Multi-Label Bayesian Active Learning with Inter-Label Relationships

Yuanyuan Qi[1] **Jueqing Lu**[1] **Xiaohao Yang**[1] **Joanne Enticott**[2] **Lan Du** [*1]

[1]Faculty of Information Technology, Monash University, Melbourne, Australia
[2]Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia

Abstract

The primary challenge of multi-label active learning, differing it from multi-class active learning, lies in assessing the informativeness of an indefinite number of labels while also accounting for the inherited label correlation. Existing studies either require substantial computational resources to leverage correlations or fail to fully explore label dependencies. Additionally, real-world scenarios often require addressing intrinsic biases stemming from imbalanced data distributions. In this paper, we propose a new multi-label active learning strategy to address both challenges. Our method incorporates progressively updated positive and negative correlation matrices to capture co-occurrence and disjoint relationships within the label space of annotated samples, enabling a holistic assessment of uncertainty rather than treating labels as isolated elements. Furthermore, alongside diversity, our model employs ensemble pseudo labeling and beta scoring rules to address data imbalances. Extensive experiments on four realistic datasets demonstrate that our strategy consistently achieves more reliable and superior performance, compared to several established methods.

1 INTRODUCTION

In recent years, extensive machine learning models and algorithms have been developed to deal with the exponential growth of real-world data. However, the significant mismatch between the rapid increase in data and the slow pace of manual data annotation underscores the imperative of active learning (AL) [Liu et al., 2021, Xie et al., 2022]. Multi-label active learning (MLAL), which considers the co-occurrence of labels and is more aligned with real-world applications, has been explored in different domains, including text classification [Han et al., 2024, Kang et al., 2020], medical imaging [Huang et al., 2024, Simao et al., 2023], remote sensing [Möllenbrok et al., 2023, Möllenbrok and Demir, 2023], and so on. The multi-label task, due to the complexity of label-wise correlation and imbalanced data distribution, remains a vital task to comprehensively examine the data, thus necessitating the development of effective query strategies [Siméoni et al., 2020, Dor et al., 2020].

To deal with the multi-label issue in active learning, earlier approaches usually transform it into multiple binary classification tasks, known as binary relevance (BR) which sums the informativeness evaluated for each individual label to obtain the final acquisition score [Zheng et al., 2021, Wang et al., 2022]. However, these approaches overlook the potential correlation of labels, such as their co-occurrence, which should be factored into the overall information assessment [Zhang et al., 2021, Min et al., 2022]. Consequently, the information inherent in the label correlation of the queried samples may not be fully explored.

Some recent works have employed co-occurrence and label correlation matrices to model these inherent label relationships [Su et al., 2023]. However, while the positive correlations, indicating strong co-occurrence between labels, have been included, few studies have explored negative correlations where labels are mutually exclusive and do not appear together. Moreover, asymmetric label-wise correlations, where one label frequently appears with another label without a reciprocal relationship, remains under-explored. This also includes the hierarchical structure of the label set, where node labels inherently belong to and serve as subsets of their corresponding root labels. The selection of overlapping labels, due to the hierarchical nature, affects the diversity of the strategy, consequently, its overall outcome [Nakano et al., 2020]. Furthermore, due to the high imbalance ratios in real-world datasets, addressing data imbalance to maintain consistent performance across different datasets highlights the critical importance of MLAL tasks [Chen et al., 2022, Arens et al., 2024].

[*]Corresponding author

Considering label co-occurrence and data imbalance, we propose a new MLAL framework, named multi-label **Co**-**R**relation-**A**ware active learning with **B**eta scoring rules (CRAB) [1] in this paper. By incorporating the Beta scoring rules to deal with data imbalance and the expected loss reduction framework to select the most informative data instance, we introduce dynamic positive and negative correlation matrices to handle the distinct and asymmetric label correlation within a Bayesian framework. This approach demonstrates robust and outstanding performance on four benchmark datasets for multi-label active learning.

2 RELATED WORK

Active learning involves selecting the most informative data from the unlabeled pool for annotation, thereby reducing the required training data while maintaining comparable performance. Two mainstream AL query strategies are uncertainty-based and diversity-based approaches [Ren et al., 2021]; the former concentrates on informative measurement at the sample-level [Abdar et al., 2021], while the latter emphasizes data distribution [Kim et al., 2023]. To quantify the sample uncertainty, methods such as proper scoring rules [Tan et al., 2024], Dirichlet distribution [Hemmer et al., 2022], and Gaussian Process [Shi et al., 2021] can be used to estimate the sample informativeness. By aligning the prior and posterior distributions with model output observations, these models can effectively capture the uncertainty.

However, focusing exclusively on uncertainty can introduce bias in sampling (i.e., selecting near-identical instances, thus wasting the annotation budget), which may lead to sub-optimal performance [Prabhu et al., 2019]. Incorporating diversity into the sampling process offers an alternative approach to enhancing generalization [Buchert et al., 2023]. Huang and Zhou [2013] utilized label cardinality inconsistency to exploit uncertainty and integrated it with the diversity-based sampling. PLVI-CE leverages average posterior probability discrepancy to measure data diversity and prediction inconsistency to assess uncertainty, thus enhancing model generalization with limited annotated instances [Gu et al., 2023]. Recently, Tan et al. [2024] proposed BESRA which uses the strictly proper scoring rules. Its acquisition function combines beta-scoring rules and k-means clustering to enhance diversity, while the Beta scoring rules also address data imbalance common in multi-label datasets. Inspired by BESRA, our framework further takes into account label correlation in the acquisition function.

