

Causality-Aware Feature Disentanglement Model for Shared and Class-Specific Information Learning

Dr. Haziq Abdullah

Department of Artificial Intelligence Technologies
Brunei Center for Digital Innovation Bandar Seri Begawan, Brunei

Dr. Nur Aisyah Rahman

School Laboratory of Machine Learning
SystemInstitute of Smart Computing Research Brunei

VOLUME 03 ISSUE 08 (2026)

Published Date: 12 August 2026 // Page no.: -09-15

ABSTRACT

Multi-label learning requires models to simultaneously capture information shared across labels and information that is uniquely associated with individual labels. Conventional feature-selection and representation-learning approaches can improve predictive performance, but they may inadequately distinguish genuine label-relevant information from redundant, correlated, or spurious associations. This paper proposes a Causality-Aware Feature Disentanglement Model (CAFDM) for learning shared and class-specific representations through causal feature analysis, correlation-aware disentanglement, and label-conditioned representation learning. The proposed framework separates an input representation into a common component that captures information useful across multiple labels and label-specific components that preserve discriminative information associated with individual labels. Causal relevance is incorporated to prioritize stable feature relationships rather than relying exclusively on statistical correlation. The framework is conceptually grounded in multi-label feature learning, causality-based feature selection, multidimensional correlation analysis, and representation separation. The proposed architecture further incorporates contrastive learning principles to increase the distinction between shared and label-specific representations while reducing redundancy. Analytical findings indicate that causal disentanglement can provide a more interpretable representation structure, improve resistance to spurious correlations, and support label-specific decision making without discarding globally useful information. The model is particularly applicable to multi-label classification, image annotation, text classification, recommendation, and biological function prediction. The study also identifies important limitations, including causal discovery uncertainty, computational complexity, dependence on label quality, and the absence of direct experimental validation. Overall, the framework provides a theoretically motivated direction for integrating causal reasoning with disentangled multi-label representation learning.

Keywords: Causal Representation Learning, Feature Disentanglement, Multi-Label Classification, Label-Specific Features, Shared Features, Causal Feature Selection, Contrastive Learning, Feature Learning.

1. INTRODUCTION

1.1 Background

Multi-label learning differs from conventional single-label classification because an instance may simultaneously belong to several semantic categories. Consequently, an effective model must identify both the information common to multiple labels and the information that distinguishes one label from another. The distinction is important because indiscriminately learning all features as a unified representation can cause label-specific information to become mixed with shared information. Existing research has demonstrated the importance of learning common and label-specific features for multi-label classification, particularly when label correlation contains useful structural information (Li et al., 2022; Zhang and Zhou, 2013).

The problem becomes more difficult when observed feature-label relationships contain redundant or non-causal associations. Correlation can be useful for identifying dependencies, but it does not necessarily establish whether a feature provides stable explanatory information for a label. Causality-based feature selection therefore offers an important theoretical foundation for identifying features whose relationships with target variables are less dependent on accidental associations (Yu et al., 2020). This distinction motivates the development of a representation-learning framework in which causal relevance guides the separation of shared and label-specific information.

Recent work also demonstrates that feature selection can benefit from multidimensional correlation and parallel learning mechanisms. Wang et al. examined weak-label-specific feature learning through multidimensional correlation, emphasizing the importance of capturing relationships between features and weak labels (Wang et al.,

2023). Similarly, Miu et al. investigated parallel dual-channel multi-label feature selection, suggesting that multiple information channels can be exploited to preserve complementary feature characteristics (Miu et al., 2023).

1.2 Problem Statement

A central limitation in multi-label representation learning is the difficulty of determining whether a feature should contribute to the common representation, a particular label-specific representation, or both. Purely correlation-driven methods can preserve redundant associations, while overly aggressive feature selection can remove information required for other labels. Therefore, the principal problem addressed in this paper is:

How can causal information be incorporated into feature disentanglement so that shared and label-specific representations are learned jointly while reducing redundancy and spurious associations?

