Towards Reliable Multi-Label Classification via Conditional Dependency Modeling

Arkapal Panda, Aditya Shankar Pal, and Utpal Garain

Indian Statistical Institute, Kolkata, West Bengal, 700108, INDIA

Abstract. Multi-label classification models are commonly trained by assuming conditional independence among labels, typically through the use of binary cross-entropy objectives. While this assumption simplifies optimization, it introduces a structural bias that ignores dependencies among labels and can adversely affect the calibration of predicted probabilities. In this work, we theoretically analyze the impact of this assumption and show that neglecting label dependencies leads to a miscalibration effect that is linked to the conditional dependency of the labels. Motivated by this observation, we propose *Pairwise Correlation Difference* (PCD), a novel auxiliary loss designed to incorporate label dependency information during training. PCD aligns the pairwise correlations of model logits with those of the ground-truth labels, thereby encouraging the network to capture conditional dependencies among labels. We combine PCD with the standard binary cross-entropy loss to form the *Correlated Multi-Label Loss* (CMLL), which serves as a objective for dependency-aware training. We provide theoretical justification for the functional form of the proposed loss and its connection to the dependency structure of the label distribution. Extensive experiments on three benchmark multi-label datasets demonstrate that CMLL consistently produces better-calibrated predictions while maintaining competitive classification accuracy compared to existing multi-label learning objectives. Additionally, we show that the proposed approach exhibits robustness to the choice of binning schemes used in calibration evaluation.

Keywords: Calibration · Conditional Dependency · Multi-Label Classification

1 Introduction

Deep Neural Networks (DNNs) including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [12] have demonstrated superior capacity in supervised learning tasks in computer vision. However, application of DNNs in safety-critical tasks requires trustworthy ML models which are not only accurate but also predicts accurate estimate of posterior probabilities i.e., the models should be well calibrated. For example, a perfectly calibrated weather prediction model would predict an 80% chance of rain on multiple days throughout the year, and it would actually rain on exactly 80% of those days. The importance

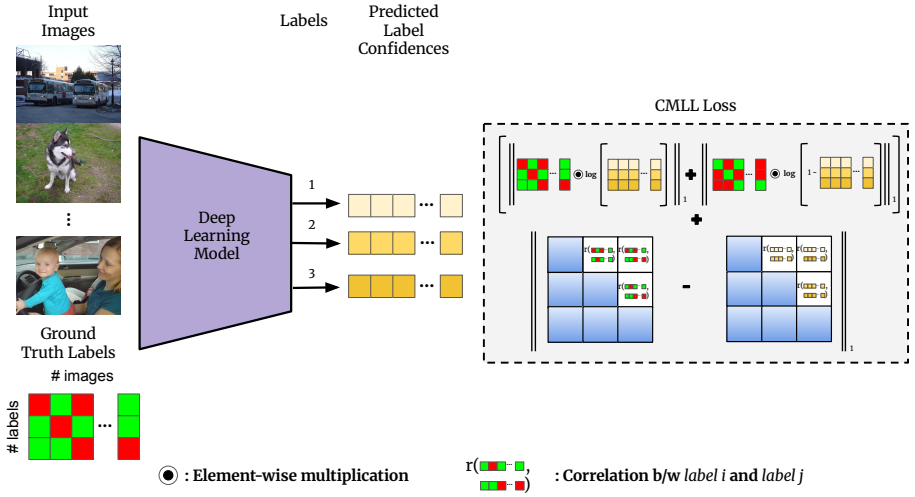

Fig. 1: Overview of the proposed CMLL loss.

of model calibration has been highlighted in several tasks. In disease diagnosis [16, 36], the predicted probabilities (confidence) of a model can be used to determine whether human intervention is necessary. But this can only be used safely if the DNN is well calibrated. Application of DNNs in cost-sensitive areas such as decision making in business [35, 45], fraud detection [2] etc, involves different costs for different misclassification errors. In these cases, building a model that predicts accurate class posterior probabilities cannot be compromised in favour of achieving Bayes optimality. Hence, calibration of DNNs has attracted substantial attention in several areas of computer vision and machine learning tasks [4, 14, 20, 29].

While various calibration methods have been proposed, most of them address binary or multi-class classification problems rather than MLC problems. However, in real world situations multiple labels are often associated with a single instance. For instance, in autonomous driving, a single image is often annotated with multiple scene instances [5] or in medical images like Chest X-Ray [3] where multiple disease condition might be associated with the same chest X-ray image. Thus, multi-label learning has garnered significant attention, with a focus on enhancing both accuracy and trustworthiness, particularly due to its applications in safety-critical domains [3, 48]. Some of the most popular approaches to multi-label learning include designing novel loss functions [19, 26, 49, 56] and leveraging auxiliary information [11, 21, 57]. For MLC tasks, DNNs can be trained using Class Probability Estimate (CPE) losses like BCE, SPA [8], ASY [49], and Focal Loss [37] to obtain calibrated predicted probability estimates but they operate under the strict assumption of conditional independence among labels. While some of these methods are proper scoring rules that enjoy Fisher Consistency

which guarantees perfectly calibrated models in the limit of infinite data, our experiments reveal that their performance is far from perfect and produces poorly calibrated results. We contend that this limitation arises because these losses do not adequately captures the dependency among the labels.