Research in active learning has gradually paid attention to the label correlation in MLAL. In light of the specific characteristics of graph data, DAMAL incorporates class-label interactions using a graph-based ranking approach, where edge weights are defined as the cosine similarity between latent features, thus quantifying the graph's informativeness in relation to label correlation [Mahapatra et al., 2024, Lu et al., 2020]. To address the uncertainty in feature correlations within standard data, Shi et al. [2021] integrated a Gaussian process with a Bernoulli Mixture model to model correlation through the covariance matrix. Correlation matrix-based weighted uncertainty, typically derived through co-occurrence or label similarity analysis, is commonly used to query the most informative label pairs by capturing the inter-label influence during label selection [Gong and Zhai, 2021, Su et al., 2023]. Han et al. [2024] propose a two-stage sample acquisition strategy, called ALMuLa-mix, utilizing inconsistency to capture label correlations with novel features as the first stage and employing the class frequency at the second stage to ensure inter-class diversity.

Although an increasing number of studies recognize the importance of correlation during data acquisition, existing approaches are often resource-intensive, requiring additional training for interrelation modeling, or struggle to maintain performance under data imbalance. Our approach effectively samples the representative data in a correlation-aware manner while maintaining consistent performance, even with highly imbalanced datasets.

3 CORRELATION-AWARE MULTI-LABEL ACTIVE LEARNING

Without loss of generality, suppose $L = \{X, Y\}, U = \{X\}$ represent the initial collection of training set and unlabeled data samples, where $|U| \gg |L|$; and $y_i \in \{-1, +1\}^k$ represents the label of the i_{th} example in the k label space. K denotes the total number of labels in the space. Firstly, our model generates the two-dimensional correlation matrix, including both positive correlation and negative correlation, based on the iteratively updated labeled dataset L. Then, given a model parameterized by $\theta \in \Theta$, the probability of label y of a data instance x is $P(y|\theta, x)$. We are able to derive the pseudo label as y^* based on $\int_\theta P(y|\theta, x)P(\theta)d\theta$, where the integration can be approximated by Monte Carlo via ensemble. Considering the model's learning capability of different categories of data, our model refines the sampling pool into a more preventative subset based on the pseudo labels. And considering the influence of label correlation in quantifying the informativeness of sample, we propose a variation of the beta scoring rule used in Tan et al. [2024]. Its key idea is introduced in Section 3.4. Finally, the clustering approach assures the diversity in sampling. Fig. 1 illustrates the overall flowchart of our proposed framework.

3.1 PRELIMINARIES

To address the multi-label active learning problem, most studies decompose it into multiple binary classification

[1] Our code is publicly available at `https://github.com/qijindou/CRAB`.

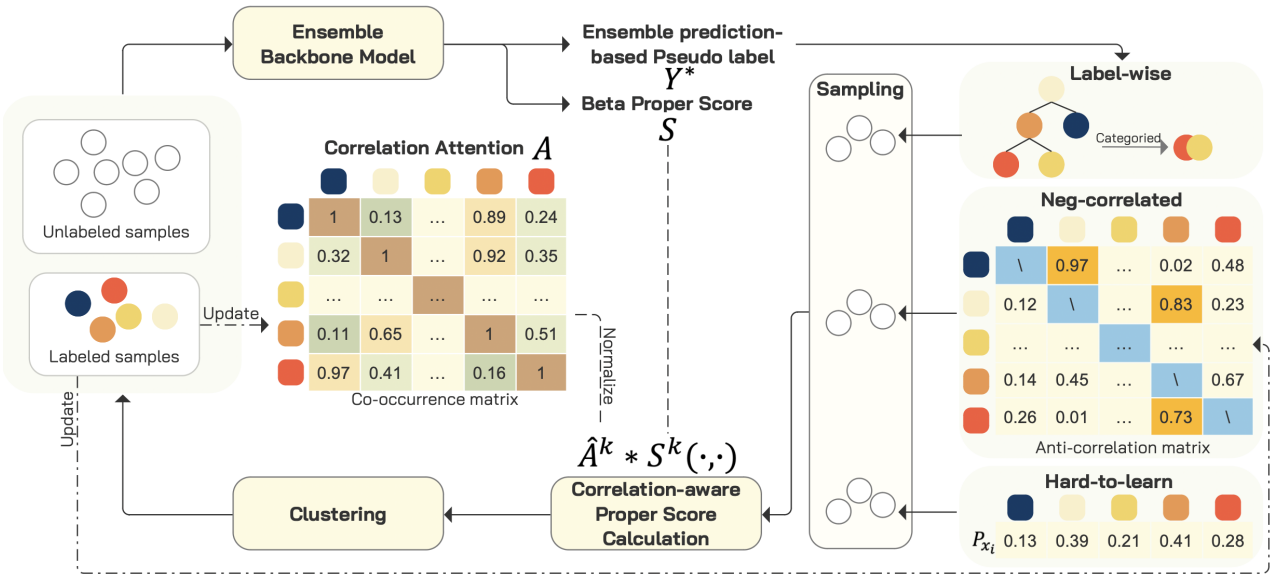


Figure 1: Overview of the CRAB framework. It is trained using an ensemble method that generates pseudo labels based on predictions and calculates the beta proper score. Firstly, positive and negative correlation matrices are updated with newly sampled instances. Then, utilizing pseudo labels, our model samples data from three categories: label-wise, negatively correlated, and hard-to-learn samples. Last, our model calculates correlation-aware proper score and subsequently clusters and labels the selected data based on this score.