The proposed Causality-Aware Feature Disentanglement Model addresses this problem by introducing a causal relevance mechanism before and during representation separation.

1.3 Research Relevance and Objectives

The relevance of the proposed framework arises from the intersection of three research directions: multi-label feature learning, causal feature selection, and disentangled representation learning. Multi-label learning research emphasizes label dependencies, whereas causal feature analysis emphasizes the stability and explanatory value of feature relationships. Integrating these perspectives can produce representations that are both predictive and structurally interpretable.

The main objective is to construct a unified architecture that learns a shared representation and multiple label-specific representations while controlling redundancy between them. A secondary objective is to establish a theoretical mechanism through which causal relevance influences feature allocation. The framework is intended to support applications where an instance may contain multiple simultaneous attributes, such as image annotation, textual categorization, recommendation, and biological function prediction.

1.4 Scope and Significance

The proposed model focuses on multi-label representation learning rather than causal discovery as an independent problem. It assumes that causal relevance can be estimated using an appropriate causal feature-selection mechanism and uses this information to constrain representation learning. The significance of the approach lies in its attempt to move beyond correlation-only feature selection toward representations in which common and label-specific information have explicit functional roles.

2. LITERATURE REVIEW

2.1 Multi-Label Feature Learning

Zhang and Zhou provided a broad examination of multi-label learning algorithms and established the importance of modeling the unique characteristics of multi-label prediction problems (Zhang and Zhou, 2013). Their work provides the general methodological foundation for treating multiple labels as structurally related rather than as independent classification targets.

Li et al. specifically investigated learning common and label-specific features for multi-label classification with correlation information (Li et al., 2022). Their work is particularly relevant to the present framework because it explicitly recognizes that multi-label data contain both shared and label-dependent information. However, correlation information alone may not distinguish stable relationships from spurious statistical dependencies. The proposed model therefore extends this conceptual distinction by introducing causal relevance as an additional criterion.

2.2 Correlation and Label-Specific Feature Selection

Miu et al. developed a parallel dual-channel strategy for multi-label feature selection, demonstrating the value of maintaining complementary feature-selection channels (Miu et al., 2023). This principle is compatible with the proposed shared-versus-specific architecture because the two representation pathways can preserve different types of information.

Wang et al. examined weak-label-specific feature learning based on multidimensional correlation (Wang et al., 2023). Their approach highlights the importance of analyzing relationships across multiple dimensions rather than treating each feature-label relationship independently. Nevertheless, correlation-based dependence remains different from causal dependence. The proposed model consequently uses correlation as a supporting signal while assigning greater structural importance to causally relevant information.

2.3 Causality-Based Feature Selection

Yu et al. reviewed causality-based feature-selection methods and evaluations, establishing the relevance of causal reasoning to feature selection (Yu et al., 2020). Causal feature selection is theoretically attractive because features associated with a target through more stable mechanisms may generalize better than features whose predictive value results primarily from contextual correlation.

The proposed framework does not assume that every causal relationship is automatically predictive. Instead, causal relevance is treated as a mechanism for prioritizing candidate information before the model learns shared and label-specific representations. This distinction is important because a causal feature can be weakly predictive in a particular dataset, while a correlated feature may temporarily provide strong predictive performance.

2.4 Representation Learning in Application Domains

Qian et al. investigated contrastive learning from label distributions in text classification, demonstrating the relevance of representation separation and discriminative learning to label-structured data (Qian et al., 2022).

Contrastive objectives are particularly suitable for the proposed model because they can encourage related representations to become closer while increasing separation between incompatible representation components.

Wei et al. addressed automatic image annotation using an improved nearest-neighbor strategy with semantic tag extension, illustrating the importance of semantic relationships in multi-label image annotation (Wei et al., 2021). Xia et al. developed a strategy for multi-class and multi-label protein-function annotation by integrating diverse deep-learning methods, demonstrating the practical breadth of multi-label prediction across biological applications (Xia et al., 2022).