Multi-Label datasets can show complex dependency structure (For example, the model might be confident about y_A and y_B individually but obliviates the fact that they can never occur together.) among the labels which if not modeled properly may lead to problems in predicting accurate posterior probability estimates that is used in further downstream tasks like decision making and uncertainty quantification. While some literatures suggest that the advantages of exploiting label dependencies are contingent upon the chosen evaluation metric [10, 27], a growing body of subsequent research [7, 32, 33, 50] demonstrates that assuming label independence often leads to suboptimal predictive performance. There are a few methods [1, 7, 10, 47] that try to model the dependency among the labels during classification but the question of whether these or similar losses also encourage to produce calibrated probability estimates is still open. Besides, statistical properties of these methods are still not well understood. This motivates us to design a novel auxiliary loss function: Pairwise Correlation Difference (PCD) loss. The proposed loss function can be used along with BCE at the time of training to produce well calibrated probability estimates. PCD exploits the inherent conditional dependency structure during training so that the model parameters are aware of it. We provide a rigorous theoretical foundation for CMLL, showing that its design is not heuristic but grounded in learning theory. Our main contributions can be listed as follows:

- We show that the common assumption of conditional independence among labels in MLC tasks introduce a structural bias in posterior probability estimation, leading to poorly calibrated probability predictions.
- We theoretically analyze this phenomenon and establish that the magnitude of the induced bias is proportional to the pairwise covariance of the occurrence of labels.
- To address this issue we propose a trainable DNN calibration method by introducing a novel auxiliary loss function, PCD, to be used along with BCE at the time of training. Together we name the objective function as Correlated Multi-Label Loss (CMLL). CMLL considers the information of conditional dependency among the labels during training. An overview of our approach can be found in figure 1.
- Extensive experiments on three standard multi-label vision benchmarks demonstrate that CMLL consistently improves calibration while maintaining competitive classification performance, outperforming existing train-time calibration methods.

2 Related Works

2.1 Calibration of DNNs

Several studies in recent years have revealed that modern neural networks are poorly calibrated i.e. over confident [20,52]. It is important to note that calibration and classification accuracy are distinct [20]. Accuracy evaluates thresholded label decisions, whereas calibration evaluates the reliability of predicted probabilities. Various methods have been proposed for calibration including post hoc methods such as Temperature Scaling [20,54], Histogram Binning [58], Gaussian Process based methods [55], calibration via splines [22] Dirichlet Calibration [28]; train-time regularization [25,43,46,53]; Bayesian Methods [39,51]. Unfortunately, results on these methods are available only for binary and multi-class problems, not for multi-label classification problems. For calibration in MLC tasks Cheng et. al [8] and Ridnik et al. [49] has suggested Asymmetric Proper Loss but their method assumes label independence. Pal et al. [42] has suggested considering pairwise predicted probabilities in the training process but their loss function is based on heuristics. Peng et al. [44] has suggested semantic aware regularization to incorporate the information of dependency in sequential confidence calibration but their work is not designed to handle MLC tasks. Some prior work has attempted to incorporate label dependency information into multi-label learning. In particular, Chen et al. [6] leverage category-specific features to adapt label smoothing [38] for multi-label classification. However, the proposed DCLR framework in [6] relies on a two-stage learning procedure and the method is restricted to a limited class of architectures.

2.2 Multi-label classification

MLC is a special case of multidimensional function learning which has long been studied in machine learning literature [34,60]. Binary Relevance [59] is one of the most popular approaches to MLC tasks but it fails to capture the dependency among the labels. Classifier Chains [47] tires to model these dependency structures explicitly but computationally very demanding. Transforming MLC tasks into binary or multi-class classification problem such as Binary Relevance [18] or adaptation based classifiers such as AdaBoost.MH [25] are popular learning algorithms. However, these methods decompose the task by treating labels independently, thereby ignoring label dependencies. Methods like CCA [23], DCCA [1] claims to model the dependency among labels but statistical properties of these methods are still not widely understood. Moreover, all of their behaviour from a calibration perspective remains largely unexplored, particularly in deep learning settings.

3 Preliminaries

Notations: Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ be a multi-label dataset sampled from an underlying distribution P defined over $\mathcal{X} \times \mathcal{Y}$. An input instance is denoted by $\mathbf{x} \in \mathcal{X}$, and

its associated label vector is

$$\mathbf{y} = [y^{(1)}, \ldots, y^{(L)}]^\top \in \mathcal{Y},$$

where $\mathcal{Y} \subset \{-1, +1\}^L$ and L denotes the total number of labels. We use $\mathbf{X}$ and $\mathbf{Y}$ to denote the corresponding random vectors of inputs and labels, respectively.

For mini-batch training, let $\mathcal{X}_B$ and $\mathcal{Y}_B$ denote the B^{th} batch of inputs and labels. Each row of $\mathcal{Y}_B$ corresponds to the ground-truths for a particular label in batch B. $\mathcal{Y}_B^{(l)}$ is the l^{th} row of the matrix that denotes the collection of ground-truth values for the l^{th} label across the batch.