tasks, aggregating individual label scores instead of assessing the entire instance holistically. This approach is formulated in Eq. (1), where S_{BR} represents a scoring function that measures the informativeness (e.g., uncertainty score) of individual samples, and S^k_{BR} denotes the score with respect to each label.

$$S_{BR}(p, y) = \sum_{k=1}^{K} S^k_{BR}(p, y^k) = \sum_{k=1}^{K} \mathcal{L}(y^k|p) \quad (1)$$

Monte Carlo estimation offers a probabilistic framework for computing acquisition scores by incorporating randomness to account for variability. Monte Carlo-based error reduction estimation [Roy and McCallum, 2001] utilizes the Monte Carlo approach to approximate the expected reduction in error resulting from the labeling of a given sample. However, while Monte Carlo-based methods estimate expected improvement in model performance, they do not assess the quality of probabilistic outputs. Rather than estimating error, proper scoring rules provide a summary measure of predictive probability, computing the positive-oriented rewards (i.e., utilities) that a classifier seeks to maximize [Gneiting and Raftery, 2007]. Eq. (2) and (3) show the core concept of expected increase in score when querying.

$$Q(L) = \mathbb{E}_{P_{(x)}} \mathbb{E}_{P(\theta|L)} \big[\mathbb{E}_{P(y|\theta,x)} [S(P(\cdot|x,\theta), y) - S(P(\cdot|x), y)]\big] \quad (2)$$

$$\begin{aligned} \Delta Q(x|L) &= Q_L - \mathbb{E}_{P_{(y|L,x)}}[Q_{L+\{x,y\}}] \\ &= \mathbb{E}_{P_{(y|L,x)}} \big[\mathbb{E}_{P(x')P(y'|L,(x,y),x')} \\ &\quad [S(P(\cdot|L,(x,y),x'),y') - S(P(\cdot|L,x'),y')]\big] \end{aligned} \quad (3)$$

S denotes the scoring function that evaluates predictive probability distribution on an event, i.e., predicting y', given x'. $Q(L)$ represents the mean proper scoring rule of the predictive probabilities obtained using Bayesian estimation based on the current labeled dataset L. The term $\Delta Q(x|L)$ denotes the increment in the score resulting from acquiring the label of a sample x, drawn from the unlabeled data pool U. And x is the point to be acquired, x' denotes the selected unlabeled anchor point for assessment. The value of $\Delta Q(x|L)$ is then used to select sample leading to a large increment in the score or reward. Since the label of unlabeled points x and x' is unknown, we derive $P(y'|L,(x,y),x')$ by calculating the posterior distribution of the ensemble models using Eq. (4).

$$P(y'|L,(x,y),x') = \sum_{\theta \in \Theta^E} P(y'|\theta,x') P(\theta|L,(x,y)) \quad (4)$$

$$P(\theta|L,(x,y)) \approx \frac{P(\theta|L)P(y|\theta,x)}{\sum_{\theta \in \Theta^E} P(\theta|L)P(y|\theta,x)} \quad (5)$$

Beta family [Buja et al., 2005], which generalizes the logarithmic score and the brier score, or the other desired cost-weighted scoring rule, is able to address the issue of imbalanced label distribution in multi-label learning. The equation below illustrates the proper scoring rules $\mathcal{L}$ of a predictive

distribution p given the expected value y^k, where y^k represents the label for class k. $\mathcal{L}(1|p)$ and $\mathcal{L}(0|p)$ represent the partial losses when p is been classified as 1 and 0, respectively.

$$S^k_{BR}(p, y^k) = y^k\mathcal{L}(1|p) + (1 - y^k)\mathcal{L}(0|p) \quad (6)$$

$$\mathcal{L}(1|p) = \mathcal{L}_1(1-p) = \int_p^1 p^{\alpha-1}(1-p)^\beta dp \quad (7)$$

$$\mathcal{L}(0|p) = \mathcal{L}_0(p) = \int_0^p p^\alpha(1-p)^{\beta-1} dp \quad (8)$$

By leveraging $I_x(\alpha, \beta)$, the Incomplete Beta Function, the closed form of the Beta distribution is obtained for $\alpha, \beta > 0$. When $\alpha = \beta = 0$, the scoring becomes log-loss, and when $\alpha = \beta = 1$, the scoring rule will transform to squared error losses. By adjusting the value of α and β, our model can effectively handle scenarios with diverse data distributions. In our research, we employ the greedy search result of BESRA as the parameter for scoring, where the $\alpha = 0.1, \beta = 3$ [Tan et al., 2024].

3.2 CORRELATION MATRIX CONSTRUCTION

In multi-label scenario, a single sample often has more than one label, and label with relatively small semantic distances frequently appear simultaneously. To leverage inherent relationships within the label space to assist in acquisition process, our framework maintains two dynamic matrices: a co-occurrence matrix representing positive correlations, and a anti-correlation matrix representing negative correlations between labels. Both matrices are updated after each acquisition iteration with newly annotated instances. By discovering the pattern of the occurrence between labels, we aim to quantify the informativeness considering influence of correlation, and maintain the diversity while take into account the imbalanced data distribution.