Liu et al. investigated multi-community graph convolution networks with decision fusion for personalized recommendation (Liu et al., 2022). Although recommendation is not identical to multi-label classification, the study supports the broader principle that structurally different information communities can contribute complementary predictive signals.

2.5 Research Gap

The reviewed literature establishes several complementary foundations but leaves an identifiable integration gap. Existing studies separately emphasize common and label-specific feature learning, multidimensional correlation, causal feature selection, contrastive learning, graph-based representation, and application-specific multi-label prediction (Li et al., 2022; Miu et al., 2023; Yu et al., 2020; Qian et al., 2022). A unified framework that explicitly uses causal relevance to regulate the separation between shared and class-specific representations remains insufficiently developed within the supplied literature.

The proposed CAFDM addresses this gap by integrating causal relevance, representation disentanglement, label conditioning, and redundancy control within one conceptual architecture.

3. METHODOLOGY

3.1 Proposed Causality-Aware Feature Disentanglement Architecture

The proposed CAFDM consists of five functional stages: input representation, causal relevance estimation, shared-feature encoder, label-specific encoders, and multi-label prediction. For an input instance (x), an initial representation ($h=f(x)$) is generated. A causal relevance module then estimates the contribution of candidate features to each label.

Let (C) represent the common representation and (S_k) represent the label-specific representation for label (k). The overall representation can be formulated as:

$$h \rightarrow \{C, S_1, S_2, \dots, S_K\}$$

]

where (K) denotes the number of labels.

The common representation is designed to capture information that contributes to multiple labels, while (S_k) captures information particularly relevant to label (k). This separation directly operationalizes the common-versus-label-specific feature distinction emphasized by Li et al. (2022).

3.2 Causal Relevance Estimation

The causal module assigns a relevance score to feature-label relationships. Conceptually, for feature (x_j) and label (y_k), a causal relevance score can be expressed as:

$$R_{jk}=F(X_j,Y_k,\mathcal{G})$$

where (F) represents the selected causal assessment mechanism and ($\mathcal{G}$) represents the assumed or estimated dependency structure.

Features with strong causal relevance can receive larger weights during representation construction. Conversely, features whose predictive association is primarily redundant or unstable can be attenuated. This approach is motivated by the distinction between causal and correlation-based feature selection discussed by Yu et al. (2020).

The causal module should not be interpreted as proving causality automatically. Its purpose is to introduce causal evidence as a structured signal for feature prioritization. This limitation is important because incorrect causal assumptions can propagate into downstream representation learning.

3.3 Shared Representation Learning

The common encoder (E_c) transforms the input into:

$$C=E_c(h)$$

The objective of (C) is to retain information useful across several labels. For example, in image annotation, visual characteristics such as object shape or scene structure may contribute to several simultaneous tags. In biological function prediction, molecular characteristics may support multiple functional annotations (Xia et al., 2022).

The shared representation should therefore maximize multi-label utility while minimizing excessive dependence on any single label. This constraint prevents the common pathway from becoming merely a duplicate of one label-specific pathway.

3.4 Label-Specific Representation Learning

For each label (k), a separate encoder generates:

$$S_k = E_k(h, R_k)$$

where (R_k) is the causal relevance information associated with label (k) .

The purpose of (S_k) is to preserve discriminative information that may be weakly represented in the common space. This is particularly important when two labels share broad semantic characteristics but require different fine-grained evidence.

The dual representation structure is consistent with the conceptual distinction between common and label-specific features reported by Li et al. (2022), while the causal weighting mechanism provides an additional criterion for determining what information should receive label-specific emphasis.

3.5 Disentanglement and Redundancy Control

A central challenge is preventing (C) and (S_k) from encoding identical information. The model therefore incorporates a disentanglement objective:

$$\mathcal{L}_{\{dis\}} = \sum_k D(C, S_k)$$

where (D) measures undesirable dependence between the shared and label-specific representations.

A practical implementation may combine covariance reduction, orthogonality constraints, mutual-information minimization, or contrastive separation. Contrastive learning is relevant because it can explicitly structure representations according to positive and negative relationships (Qian et al., 2022).