Let $\mathcal{H}$ denote the hypothesis class under consideration. The goal of multi-label classification (MLC) is to learn a classifier $\mathbf{h} \in \mathcal{H}$, $\mathbf{h} : \mathcal{X} \to \mathbb{R}^L$. In practice, we parameterize $\mathbf{h}$ using a deep neural network (DNN) with parameters θ, denoted by $\mathbf{h}_\theta = \left(h_\theta^{(1)}, \ldots, h_\theta^{(L)}\right) : \mathcal{X} \to \mathbb{R}^L$, where $h_\theta^{(l)}$ represents the output corresponding to the l^{th} label.

For a given batch B, we define $\mathbf{h}_\theta^B = \mathbf{h}_\theta(\mathcal{X}_B)$ as the matrix of model outputs over batch B. The l^{th} row of this matrix is denoted by $\mathbf{h}_\theta^{B,l}$. The predicted label vector for an input is written as $\hat{\mathbf{y}} = [\hat{y}^{(1)}, \ldots, \hat{y}^{(L)}]^\top$.

Given $\mathbf{x} \in \mathcal{X}$, the goal of calibrated multi-label learning is to predict the vector of posterior probability $\rho(\mathbf{x}) = [\rho^{(1)}(\mathbf{x}), \ldots, \rho^{(L)}(\mathbf{x})]^\top$ accurately

$$\rho^{(l)}(\mathbf{x}) = P(y^{(l)} = 1|\mathbf{x}) \; ; \; \forall l \in \{1, \ldots L\}$$

where $P(\cdot|\mathbf{x})$ is the true conditional distribution.

To estimate the posterior probability of a label, a multi-label DNN maps the input to an embedding $\mathbf{h}_\theta(\mathbf{x}) = [h_\theta^{(1)}(\mathbf{x}), \ldots, h_\theta^{(L)}(\mathbf{x})]^\top$. Then each $h_l(\mathbf{x})$ is mapped to a label probability estimate through sigmoid function. So,

$$\hat{\rho}^{(l)}(\mathbf{x}) = Q_\theta(y^{(l)} = 1|\mathbf{x}) = \sigma(h_l(\mathbf{x})) \; ; \; \forall l \in \{1, \ldots, L\}$$

where $Q_\theta(\cdot|\mathbf{x})$ is the conditional distribution induced by the DNN.

Definition 1 (Conditional Independence [10]). *A random vector of labels* $\mathbf{Y} = [Y^{(1)}, \ldots, Y^{(L)}]$ *will be conditionally independent given* $\mathbf{x} \in \mathcal{X}$ *if*

$$P(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \prod_{l=1}^{L} P(Y^{(l)}|\mathbf{X} = \mathbf{x}) \tag{1}$$

By dependency we will consider conditional dependency in our work. The random vector will be conditionally dependent if the equality in equation 1 does not hold.

Definition 2 (Perfect Calibration [8]). *We can say that a DNN for MLC tasks is perfectly calibrated if* $\hat{\rho}(\cdot) = \rho(\cdot)\ ; \forall \mathbf{x} \in \mathcal{X}$.

Definition 3 (Calibration Error). *The miss-calibration or Calibration Error [42, 54] of a multi-label DNN* $\mathbf{h}_\theta$ *can be written as*

$$CE^{p}_{\mathbf{h}_\theta} = \mathbb{E}\left[\left|\left|\mathrm{E}[\mathbf{y}|\boldsymbol{\sigma} \circ \mathbf{h}_\theta(\mathbf{x})] - \boldsymbol{\sigma} \circ \mathbf{h}_\theta(\mathbf{x})\right|\right|^{\mathrm{p}}_{\mathrm{p}}\right]^{\frac{1}{p}}$$

Estimating Calibration Error: The expectations in Definition 3 cannot be computed analytically since we only have access to a finite number of samples. Therefore, to estimate the calibration error in a non-parametric manner, we partition the interval $[0, 1]$ into M disjoint bins defined as $\mathcal{B}_m = \{(\frac{m-1}{M}, \frac{m}{M}]\ ; m \in \{1, \ldots, M\}\}$. Let $\mathcal{O}_m$ be the fraction of true positive instances within bin $\mathcal{B}_m$, and let $\mathcal{P}_m$ denote the average predicted confidence score of the samples assigned to bin $\mathcal{B}_m$. Using this binning scheme, we define two widely used calibration metrics: Maximum Calibration Error (MCE) [39] and Average Calibration Error (ACE) [40], as follows:

$$ACE = \frac{1}{M^+} \sum_{m=1}^{M} |\mathcal{O}_m - \mathcal{P}_m|$$

$$MCE = \max_{m \in \{1,\ldots,M\}} |\mathcal{O}_m - \mathcal{P}_m|$$

M^+ is the number of non-empty bins.

It can be noted that Definition 2 is stronger, as it aims to recover the full posterior distribution $P(\mathbf{Y} \mid \mathbf{X})$. In contrast, Definition 3 provides an operational notion of calibration error used for evaluation. Since posterior recovery implies minimization of calibration error, optimizing toward the stronger conditional-distribution objective is consistent with improving calibration. In the following sections, we move in this direction by incorporating information about the conditional dependency structure into the training process.

4 Structural Bias

Perfect calibration is achieved when the predicted conditional distribution coincides with the true conditional distribution. Proper Scoring Rules (PSR) [17] are known to be minimized by the true conditional distribution thus asymptotically recovers calibrated marginals under correct model specification [37]. However, standard PSRs applied in MLC tasks does not consider conditional dependence among labels.