Positive correlation matrix: The positive correlation matrix A is constructed based on the label-wise dependence. A is a $K \times K$ two dimensional matrix, where each element $A(m, n)$ quantifies the dependency of the presence of label m on label n. This dependency is formally computed using Eq. (9). Specifically, $\sum_{i=1}^{L} N(y_i^m = +1, y_i^n = +1)$ refers to the count of labeled instances in which both labels m and n appear simultaneously. $P(y^m|y^n)$ gives the likelihood of label m occurring when label n is present. When $m = n$, $A(m, n)$ equals 1. The value of $A(m, n)$ reflects the probability of one label's existence conditioned on another.

$$\begin{aligned} A(m,n) &= P(y^m|y^n), \quad \text{where } m \neq n \\ &= \frac{\sum_{i=1}^{L} N(y_i^m = +1, y_i^n = +1)}{\sum_{i=1}^{L} N(y_i^n = +1)} \end{aligned} \quad (9)$$

The positive correlation matrix is constructed to characterize the pattern of the co-occurrence between labels and capture the asymmetric correlation between labels, including hierarchical relationships.

Negative correlation matrix: Despite the positive correlations, negative correlations between labels have rarely been addressed in previous research. However, in real-world scenarios, negative correlations are instrumental in enabling the model differentiate between mutually exclusive classes and contribute to more accurate decisions by clarifying the model's decision boundaries [Yang et al., 2024].

To effectively model these negative correlations, we construct an updated anti-correlation matrix, $NegA$. Maintaining the same format as the positive correlation matrix A, $NegA$ is a $K \times K$ two dimensional matrix, where each element $Neg(m, n)$ quantifies the confidence in the absence of label m given the presence of label n, as defined in Eq. (10). Specifically, $\sum_{i=1}^{L} N(y_i^m = -1, y_i^n = +1)$ represents the count of instances where labels m and n do not co-occur. This asymmetry allows the matrix to capture the nuanced conditional negative relationships, and provide a new perspective towards the label dependencies.

$$\begin{aligned} NegA(m,n) &= P(\overline{y^m}|y^n), \quad \text{where } m \neq n \\ &= \frac{\sum_{i=1}^{L} N(y_i^m = -1, y_i^n = +1)}{\sum_{i=1}^{L} N(y_i^n = +1)} \end{aligned} \quad (10)$$

3.3 CORRELATION-BASED SAMPLING

Refining the unlabeled pool to ensure that selected instances concentrate on specific representative criteria is a common strategy in active learning [Kang et al., 2020]. However, current research predominantly based on informativeness analysis, neglecting the critical role of data correlation in MLAL. To address this limitation and provide more representative and evenly distribution samples for continuous process, our model refines the unlabeled pool from three perspectives based on the correlation properties to generate a subset to be used in acquisition. And the pseudo label y^*, obtained by averaging the prediction result of ensemble models through Eq. (11), is used for the following correlation-based sampling.

$$y^* = \mathbb{I}[P(y \mid x, L) > 0.5] \quad (11)$$

$$\begin{aligned} P(y \mid x, L) &= \int_\theta P(y|x,\theta)p(\theta \mid L) \\ &\approx \sum_{e=1}^{E} P(y|x,\theta_e)/E \end{aligned} \quad (12)$$

3.3.1 Label-wise sampling

In multi-label scenarios, labels often exhibit asymmetric correlations, where the present of one label, m, is highly correlated with another label, n, but not vice versa. Hierarchical structures within labels are a common example of

this kind of relationship. To illustrate the impact on performance, we can consider hierarchical data: when asymmetric correlations exist, the selection of root-node labels often inevitably overlaps with that of corresponding leaf-node labels, while rarely occurring independently, which reduces the representativeness of the selected root labels [Nakano et al., 2020].

To address this, our strategy introduces a new mechanism for the label-wise selection. If one label n is highly dependent on the presence of another m, while the reverse is not necessarily true—otherwise, they would effectively be considered the same label in most cases—and their correlation exceeds a predefined threshold, σ, set as the standard deviation of a two-tailed normal distribution, we classify the label pair as asymmetrically correlated. To improve label-wise sampling, we then refine the label space of instances that contain both labels m and n by removing label n, ensuring that sampling prioritizes the most independent label, m. The model then performs evenly sampling based on the refined pseudo labels, ensuring a more balanced sampling pool.

3.3.2 Negative-correlated label sampling

In multi-label learning, it is essential for the model to respect the exclusivity of certain labels to ensure accurate predictions [Huang et al., 2017]. When mutually exclusive labels, such as those that should not logically co-occur, are predicted together, it is often an indication of model bias or misguided learning [Huang and Kang, 2021, Perales-González et al., 2020]. This misalignment can reduce model's effectiveness. As the second subset for concentrated sampling, we select instances with negatively correlated pseudo labels that are not expected to co-occur in the label space.

To formalize this, we consider a pair of labels as mutually exclusive when the negative correlation coefficient $NegA(m, n)$ exceeds a predefined threshold, set as 2σ, where σ represents the standard deviation. Based on predictions with pseudo labels, our model selects samples with the predicted labels that are unlikely to co-occur, according to the negative correlation matrix, as those samples are at high risk of incorrect predictions. These selected samples are then added to the refined subset of the unlabeled pool, ensuring that the model better accounts for negative correlations.

