.68±3.66 | 56.03±2.33 | **62.26±3.20** |
| Birds | 23.11±1.07 | 21.00±0.75 | 22.08±1.56 | 16.22±4.14 | 21.82±1.31 | 22.22±2.46 | 18.47±0.79 | **24.03±2.04** |
| Scene | 40.85±5.83 | 55.70±6.73 | 51.33±6.44 | 51.04±8.43 | 51.41±7.04 | **58.52±3.35** | 54.67±4.21 | 56.93±4.76 |
| CAL500 | 98.17±0.11 | 98.09±0.14 | 98.18±0.14 | 97.92±0.12 | 98.32±0.14 | 98.13±0.17 | 97.94±0.19 | **98.34±0.16** |
| Music_style | 25.56±0.08 | 25.39±0.06 | 25.68±0.13 | 25.39±0.17 | 25.61±0.07 | 25.77±0.09 | 25.29±0.10 | **25.88±0.12** |
| Mirflickr | 71.10±2.39 | 69.13±2.41 | 71.09±2.38 | 66.28±1.58 | 71.59±1.92 | 66.72±1.65 | 66.64±1.68 | **71.98±1.96** |
| Music_emotion | 42.34±0.14 | 39.56±1.03 | 42.34±0.39 | 43.32±0.52 | 42.26±0.31 | 43.25±0.23 | 43.57±0.29 | **43.82±0.24** |

Table 6: Experimental results (mean ± std) (%) of each method under the MIF metric. (The best results are marked in bold.)

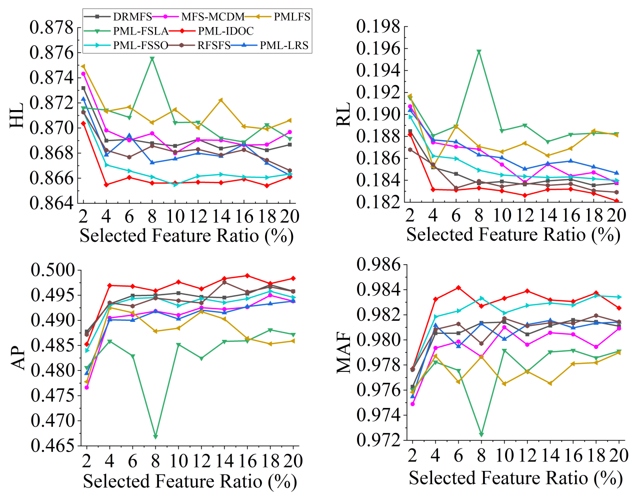


Figure 4: Performance curves of each method on CAL500 dataset in terms of HL, RL, AP and MAF.

| Algorithms | Birds | Emotion | Yeast | Mirflickr |
|---|---|---|---|---|
| PML-IDOC | **14.18** | **59.74** | **43.67** | **45.83** |
| PML-IDOC ($\beta = 0$) | 7.98 | 52.53 | 30.38 | 43.59 |
| PML-IDOC ($\gamma = 0$) | 7.43 | 51.41 | 30.48 | 40.42 |
| PML-IDOC ($\{\beta, \gamma\} = 0$) | 5.30 | 46.15 | 30.15 | 38.69 |

Table 7: Results (%) of ablative studies under the MAF metric.

| Algorithms | Birds | Emotion | Yeast | Mirflickr |
|---|---|---|---|---|
| PML-IDOC | **24.03** | **62.26** | **63.13** | **71.98** |
| PML-IDOC ($\beta = 0$) | 18.22 | 57.62 | 59.45 | 70.45 |
| PML-IDOC ($\gamma = 0$) | 14.45 | 56.84 | 59.30 | 71.19 |
| PML-IDOC ($\{\beta, \gamma\} = 0$) | 12.58 | 53.10 | 59.03 | 70.74 |

Table 8: Results (%) of ablative studies under the MIF metric.

ability of selected features to identify positive labels.

## 5.5 Parameter Analysis

In PML-IDOC, three hyperparameters $\alpha$, $\beta$, and $\gamma$ are adopted to balance the penalty strengths among different regularization terms. Fig. 5 presents the parameter analysis results of PML-IDOC on the Scene dataset in terms of the AP metric. In our experiments, we tune these parameters within the range from 0.001 to 1000 and record the corresponding

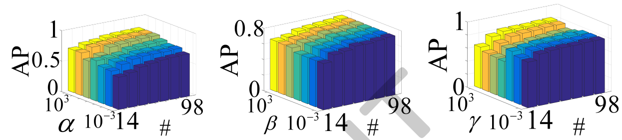


Figure 5: Parameter analysis of PML-IDOC on Scene dataset.

performance when the number of the selected features varies from 14 to 98. It can be observed that PML-IDOC maintains relatively stable performance within a limited range of each parameter. In addition, the AP value consistently rises as the number of selected features increases under all parameter settings. These results highlight the effectiveness and robustness of PML-IDOC. Furthermore, we summarize the explicit optimal range for each parameter and analyze their influence mechanisms in depth. Specifically, $\alpha$ and $\beta$ balance the label decomposition error and regulate the global orthogonal constraint on the feature weight matrix, respectively, both with an optimal range of $[10, 10^3]$. $\gamma$ controls the local orthogonal constraint in the label space, with an optimal range of $[10^2, 10^3]$. Moreover, overly small parameter values result in insufficient constraints and stronger noise interference, which lowers the feature discriminability.