The objective is not complete statistical independence. Excessive separation could remove useful information that naturally belongs to both common and specific representations. The model should instead seek functional disentanglement, where each pathway has a distinct predictive role.

3.6 Label Prediction

The final prediction for label (k) can be obtained from the combined representation:

$$z_k = G_k(C, S_k)$$

followed by a sigmoid probability:

$$\hat{y}_k = \sigma(z_k)$$

The classification loss can be represented by a multi-label binary cross-entropy objective:

$$\mathcal{L}_{\{cls\}} = -\sum_{k=1}^K [y_k \log(\hat{y}_k) + (1 - y_k) \log(1 - \hat{y}_k)]$$

The total training objective is conceptually defined as:

$$\mathcal{L} = \mathcal{L}_{\{cls\}} + \lambda_1 \mathcal{L}_{\{causal\}} + \lambda_2 \mathcal{L}_{\{dis\}} + \lambda_3 \mathcal{L}_{\{con\}}$$

where $(\mathcal{L}_{\{causal\}})$ controls causal relevance, $(\mathcal{L}_{\{dis\}})$ promotes representation disentanglement, and $(\mathcal{L}_{\{con\}})$ represents a contrastive or representation-structuring objective.

3.7 Integration with Label Correlations

Multi-label dependencies cannot be ignored simply because the framework separates label-specific representations. The common representation can capture shared dependencies, while label-specific encoders retain distinctions. This design avoids the two extreme cases of treating all labels independently or forcing all labels into a single representation.

The graph-based perspective explored by Liu et al. (2022) further suggests that structurally related information can be represented through interconnected communities. Accordingly, label relationships can optionally be modeled through a label graph, with the shared encoder representing global structure and specific encoders representing localized label information.

3.8 Training and Operational Workflow

Training begins with feature extraction and normalization, followed by causal relevance estimation. The model then simultaneously optimizes shared and label-specific encoders. During each iteration, the classification objective rewards predictive representations, while causal and disentanglement objectives regulate feature allocation.

A hypothetical text-classification example illustrates the mechanism. Suppose a document contains labels representing several related topics. A word or phrase associated broadly with the subject may enter the common representation,

whereas a specialized term that strongly distinguishes one topic may receive greater weight in its corresponding label-specific representation. If the specialized term is only spuriously associated with the label, causal relevance estimation can reduce its influence.

3.9 Theoretical Positioning

The theoretical contribution of CAFDM is the integration of three principles. First, multi-label learning requires simultaneous modeling of shared and label-specific information (Zhang and Zhou, 2013; Li et al., 2022). Second, causal feature selection can distinguish potentially stable relationships from simple statistical association (Yu et al., 2020). Third, contrastive and multi-channel learning can improve separation of complementary representations (Miu et al., 2023; Qian et al., 2022).

The resulting framework is therefore not simply a feature-selection algorithm. It is a representation architecture in which causal information regulates how feature information is distributed across shared and label-conditioned latent spaces.

4. RESULTS

The analytical findings of the proposed framework indicate that causal-aware disentanglement provides a coherent mechanism for resolving the tension between shared information preservation and label-specific discrimination. The first major finding is that the explicit separation of (C) and (S_k) creates a more interpretable representation structure than a single undifferentiated feature space. The common pathway can represent information relevant to several labels, whereas the specific pathways can preserve discriminative characteristics that would otherwise be diluted.

The second finding concerns redundancy. Correlation-based feature learning can identify useful relationships, but highly correlated features may repeatedly encode the same underlying signal. By introducing a disentanglement objective, CAFDM attempts to reduce unnecessary overlap between shared and label-specific representations. This extends the dual-channel perspective of Miu et al. (2023) by assigning explicit representational roles to the two information pathways.

Third, causal relevance introduces a stability-oriented criterion. Yu et al. (2020) demonstrate the methodological importance of causality-based feature selection. Within CAFDM, this principle becomes operational rather than remaining an independent preprocessing step. A feature with strong statistical association but weak causal support can be assigned lower representational importance, reducing the possibility that the model learns an unstable shortcut.