3.3.3 Hard-to-learn label sampling

The third set, which our model uses to further expand the subset, consists of hard-to-learn samples. These samples are typically characterized by low confidence and low variability, indicating instances where the model has difficulty making accurate predictions. Such samples often contain ambiguous or noisy features or lie near decision boundaries [Chang et al., 2017, Yang and Xu, 2020]. In this study, we define samples without any predicted pseudo labels as hard-to-learn samples. Specifically, if the pseudo labels obtained through Eq. (11) for all classes of a given instance falls below the threshold, 0.5, the classifier cannot make any prediction, thus the sample is classified as hard to learn. To improve performance on these challenging instances while maintaining diversity in the sampling process, our model dynamically adjusts the sample size using a polynomial decay function, enabling more focused learning on difficult cases over time.

Algorithm 1 CRAB Update Strategy for MLAL

Input: Labeled pool: L; Unlabeled pool: U; Model: $\Theta^E = \theta_1, ..., \theta_E$; Query size: N; Per-label query size: N; Hard to learn query size: Z.
Output: Updated labeled and unlabeled pool.

for $\theta_e \in \Theta^E$ **do**
  Get the prediction $P(y|\theta_e, x)$ and corresponding Beta score S_{BR}
end for
Update correlation matrix A and $NegA$ with L via Eq. (9) and Eq. (10)
Get the pseudo label Y^* via Eq. (11)
Select and add N label-wise samples per label to U^*
Select and add N negative-correlated samples with $NegA$ to U^*
Select and add Z hard samples to U^*
for $u \in U^*$ **do**
  Get the attention beta score, S_{AB}, via Eq. (13). Update to $Q(x|L, x')$ with Eq. (3)
end for
L^+=k-Means Centers($Q(x|L, x')$, N)
$L \leftarrow L + L^+$; $U \leftarrow U - L^+$
return L, U

3.4 CORRELATION-AWARE QUERYING

With the correlation-based sampling strategy described in section 3.3, our model obtains a refined subset of the original unlabeled data pool. Then, our model calculates correlation-aware beta scores for these samples in the selected subset, and use those scores to cluster the samples for acquisition. This score computation method is inspired by the attention mechanism introduced in transformer models [Vaswani et al., 2017], which is defined as follows.

$$S_{AB}(f_L(x), y) = \sum_{m=1}^{K} \hat{A}(m, :) S^m_{BR}(f_L(x), y^m) \quad (13)$$

$$\begin{aligned} \hat{A}(m, n) &= \text{norm}(A(m, n)), \quad \text{where } m \neq n \\ &= \frac{A(m, n)}{\gamma \cdot \max(A(:, n))} \end{aligned} \quad (14)$$

Using Eq. (13), we score each prediction, where S_{AB} incorporates the influence of other labels' scores through the attention coefficient $\hat{A}$ as the final score, accounting the correlation. Additionally, we introduce γ, a normalization parameter set to 2, to prevent over-estimating correlated uncertainty while preserving the original significance of each label's score. This approach allows our model consider the impact of neighboring labels on informativeness, and the refined unlabeled pool enhances computational efficiency and deepens the analysis of label correlations. Algorithm 1 details the procedure for one iteration of our framework.

4 EXPERIMENTS

We collected four benchmark multi-label text datasets to analyze the performance and robustness of our framework [Kementchedjhieva and Chalkidis, 2023]. Those datasets include: RCV1 [Lewis et al., 2004], UKLEX [Chalkidis and Søgaard, 2022], EURLEX [Chalkidis et al., 2021], and MIMIC3 [Johnson et al., 2016]. RCV1 is a news articles dataset from Reuters; UKLEX is a collection of legal documents sourced from various categories within UK law; EURLEX is a set of descriptors from European legal information thesaurus extracted from the European Union's legal database; and MIMIC3 is a set of de-identified health records for medical diagnosis. Following the method by Charte et al. [2015], we used mean imbalance ratio (MeanIR) to create synthetic datasets with varying imbalance ratios based on the modified RCV1 dataset, reduce the label size of RCV1 to ten by selecting the most frequently occurring labels, enabling an evaluation of the model's performance across different degrees of imbalance. Table 1 and Table 2 offer a detailed summary of these four datasets. We also introduced a new metric, termed **CorrAvg**, defined as $\sum_{m=1}^{K}\sum_{n=1}^{K} A(m,n)/(K \times K), m \neq n$, to quantify the degree of inter-correlation within label set.

Dataset	#Document Train/Test	#Vocab./ #Label	#MeanIR Train/Test	#CorrAvg Train/Test
RCV1	24,891/6223	104,619/102	402/197	0.137/0.137
UKLEX	20,000/8500	63,157/18	7/6	0.026/0.024
EURLEX	55,000/5000	160,211/21	16/15	0.131/0.147
MIMIC	29,999/10000	137,678/19	127/101	0.321/0.320

Table 1: Benchmark datasets with corresponding imbalance level and correlation level statistics.