## 6 Conclusion

Based on the latent orthogonality under the label decomposition assumption, this paper proposes a novel PMLFS method, PML-IDOC, to address the label disambiguation problem in PML. The method decomposes partial label information into ground-truth and noisy labels, and achieves the decoupling of feature contributions to the two types of labels by imposing local and global orthogonality constraints, thereby significantly improving the accuracy of feature screening. Extensive experimental results show that PML-IDOC outperforms existing methods in various complex scenarios, fully verifying its effectiveness and robustness.

## Ethical Statement

There are no ethical issues.

## Acknowledgments

This work was supported by the Natural Science Foundation of Anhui Province (No. 2308085QF228) and the National Natural Science Foundation of China (No. 62306004).

## References

[Han *et al.*, 2025] Qingqi Han, Liang Hu, and Wanfu Gao. Integrating label confidence-based feature selection for partial multi-label learning. *Pattern Recognit.*, 161:111281, 2025.

[Hang and Zhang, 2023] Jun Yi Hang and Min Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 1339–1351, 2023.

[Hao *et al.*, 2023] Pingting Hao, Liang Hu, and Wanfu Gao. Partial multi-label feature selection via subspace optimization. *Inf. Sci.*, 648:119556, 2023.

[Hashemi *et al.*, 2020] Amin Hashemi, Mohammad Bagher Dowlatshahi, and Hossein Nezamabadi-Pour. MFS-MCDM: Multi-label feature selection using multi-criteria decision making. *Knowl. Based Syst.*, 206:106365, 2020.

[Hou *et al.*, 2016] Peng Hou, Xin Geng, and Min Ling Zhang. Multi-label manifold learning. In *Proceedings of the 30th AAAI conference on artificial intelligence*, volume 30, pages 1680–1686, 2016.

[Hu *et al.*, 2020] Juncheng Hu, Yonghao Li, Wanfu Gao, and Ping Zhang. Robust multi-label feature selection with dual-graph regularization. *Knowl. Based Syst.*, 203:106126, 2020.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *proceeding of the 24th international joint conference on artificial intelligence*, pages 3569–3575, 2015.

[Komeili *et al.*, 2017] Majid Komeili, Wael Louis, Narges Armanfard, and Dimitrios Hatzinakos. Feature selection for nonstationary data: Application to human recognition using medical biometrics. *IEEE Trans. Cybern.*, 48(5):1446–1459, 2017.

[Lee and Seung, 1999] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Li *et al.*, 2023] Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection via robust flexible sparse regularization. *Pattern Recognit.*, 134:109074, 2023.

[Li *et al.*, 2024] Ximing Li, Yuanchao Dai, Bing Wang, Changchun Li, Renchu Guan, Fangming Gu, and Jihong Ouyang. WPML$^3$CP: Wasserstein Partial Multi-Label Learning with Dual Label Correlation Perspectives. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4479–4487, 2024.

[Li *et al.*, 2025a] Yonghao Li, Xiangkun Wang, Xin Yang, Wanfu Gao, Weiping Ding, and Tianrui Li. Fusion-enhanced multi-label feature selection with sparse supplementation. *Inf. Fusion*, 117:102813, 2025.

[Li *et al.*, 2025b] Zhuoming Li, Yuheng Jia, Mi Yu, and Zicong Miao. Calibrated disambiguation for partial multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18620–18628. AAAI Press, 2025.

[Liu *et al.*, 2025] Guanghui Liu, Qiaoyan Li, Xiaofei Yang, Zhiwei Xing, and Yingcang Ma. Partial multi-label feature selection based on label matrix decomposition. *Neural Comput. Appl.*, 37(6):4207–4227, 2025.

[Pan *et al.*, 2025] Hanlin Pan, Kunpeng Liu, and Wanfu Gao. Reconsidering Feature Structure Information and Latent Space Alignment in Partial Multi-label Feature Selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19786–19794, 2025.

[Pham *et al.*, 2022] Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Yuefeng Li, and Haoran Xie. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowl. Based Syst.*, 235:107662, 2022.

[Sun *et al.*, 2019] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5016–5023, 2019.

[Sun *et al.*, 2024] Zhenzhen Sun, Zexiang Chen, Jinghua Liu, Yewang Chen, and Yuanlong Yu. Partial multi-label feature selection via low-rank and sparse factorization with manifold learning. *Knowl. Based Syst.*, 296:111899, 2024.

[Wang *et al.*, 2019] Haobo Wang, Weiwei Liu, Yang Zhao, Chen Zhang, Tianlei Hu, and Gang Chen. Discriminative and Correlative Partial Multi-Label Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3691–3697, 2019.

[Wang *et al.*, 2022] Jing Wang, Peipei Li, and Kui Yu. Partial multi-label feature selection. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2022.

[Wang *et al.*, 2024] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. Limited-supervised multi-label learning with dependency noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15662–15670. AAAI Press, 2024.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4302–4309. AAAI Press, 2018.

[Xie and Huang, 2021] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3676–3687, 2021.

[Yilmaz *et al.*, 2021] Selim F Yilmaz, E Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[Yu *et al.*, 2022] Kui Yu, Yajing Yang, and Wei Ding. Causal feature selection with missing data. *ACM Trans. Knowl. Discov. Data*, 16:1–24, 2022.

[Zhang and Ma, 2022] Yao Zhang and Yingcang Ma. Non-negative multi-label feature selection with dynamic graph constraints. *Knowl. Based Syst.*, 238:107924, 2022.

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

[Zhang *et al.*, 2025] Yao Zhang, Jun Tang, Ziqiang Cao, and Han Chen. Sparse multi-label feature selection via pseudo-label learning and dynamic graph constraints. *Inf. Fusion*, 118:102975, 2025.