Below we analyse Kullback–Leibler (KL) divergence between the true joint conditional distribution and the factorized distribution induced by independence assumption. We show that this factorization introduces a *structural bias* whenever labels are conditionally dependent. In our context, the term *structural* refers

to the functional form imposed on the joint conditional distribution. Under the conditional independence assumption, the joint distribution is constrained to factorize as

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_i P\Big(y^{(i)} \mid \mathbf{x}\Big)$$

which enforces a specific structural form of the distribution.

Let $D_{KL}(\cdot||\cdot)$ denote the Kullback-Leibler divergence. The expected KL divergence between P and Q_θ can be written as

$$R_{f_\theta} = E_{\mathbf{x}\sim\mathcal{D}}[D_{KL}(P(\mathbf{Y}|\mathbf{X})||Q_\theta(\mathbf{Y}|\mathbf{X}))] \tag{2}$$

The Kullback–Leibler (KL) divergence quantifies the discrepancy between two probability distributions. If the KL divergence between the true conditional distribution and the induced conditional distribution is zero, then the two distributions coincide almost everywhere. In this case, the induced conditional distribution exactly matches the true conditional distribution, which corresponds to perfectly calibrated probabilistic predictions. For notational simplicity we have written $\mathbf{Y}|\mathbf{X}$ instead of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$.

As the term $Q_\theta(\mathbf{y}|\mathbf{x})$ is the distribution induced by the DNN on the assumption of conditional independence, the term can be written as product of independent marginals i.e.

$$Q_\theta(\mathbf{y}|\mathbf{x}) = \prod_i Q_\theta(y^{(i)}|\mathbf{x}) \tag{3}$$

Lemma 1. *For multi-label classification with L number of labels the KL divergence between true conditional distribution and distribution induced by the DNN on the assumption of conditional independence can be written as*

$$\begin{aligned} &D_{KL}(P(\mathbf{Y}|\mathbf{X})||Q_\theta(\mathbf{Y}|\mathbf{X})) \\ &\quad = D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x})) + \sum_{i=1}^{L} D_{KL}(P(y^{(i)}|\mathbf{x})||Q_\theta(y^{(i)}|\mathbf{x})) \end{aligned} \tag{4}$$

Proof. We have

$$\begin{aligned} &D_{KL}(P(\mathbf{Y}|\mathbf{X})||Q_\theta(\mathbf{Y}|\mathbf{X})) \\ &\quad = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \log \frac{P(\mathbf{y}|\mathbf{x})}{\prod_i Q_\theta(y^{(i)}|\mathbf{x})} \end{aligned}$$

By adding and subtracting the *log* product of actual marginals $\prod_i P(y^{(i)}|\mathbf{x})$ we get

$$
\begin{aligned}
&D_{KL}(P(\mathbf{Y}|\mathbf{X})||Q_\theta(\mathbf{Y}|\mathbf{X})) \\
&= \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \log \frac{P(\mathbf{y}|\mathbf{x})}{\prod_i P(y^{(i)}|\mathbf{x})} + \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \log \frac{\prod_i P(y^{(i)}|\mathbf{x})}{\prod_i Q_\theta(y^{(i)}|\mathbf{x})} \\
&= D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x})) + \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \sum_{i=1}^{L} \log \frac{P(y^{(i)}|\mathbf{x})}{Q_\theta(y^{(i)}|\mathbf{x})} \\
&= D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x})) + \sum_{i=1}^{L} \sum_{y^{(i)}} \Big(\sum_{\mathbf{y}^{(-i)}} P(y^{(i)}, \mathbf{y}^{(-i)}|\mathbf{x}) \Big) \log \frac{P(y^{(i)}|\mathbf{x})}{Q_\theta(y^{(i)}|\mathbf{x})} \\
&= D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x})) + \sum_{i=1}^{L} D_{KL}(P(y^{(i)}|\mathbf{x})||Q_\theta(y^{(i)}|\mathbf{x}))
\end{aligned}
$$

The quantity $D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x}))$ characterizes the *structural bias* induced by the assumption of conditional independence. Specifically, it measures the divergence between the true conditional distribution and its product of marginals approximation obtained by treating labels as independent. If the labels are truly independent then $P(\mathbf{y}|\mathbf{x}) = \prod_i P(y^{(i)}|\mathbf{x})$ and $D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x}))$ becomes zero, introducing no structural bias.

The second term on the right-hand side of equation 4 corresponds to the marginal discrepancy, as it decomposes into a sum of Kullback–Leibler divergences between the true and predicted marginal distributions for each label. This term can be minimized by optimizing a strictly proper scoring rule (e.g., BCE loss) independently for each label, since strictly proper scoring rules are known to minimize the KL divergence between the predicted and true marginals [8,37].

5 Correlated Multi-Label Loss

From the previous section, we observed that the quantity

$$D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x}))$$

corresponds to the degree of conditional dependence among the labels. When labels are conditionally dependent, this term is strictly positive and quantifies the structural error incurred under an independence assumption. This implies that this term contains the information of conditional dependency. However, directly using the information revealed by this term in the objective function is not feasible since the true conditional distributions $P(\mathbf{y} \mid \mathbf{x})$ and $P(y^{(i)} \mid \mathbf{x})$ are unknown. Consequently, the functional form of the above KL divergence cannot be evaluated explicitly. To address this limitation, we seek a tractable approximation of $D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x}))$ and show that it is proportional to the

empirical covariance between pairs of labels. The following theorem formalizes this.