Dataset	#Document Train/Test	#Vocab./ #Label	#MeanIR Train/Test	#CorrAvg Train/Test
RCV1-T10-5	1,200/600	25,254/10	5/10	0.133/0.138
RCV1-T10-10	1,200/600	25,289/10	10/10	0.135/0.138
RCV1-T10-20	1,200/600	24,170/10	20/10	0.137/0.138
RCV1-T10-50	1,200/600	25,280/10	50/10	0.142/0.138

Table 2: Synthetic datasets with corresponding imbalance level and correlation level statistics.

4.1 IMPLEMETATION

We used Neural-Classifier [Liu et al., 2019], implemented in Pytorch [Paszke et al., 2019], as the code base. In our study, we exployed three mainstream models, TextCNN [Zhang and Wallace, 2017], TextRNN [Liu et al., 2016], and DistilBERT [Sanh, 2019], as the backbone classifiers. To enhance efficiency and performance, we applied the cold start strategy [Zhu et al., 2019] with random initialization at the beginning of each active learning iteration, a method known for its applicability to real-world scenarios [Frankle and Carbin, 2018]. All experiments were conducted on a single RTX3090 GPU. Following the setting of Tan et al. [2024], the maximum sequence length for the text data was set to 256, with each training iteration consisting of 80 epochs. The initial training set and validation set sizes are set to 100 and 1000, respectively, and are sampled from the training set. We implemented an early stopping criterion with the patience of 30 epochs to prevent the model from falling into local optima or overfitting [Du et al., 2019, Ying, 2019]. AdamW was used as the optimizer [Loshchilov and Hutter, 2019], with the learning rate tailored for each model: 5e-2 for TextCNN and TextRNN, and 5e-5 for DistilBERT. The hard-to-learn query size was set to 300 for benchmark datasets and 200 for synthetic datasets, while the per-label query size was set to 50 for RCV1 and 100 for other datasets, due to differences in label space size.

4.2 BASELINES

To conduct a comparative performance analysis, we adopted five state-of-the-art MLAL methods as baselines, including random sampling. Each baseline uses the same query parameters and backbone classifier to maintain consistency across experiments. Specifically, **MMC** [Yang et al., 2009] applies maximal confidence to selecting data that induces the largest reduction in expected model loss. **AUDI** [Huang and Zhou, 2013] explores uncertainty and diversity in both instance and label spaces through label ranking and threshold learning. **ADAPTIVE**[Li and Guo, 2013] integrates max-margin prediction uncertainty with label cardinality inconsistency to assess the unified informativeness of multi-label instances. **BESRA** [Tan et al., 2024] utilizes the beta scoring rules within an expected loss reduction framework to evaluate informativeness and employs vector representations to maintain diversity. **CMAL** [Yu et al., 2020] leverages global label correlation matrix and label space sparsity with uncertainty to query the most informative example-label pairs.

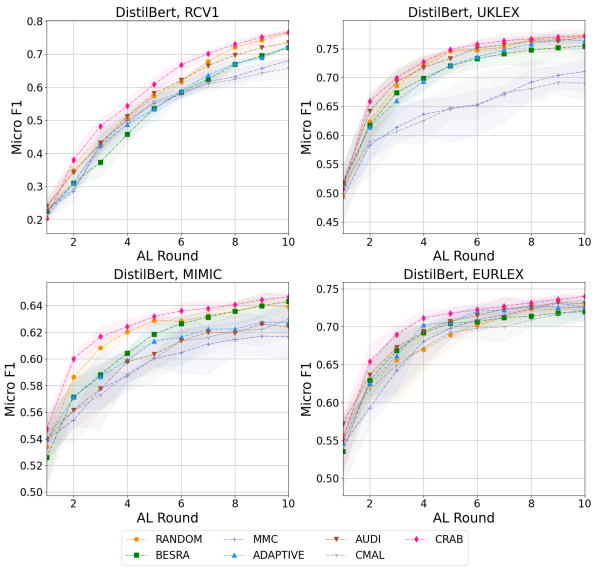


Figure 2: Averaged micro-F1 score on DistilBERT, averaged results with 5 random seeds.

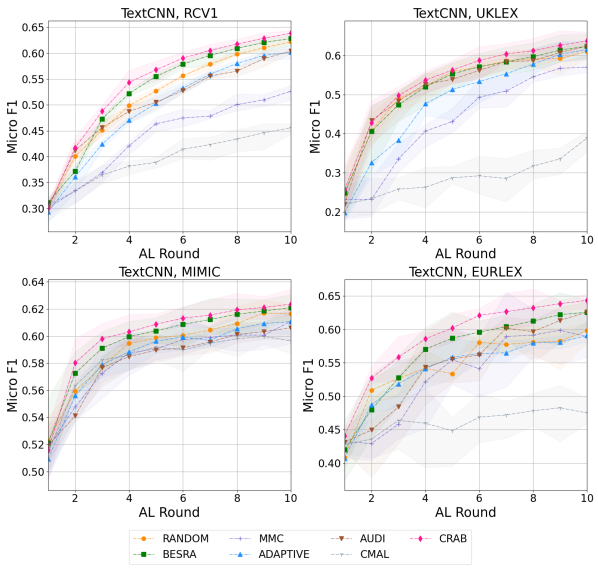


Figure 3: Averaged micro-F1 score on TextCNN, averaged results with 5 random seeds.