Fourth, the architecture is compatible with diverse multi-label applications. The semantic structure of automatic image annotation, the complex functional relationships in protein annotation, and label-dependent textual representations can all be expressed through shared and specific feature spaces (Wei et al., 2021; Xia et al., 2022;

Qian et al., 2022). This indicates that the framework is not inherently tied to a single data modality.

Finally, the conceptual architecture suggests a stronger interpretability pathway. Because predictions can be decomposed into contributions from common and label-specific representations, researchers can investigate whether a prediction depends primarily on globally shared information or on specialized evidence. Such decomposition can complement broader self-adaptive reasoning architectures that emphasize structured computational decision processes (Ramamurthy et al., 2026).

However, these findings are analytical rather than experimentally verified. No benchmark dataset, baseline implementation, ablation study, or numerical evaluation was supplied for this paper. Consequently, claims of improved accuracy, F1-score, robustness, or generalization should be treated as hypotheses requiring empirical validation rather than established experimental outcomes.

5. DISCUSSION

The proposed CAFDM framework has several theoretical implications. Most importantly, it reframes multi-label feature learning as a structured allocation problem rather than a simple feature-selection problem. Instead of asking which features are globally useful, the framework asks which information should be shared across labels and which information should remain label-specific. This distinction is consistent with the findings of Li et al. (2022) while extending the underlying logic through causal relevance.

The causal component provides another important theoretical contribution. Traditional correlation-based approaches may perform well when the training distribution and deployment distribution are similar, but their representations can become vulnerable when correlations change. Causal feature selection offers a conceptual mechanism for reducing dependence on unstable associations (Yu et al., 2020). Nevertheless, causal inference is itself sensitive to assumptions, observational limitations, and model misspecification. Therefore, causal awareness should not be equated with guaranteed causal validity.

The framework also reveals a fundamental trade-off between disentanglement and predictive completeness. If the model forces shared and label-specific spaces to become completely independent, it may remove information that is legitimately relevant to both. Conversely, insufficient separation causes the two pathways to collapse into redundant representations. The appropriate objective is therefore controlled disentanglement rather than absolute independence.

The proposed architecture also complements correlation-oriented methods. Multidimensional correlation remains valuable for identifying label relationships, particularly in weak-label environments (Wang et al., 2023). Similarly, contrastive learning can strengthen representation separation by constructing structured similarities and differences (Qian et al., 2022). Thus, causal information should be viewed as an additional structural constraint rather than a replacement for correlation and representation learning.

From a practical perspective, CAFDM could support

applications in image annotation, recommendation, text classification, and biological function prediction. Image annotation systems, for example, may benefit from separating scene-level characteristics from object-specific evidence, while biological annotation may distinguish molecular properties shared by several functions from characteristics associated with a particular function (Wei et al., 2021; Xia et al., 2022). Recommendation systems may similarly benefit from separating community-level preferences from user- or item-specific characteristics (Liu et al., 2022).

A further implication concerns interpretability. A conventional classifier may identify a label without indicating whether the decision arises from common or label-specific evidence. CAFDM provides an architectural basis for tracing predictions to these two information sources. Such decomposition may become increasingly useful in complex intelligent systems where transparent reasoning and adaptive representation are important (Ramamurthy et al., 2026).

Several limitations remain. First, causal relevance estimation can be computationally demanding, particularly for high-dimensional data and large label spaces. Second, inaccurate causal assumptions may incorrectly suppress useful features. Third, label noise can contaminate the label-specific representation and cause the causal module to learn unreliable relationships. Fourth, the framework introduces multiple loss components and hyperparameters, increasing optimization complexity. Finally, the present study does not provide empirical validation; therefore, the proposed advantages must be tested against established multi-label feature-selection and representation-learning baselines.

Future empirical research should evaluate the model using multiple modalities and conduct ablation experiments comparing correlation-only, causal-only, disentanglement-only, and integrated configurations. Performance should be assessed using predictive metrics together with representation redundancy, robustness under distribution changes, and interpretability measures.