Theorem 1. *Let* $\mathbf{Y} = [Y^{(1)}, \ldots, Y^{(L)}]$ *represent the random vector of labels and the KL divergence admits a local quadratic expansion. Then it can be shown that*

$$D_{KL}(P(\mathbf{y}|\mathbf{x})||\prod_i P(y^{(i)}|\mathbf{x})) \propto \mathrm{Cov}(Y^{(m)}, Y^{(n)})$$
$$m \neq n\,;\, \forall m, n \in \{1, \ldots, L\}$$

Proof is available in the supplementary material.

As we can see from Theorem 1, if we can model the pairwise covariance then we will be able to capture the dependency structure of the labels. Motivated by this observation, and to exploit the dependency information captured by the first term on the right-hand side of Eq. 4, we introduce a novel tractable auxiliary surrogate objective, termed as *Pairwise Correlation Difference (PCD)* loss. For clarity and consistency with practical implementation, we define the PCD loss in batch form as follows.

$$\mathcal{L}_{PCD}(B) = \sum_{i=1}^{L} \sum_{j=i+1}^{L} \left| \tau(\phi \circ \mathbf{h}_\theta^{B,i}, \phi \circ \mathbf{h}_\theta^{B,j}) - \tau(\mathcal{Y}_B^{(i)}, \mathcal{Y}_B^{(j)}) \right| \tag{5}$$

τ is the Pearson's correlation and $\phi : \mathbb{R} \to [-1, 1]$ is an increasing function.

Intuitively, if a DNN perfectly captures the conditional dependence structure among labels, then the pairwise correlation between the predicted logits of two labels should align with the corresponding pairwise correlation of the ground-truth labels. Therefore, minimizing the discrepancy between these two correlations provides an auxiliary objective that encourages the model to learn conditional dependency during training which in turn helps to produce better calibrated results. We thus define the PCD loss as the difference between the pairwise correlations computed from model logits and those computed from ground-truth labels, thereby explicitly regularizing the model toward capturing label interactions.

We use PCD loss in conjunction with BCE loss and we name it as *Correlated Multi-Label Loss* (CMLL) which we use as the objective function. The BCE loss helps the model to learn the marginals and PCD loss helps to learn the conditional dependency during training. The loss can be written as

$$\begin{aligned}
\mathcal{L}_{CMLL}(B) &= \mathcal{L}_{BCE}(B) + \mathcal{L}_{PCD}(B) \\
&= \frac{1}{|B|} \sum_{n=1}^{|B|} \sum_{i=1}^{L} \left[y_n^{(i)} \log(\hat{\rho}_n^{(i)}) + (1 - y_n^{(i)})(\log(1 - \hat{\rho}_n^{(i)})) \right] \\
&\quad + \lambda \cdot \left[\sum_{i=1}^{L} \sum_{j=i+1}^{L} \left| \tau(\phi \circ \mathbf{h}_\theta^{B,i}, \phi \circ \mathbf{h}_\theta^{B,j}) - \tau(\mathcal{Y}_B^{(i)}, \mathcal{Y}_B^{(j)}) \right| \right]
\end{aligned} \tag{6}$$

where λ is a hyper-parameter controlling the relative weight of auxiliary PCD loss.

We set $\lambda = 1$ in our experiments. This choice preserves the original scaling between the primary loss and the PCD term. Also, this selection is theoretically motivated by equation 4 in Lemma 1, where both quantities appear in their original proportional form. Setting $\lambda = 1$ therefore maintains consistency with the theoretical results and reflects the balance suggested by the theoretical results.

6 Experiments and Results

6.1 Models and Datasets

To evaluate the benefits of the proposed loss function, we conducted experiments on three publicly available datasets – PASCAL VOC, MSCOCO, and WIDER-A to test the effectiveness of our models. A brief summary of the datasets is outlined below:

- PASCAL VOC 2012 [13]: The PASCAL VOC 2012 dataset extends the original PASCAL VOC challenge and comprises 5717 training images and 5823 test images annotated across 20 object categories.
- MS-COCO [31]: The Microsoft Common Objects in Context (MS-COCO) dataset is a large-scale benchmark in computer vision, containing 82081 training images and 40137 test images. Unlike traditional single-label datasets, MS COCO reflects real-world scenarios by including images with complex scenes and multiple objects from 80 object categories.
- WIDER-A [30]: The WIDER Attribute (WIDER-A) dataset is a large-scale benchmark designed for human attribute recognition. It consists of images collected from diverse scenarios with significant variations in pose, appearance, illumination, and occlusion. The dataset contains 28345 train images and 29179 test images. Each person instance in the dataset is annotated with 14 attributes, covering aspects such as clothing style, accessories, and physical characteristics.

The proposed loss function is tested on both CNN-based models (ResNet-50 [24]) and transformer-based models (ViT-B/32 [12]). For the multi-label classification task, the standard softmax prediction head of the models is replaced with a sigmoid activation function to allow probability estimation for each class.

6.2 Baselines

We use BCE, Focal Loss (FL) [37], ASY [49] and three recently proposed losses TWL [26], LDACE-CCL [42], SPA [8] to compare our result with. Among all these losses BCE and SPA are proper losses. The parameters are selected as proposed in their respective original works.