4.3 RESULTS

Figures 2–5 present the quantitative performance of our proposed framework. Figure 2 compares the performance of CRAB and baseline methods across four datasets using DistilBERT. Figure 3 and Figure 4 present supplementary performance comparisons based on TextCNN and TextRNN. Following Tan et al. [2024], we obtained the predictive distribution by training five ensemble models independently, each initialized with the same parameters for every AL iteration. The micro-F1 results show that our proposed model, CRAB, has consistently outperforms other AL methods across different text domains and network structures. Additionally, CRAB demonstrates robust performance on datasets with varying degrees of correlation, particularly compared with BESRA, suggesting that our strategy effectively models correlation during data selection.

Among the baseline methods, BESRA achieves strong results and shows relatively robustness on different datasets. However, its performance on the highly correlated MIMIC dataset is less stable, likely due to the absence of correlation consideration. AUDI, which incorporates both uncertainty and diversity at both data and instance level, presents notable performance on three of the datasets, RCV1, UKLEX, and EURLEX, but struggles on MIMIC. Adaptive and MMC yield similar results over four datasets, as both utilize the max-margin as the selection criterion. Although CMAL considers global label correlation, it only performs optimally on the highly correlated MIMIC dataset and does not maintain stable performance on all datasets. To examine robustness of the model on imbalanced datasets, we conducted comparative experiments on synthetic datasets, with results shown in Figure 5. CRAB maintains superior performance across synthetic datasets with different MeanIR values, demonstrating its capability to handle imbalanced datasets.

To further investigate the effectiveness of our model, Figures 6 and 7 present a qualitative analysis of CRAB. Figure 6 shows the MeanIR of the selected samples across AL iterations. Since MeanIR indicates imbalance, with lower values reflecting a more even data distribution, we observe that CRAB demonstrates a more balanced sample selection compared to other baseline models, which underscores its capacity to address data imbalance effectively. Figure 7 illustrates the trend of two categories of data within the unlabeled data pool: hard-to-learn data and negatively correlated data. Unlike random selection, CRAB strategically selects data that enhances model learning, thereby reducing misclassification of negatively correlated data. Additionally, CRAB improves performance on hard-to-learn data, helping the model becomes more robust and accurate. This targeted data selection contributes to a more balanced and adaptive learning process, ultimately leading to improved generaliztion across diverse data types.

4.4 PARAMETER SENSITIVITY ANALYSIS

To validate the effectiveness and generalizability of CRAB, we conduct two experiments to analyze its performance under different parameter settings. Figure 8a demonstrates the performance with varying sizes of hard-to-learn samples for refined unlabeled pool selection. With an acquisition size of 100 per iteration, the model achieves the optimal performance. However, including any hard-to-learn samples in the sampling pool leads to a performance decline. If the sample size is set too large, such as 200, the model initially shows

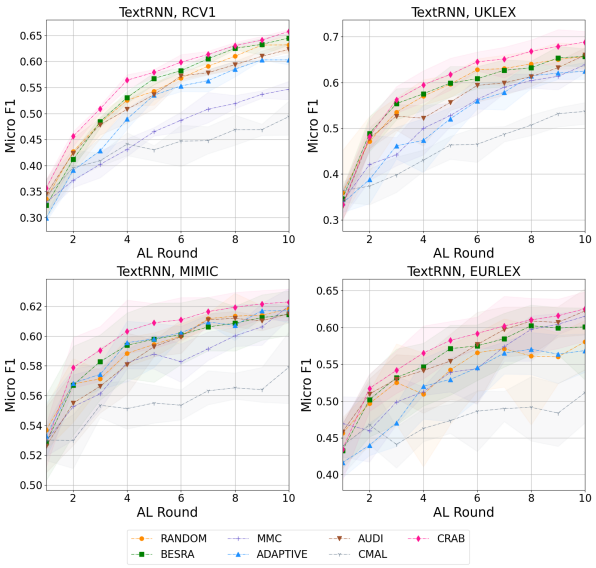


Figure 4: Averaged micro-F1 score on TextRNN, averaged results with 5 random seeds.

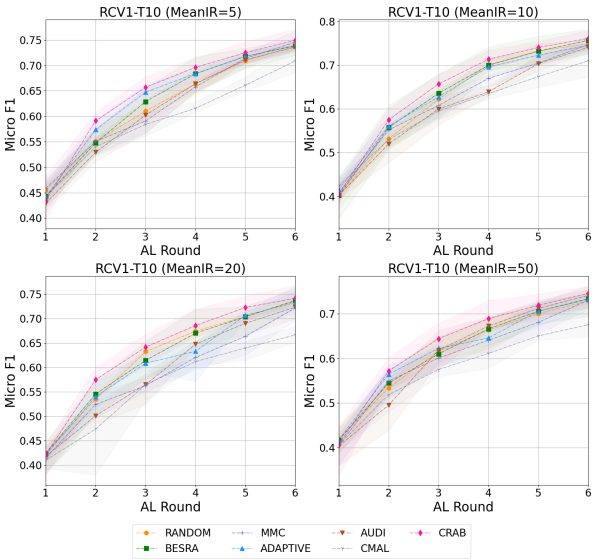


Figure 5: Averaged micro-F1 score on synthetic dataset using TextCNN, averaged results with 5 random seeds.