6. CONCLUSION

This paper proposed a Causality-Aware Feature Disentanglement Model for jointly learning shared and class-specific information in multi-label environments. The framework integrates causal feature relevance, shared representation learning, label-specific encoders, disentanglement constraints, and contrastive representation principles into a unified architecture. Its central contribution is the explicit treatment of feature learning as a structured separation between globally useful information and label-dependent information.

The theoretical analysis indicates that causal awareness can complement correlation-based multi-label learning by providing a mechanism for reducing reliance on potentially unstable associations. At the same time, disentanglement can improve the structural interpretability of learned representations by assigning distinct roles to common and specific feature spaces. The

framework is applicable across textual, visual, recommendation, and biological multi-label scenarios.

The principal limitation is the absence of experimental validation. Accordingly, the proposed model should be considered a theoretically grounded research framework rather than an empirically established superior method. Future work should implement CAFDM on benchmark multi-label datasets, compare it with existing feature-selection and representation-learning approaches, and investigate causal robustness under distribution shifts. Such validation would determine whether the proposed combination of causal reasoning and feature disentanglement can consistently improve multi-label prediction, interpretability, and generalization.

REFERENCES

1. J. H. Li, P. P. Li, X. G. Hu, and K. Yu, "Learning common and label-specific features for multi-Label classification with correlation information," *Pattern Recogn.*, vol. 121, no. 8, pp. 108257, 2022.
2. S. H. Liu, B. Wang, B. Liu, and L. T. Yang, "Multi-community graph convolution networks with decision fusion for personalized recommendation," in *Pacific-Asia Conf. Knowl. Discov. Data Min.*, Chengdu, China, 2022, pp. 16–28.
3. J. L. Miu, Y. B. Wang, Y. S. Cheng, and F. Chen, "Parallel dual—channel multi-label feature selection," *Soft Comput.*, vol. 27, no. 11, pp. 7115–7130, 2023.
4. T. Qian, F. Li, M. S. Zhang, G. N. Jin, P. Fan and W. Dai, "Contrastive learning from label distribution: A case study on text classification," *Neurocomput.*, vol. 507, no. 7, pp. 208–220, 2022.
5. Y. B. Wang, W. X. Ge, Y. S. Cheng, and H. F. Wu, "Weak-label-specific features learning based on multidimensional correlation," *J. Nanjing Univ. (Natural Sci.)*, vol. 59, no. 4, pp. 690–704, 2023 (In Chinese).
6. W. Wei et al., "Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model," *Procedia Comput. Sci.*, vol. 183, no. 24, pp. 616–623, 2021.
7. W. Q. Xia et al., "PFmulDL: A novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods," *Comput. Biol. Med.*, vol. 145, pp. 105465, 2022.
8. K. Yu et al., "Causality-based feature selection: Methods and evaluations," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–36, 2020.
9. M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2013.
10. K. Ramamurthy, R. K. Konduru and N. Amanmadv, "EvoGraphCoder: An Evolutionary Graph-Reasoning Framework for Self-Adaptive Software Engineering,"

in IEEE Access, vol. 14, pp. 63063-63076, 2026,
doi: 10.1109/ACCESS.2026.3686019.

- 11.** Geo Philip, Paulson, Integrated Intelligent Building Energy Management: A Multi-LayerFramework for Renewable Energy, HVAC Optimization, and SmartElectrical Network Coordination. Available at SSRN: <https://ssrn.com/abstract=6993209> or <http://dx.doi.org/10.2139/ssrn.6993209>
- 12.** M. R. Marri, S. B. Kurada, S. Gupta, S. Dey, S. Kumar and R. Chauhan, "Graph-Based Deep Learning Model for Identifying Cyber Threats in Cloud Platforms," 2025 International Conference on Computational Intelligence, Security, and Artificial Intelligence (IntelliSecAI), Al-Khobar, Saudi Arabia, 2025, pp. 1-7, doi: 10.1109/IntelliSecAI66368.2025.11472831.