6.3 Evaluation Metrics

To measure the accuracy in classification tasks we follow [15] and [49] and use Hamming Loss (HL) and mean Average Precision (mAP). By following Cheng et al. [8], to measure calibration error of DNNs trained with aforementioned losses, we employ two metrics based on reliability diagram [9, 41] – Average Calibration Error (ACE) [40] and Maximum Calibration Error (MCE) [39]. We have taken the number of bins $M = 10$ to produce our results.

6.4 Implementation Details

Both the ResNet-50 and the ViT-B/32 models, together with the loss functions, are implemented in Pytorch and are trained using a single NVIDIA A6000 GPU. The ResNet-50 model is optimized using Adam optimizer with a learning rate of $1e-4$ and weight decay of $1e-5$. The ViT B/32 model is trained using AdamW optimizer, employing the same learning rate of $1e-4$ and a weight decay of $1e-2$. The input image resolution for the models is set to 224×224. All the models are trained for 100 epochs with a batch size of 32. For evaluation, we select the model corresponding to the best validation loss. For all the experiments, the value of λ in CMLL loss (Equation 6) is set to 1. The source code, including data preprocessing scripts, model definitions, and training codes are given in `https://github.com/theimageprocessingguy/CMLL`.

	Model	Metric	Method						
			BCE	**TWL**	**FL**	**ASY**	**SPA**	**LDACE - CCL**	**CMLL (Ours)**
Classif.	**RN50**	H. Loss ↓	<u>0.0663</u>	0.1142	0.0687	0.1239	0.0845	0.0711	**0.0219**
		mAP ↑	0.9025	0.8105	**0.9273**	0.8853	0.9152	0.9045	<u>0.9251</u>
	ViT	H. Loss ↓	0.1063	0.1890	0.1038	0.1520	0.1248	**0.0745**	<u>0.0761</u>
		mAP ↑	0.8972	0.8918	0.9240	0.8983	0.9158	<u>0.9274</u>	**0.9368**
Calib.	**RN50**	ACE ↓	0.1533	0.3363	<u>0.1403</u>	0.2595	0.2807	0.2514	**0.1265**
		MCE ↓	0.3215	0.4703	<u>0.2825</u>	0.3804	0.4287	0.4002	**0.2189**
	ViT	ACE ↓	<u>0.0824</u>	0.2284	0.1197	0.2917	0.3182	0.3992	**0.0247**
		MCE ↓	0.2346	0.3563	0.2773	0.3411	0.4867	<u>0.2132</u>	**0.0761**

Table 1: The performance results corresponding to the PASCAL VOC 2012 dataset.

6.5 Results and Discussion

Table 1, 2, 3 summarizes our results. We can see that our proposed loss CMLL performs better in all three cases in terms of producing calibrated probabilities.

	Model	Metric	Method						
			BCE	TWL	FL	ASY	SPA	LDACE - CCL	CMLL (Ours)
Classif.	RN50	H. Loss ↓	0.0324	0.0479	**0.0214**	0.0271	0.0264	0.0284	0.0241
		mAP ↑	0.9385	0.9184	0.9005	0.9406	0.9461	0.9666	**0.9698**
	ViT	H. Loss ↓	0.0324	0.1156	**0.0295**	0.0590	0.1159	0.0306	0.0310
		mAP ↑	0.9207	0.9253	0.9147	0.9417	0.9245	0.9531	**0.9775**
Calib.	RN50	ACE ↓	0.0654	0.2057	0.1254	0.2137	0.1277	0.1385	**0.0030**
		MCE ↓	0.1766	0.4170	0.2366	0.3125	0.1907	0.2347	**0.0052**
	ViT	ACE ↓	0.0829	0.2573	0.1143	0.2137	0.1584	0.1582	**0.0020**
		MCE ↓	0.2426	0.4371	0.2607	0.3925	0.2935	0.2809	**0.0038**

Table 2: The performance results corresponding to the MS-COCO dataset.

	Model	Metric	Method						
			BCE	TWL	FL	ASY	SPA	LDACE - CCL	CMLL (Ours)
Classif.	RN50	H. Loss ↓	0.1979	0.1753	0.1912	0.2248	0.2129	0.2044	**0.1741**
		mAP ↑	0.8198	0.6334	0.8087	0.8195	0.8049	0.7839	**0.8248**
	ViT	H. Loss ↓	**0.1891**	0.3312	0.2067	0.1976	0.1910	0.2026	0.1936
		mAP ↑	0.8180	0.7546	0.7932	0.7935	0.7921	0.7919	**0.8226**
Calib.	RN50	ACE ↓	0.1352	0.2194	0.1235	0.2729	0.2538	0.2480	**0.1232**
		MCE ↓	0.3209	0.3995	0.2492	0.3255	0.3215	0.3773	**0.2256**
	ViT	ACE ↓	0.0795	0.2787	0.2124	0.2301	0.2214	0.1864	**0.0145**
		MCE ↓	0.2783	0.3768	0.2824	0.2885	0.3215	0.3051	**0.0233**

Table 3: The performance results corresponding to the WIDER-A dataset.

The results of ACE and MCE endorses that. BCE, ASY and SPA performs worse than CMLL because none of them cannot capture the dependency between the labels. This result supports our argument that incorporating label interdependencies is essential for achieving better calibration. FL, TWL and LDACE-CCL both performs far worse in terms of ACE and MCE, owing to the fact that none of them are strictly proper. This observation supports our initial claim that explicitly modeling conditional dependencies during training is crucial for designing loss functions that improve calibration. In terms of HL and mAP, CMLL performs well in most of the cases. This implies that CMLL does not affect the accuracy to produce well calibrated posterior probabilities.

In the tables 1, 2, 3, the best performances are highlighted in bold and the second best performance is highlighted with underlines. Across all three datasets and both architectures, CMLL consistently outperforms other loss functions in

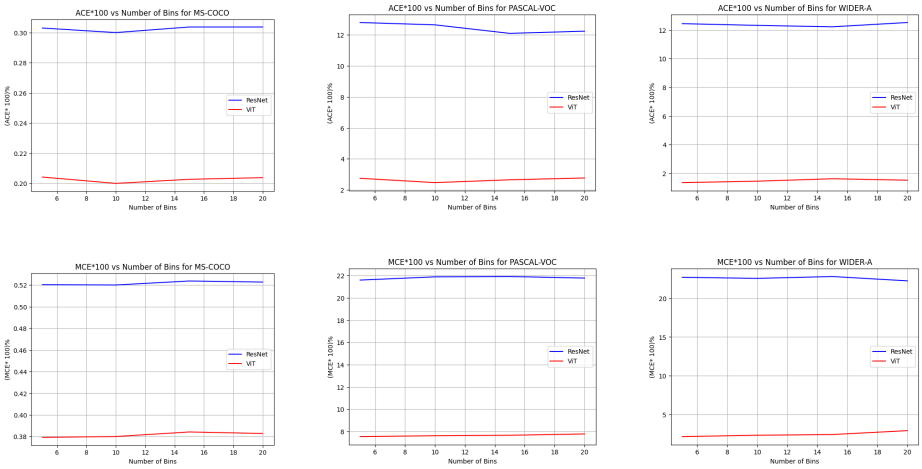

Fig. 2: The first row shows how the Average Calibration Error (ACE) varies with the number of bins, while the second row shows the variation of the Maximum Calibration Error (MCE) with respect to the number of bins.

terms of calibration while maintaining competitive accuracy. For the PASCAL-VOC 2012 dataset, training ViT-B/32 with CMLL yields the lowest ACE and MCE values of 0.0247 and 0.0761, respectively, representing a significant improvement. The HL and mAP also achieves mostly better results. On the MS-COCO dataset, both Resnet-50 and ViT-B/32 show substantial gains in calibration when trained with CMLL, without compromising accuracy. Similarly, on WIDER-A, ViT trained with CMLL demonstrates remarkable improvements in both calibration and accuracy.

In Tables 1, 2, and 3, we report calibration results using a fixed number of bins $M = 10$. Although the quantitative results are presented for this specific choice, we further demonstrate that the calibration metrics ACE and MCE are robust with respect to the number of bins. As shown in Figure 2, the values of both metrics remain relatively stable across different bin settings, thereby confirming that our findings are not sensitive to the choice of number of bins.

7 Conclusion

In this work, we theoretically show that assuming conditional independence among labels during training introduces a structural bias in multi-label classification models, which in turn negatively affects their calibration. To address this issue, we propose *Pairwise Correlation Difference* (PCD), a novel auxiliary loss designed to incorporate information about conditional label dependencies during training. We combine PCD with the standard binary cross-entropy (BCE) loss to form the *Correlated Multi-Label Loss* (CMLL), which enables the model

to learn dependency-aware representations while optimizing the primary classification objective.

We provide theoretical justification for the functional form of both PCD and CMLL, showing how the proposed loss captures second-order dependency information among labels. Extensive experiments on three multi-label benchmark datasets demonstrate that CMLL consistently produces better-calibrated predictions compared to standard training objectives. Furthermore, we show that the proposed calibration approach is robust to the choice of binning schemes used for evaluation. Importantly, CMLL achieves competitive classification performance and performs on par with state-of-the-art multi-label loss functions in terms of accuracy.

In future work, we plan to extend this framework to study the behaviour of CMLL under covariate shift and out-of-distribution settings, as well as derive theoretical generalization bounds for the proposed loss function.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International conference on machine learning. pp. 1247–1255. PMLR (2013)
2. Bahnsen, A.C., Stojanovic, A., Aouada, D., Ottersten, B.: Cost sensitive credit card fraud detection using bayes minimum risk. In: 2013 12th international conference on machine learning and applications. vol. 1, pp. 333–338. IEEE (2013)
3. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9**(1), 6381 (2019)
4. Bouniot, Q., Mozharovskyi, P., d'Alché Buc, F.: Tailoring mixup to data for calibration. arXiv preprint arXiv:2311.01434 (2023)
5. Chen, L., Zhan, W., Tian, W., He, Y., Zou, Q.: Deep integration: A multi-label architecture for road scene recognition. IEEE Transactions on Image Processing **28**(10), 4883–4898 (2019)
6. Chen, T., Wang, W., Pu, T., Qin, J., Yang, Z., Liu, J., Lin, L.: Dynamic correlation learning and regularization for multi-label confidence calibration. IEEE Transactions on Image Processing (2024)
7. Chen, Y.N., Lin, H.T.: Feature-aware label space dimension reduction for multi-label classification. Advances in neural information processing systems **25** (2012)
8. Cheng, J., Vasconcelos, N.: Towards calibrated multi-label deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27589–27599 (2024)
9. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. Journal of the Royal Statistical Society: Series D (The Statistician) **32**(1-2), 12–22 (1983)
10. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: On label dependence and loss minimization in multi-label classification. Machine Learning **88**(1), 5–45 (2012)
11. Ding, Z., Wang, A., Chen, H., Zhang, Q., Liu, P., Bao, Y., Yan, W., Han, J.: Exploring structured semantic prior for multi label recognition with incomplete labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3398–3407 (2023)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (2012)
14. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
15. Gao, W., Zhou, Z.H.: On the consistency of multi-label learning. In: Proceedings of the 24th annual conference on learning theory. pp. 341–358. JMLR Workshop and Conference Proceedings (2011)
16. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)

17. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association **102**(477), 359–378 (2007)
18. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 22–30. Springer (2004)
19. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894 (2013)
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
21. Guo, Z., Dong, B., Ji, Z., Bai, J., Guo, Y., Zuo, W.: Texts as images in prompt tuning for multi-label image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2808–2817 (2023)
22. Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., Hartley, R.: Calibration of neural networks using splines. arXiv preprint arXiv:2006.12800 (2020)
23. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16**(12), 2639–2664 (2004)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hebbalaguppe, R., Prakash, J., Madan, N., Arora, C.: A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16081–16090 (2022)
26. Kobayashi, T.: Two-way multi-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7476–7485 (2023)
27. Koyejo, O.O., Natarajan, N., Ravikumar, P.K., Dhillon, I.S.: Consistent multilabel classification. Advances in Neural Information Processing Systems **28** (2015)
28. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems **32** (2019)
29. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
30. Li, Y., Huang, C., Loy, C.C., Tang, X.: Human attribute recognition by deep hierarchical contexts. In: European conference on computer vision. pp. 684–700. Springer (2016)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
32. Liu, W., Tsang, I.: Large margin metric learning for multi-label prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
33. Liu, W., Tsang, I.: On the optimality of classifier chain for multi-label classification. Advances in Neural Information Processing Systems **28** (2015)
34. Liu, W., Wang, H., Shen, X., Tsang, I.W.: The emerging trends of multi-label learning. IEEE transactions on pattern analysis and machine intelligence **44**(11), 7955–7974 (2021)

35. Manzoor, A., Qureshi, M.A., Kidney, E., Longo, L.: A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE access **12**, 70434–70463 (2024)
36. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020)
37. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in neural information processing systems **33**, 15288–15299 (2020)
38. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems **32** (2019)
39. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)
40. Neumann, L., Zisserman, A., Vedaldi, A.: Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection (2018)
41. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632 (2005)
42. Pal, A.S., Panda, A., Garain, U.: Label dependency aware loss for reliable multi-label medical image classification. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
43. Park, H., Noh, J., Oh, Y., Baek, D., Ham, B.: Acls: Adaptive and conditional label smoothing for network calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3936–3945 (2023)
44. Peng, Z., Luo, Y., Chen, T., Xu, K., Huang, S.: Perception and semantic aware regularization for sequential confidence calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10658–10668 (2023)
45. Petrides, G., Moldovan, D., Coenen, L., Guns, T., Verbeke, W.: Cost-sensitive learning for profit-driven credit scoring. Journal of the Operational Research Society **73**(2), 338–350 (2022)
46. Popordanoska, T., Sayer, R., Blaschko, M.: A consistent and differentiable lp canonical calibration error estimator. Advances in Neural Information Processing Systems **35**, 7933–7946 (2022)
47. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333–359 (2011)
48. Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R.: Every annotation counts: Multi-label deep supervision for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9532–9542 (2021)
49. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 82–91 (2021)
50. Shen, X., Liu, W., Tsang, I.W., Sun, Q.S., Ong, Y.S.: Multilabel prediction via cross-view search. IEEE transactions on neural networks and learning systems **29**(9), 4324–4338 (2017)
51. Subedar, M., Krishnan, R., Meyer, P.L., Tickoo, O., Huang, J.: Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6301–6310 (2019)

52. Tao, L., Zhu, Y., Guo, H., Dong, M., Xu, C.: A benchmark study on calibration. arXiv preprint arXiv:2308.11838 (2023)
53. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in neural information processing systems **32** (2019)
54. Tomani, C., Cremers, D., Buettner, F.: Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In: European conference on computer vision. pp. 555–569. Springer (2022)
55. Wenger, J., Kjellström, H., Triebel, R.: Non-parametric calibration for classification. In: International Conference on Artificial Intelligence and Statistics. pp. 178–190. PMLR (2020)
56. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European conference on computer vision. pp. 162–178. Springer (2020)
57. Yazici, V.O., Gonzalez-Garcia, A., Ramisa, A., Twardowski, B., Weijer, J.v.d.: Orderless recurrent models for multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13440–13449 (2020)
58. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Icml. vol. 1 (2001)
59. Zhang, M.L., Li, Y.K., Liu, X.Y., Geng, X.: Binary relevance for multi-label learning: an overview. Frontiers of Computer Science **12**(2), 191–202 (2018)
60. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering **26**(8), 1819–1837 (2013)