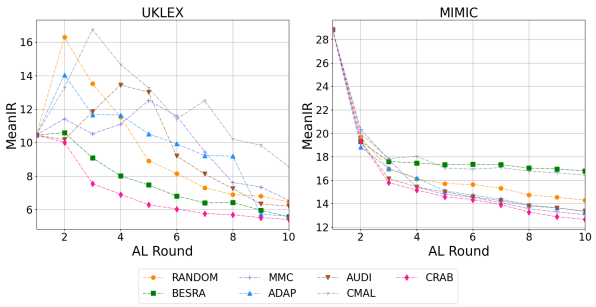


Figure 6: The averaged MeanIR of selected samples, averaged the results with 5 random seeds.

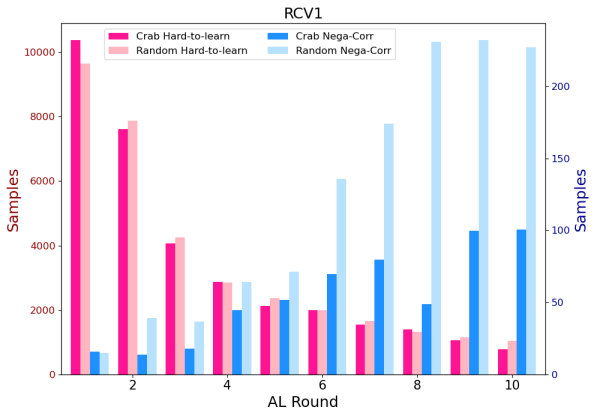
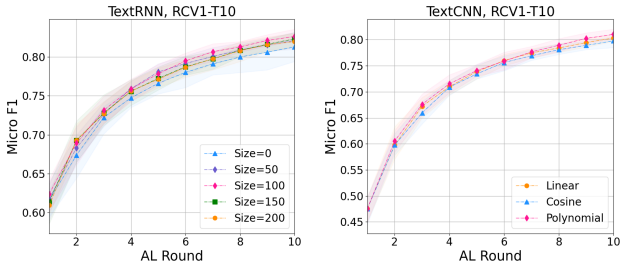


Figure 7: Trend of hard-to-learn and negative-correlated data, with the red bar axis on left and the blue bar axis on right, averaged the results with 5 random seeds.



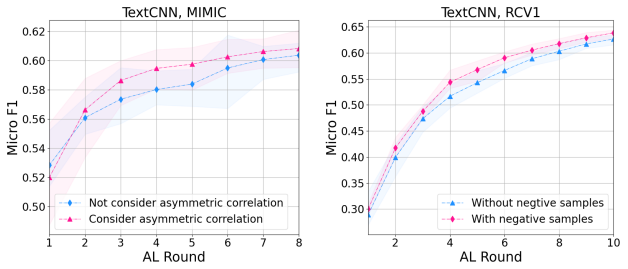
(a) (b)

Figure 8: (a) Performance for different size of hard-to-learn samples. (b) Performance for different decay functions of the hard-to-learn samples.

relatively better performance due to the higher proportion of hard-to-learn samples in the early stages. However, as annotated data increases, performance declines because the hard-to-learn samples become less influential, necessitating a reduction in their selection. This parameter is adjustable across different datasets and model structures to ensure compatibility with the learning capabilities across varying scenarios. To deal with the problem of the amount of hard-to-learn samples decreasing with increased annotated data, CRAB adopts a decay function for the size of hard-to-learn samples to adapt to the training process. Figure 8b presents performance of three decay approaches, linear decay, cosine decay, and polynomial decay. Among these, polynomial decay achieves superior performance in terms of micro-F1 score, as it produces an accelerated decrease in output for sampling, better aligning with the trend in the size of hard-to-learn samples.

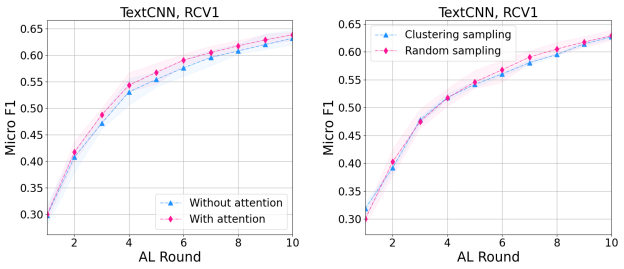
4.5 ABLATION STUDY

We conducted four experiments to examine whether the structure of CRAB improves MLAL performance by con-



(a) (b)

Figure 9: (a) Ablation study of the asymmetric-correlated label. (b) Ablation study of considering negatively correlated label pairs.



(a) (b)

Figure 10: (a) Ablation study of the correlation attention. (b) Performance for random sampling or cluster-based sampling for refined sampling pool selection.

sidering correlation. Taking into consideration the asymmetrically correlated label relationships, CRAB selects only the initial label in the correlation chain for per-label selection, thus avoiding duplicate selection of correlated labels. Figure 9a compares performance with and without considering asymmetrical correlations on the MIMIC dataset. Results illustrate that CRAB demonstrates superior performance and is more effective in querying indicative samples than when treating all labels equally. Figure 9b shows the benefits of sampling conflicted labels, with performance improvements becoming more pronounced in later training stages.

Figure 10a illustrates how the correlation attention impact MLAL accuracy. To assess performance without correlation in score evaluation, we removed the correlation attention in Eq. (13), with the results shown by the blue line. Evidently, when positive label correlations are incorporated, CRAB performs more consistently throughout the experiment, indicating that our strategy effectively models inter-label relationships to make more informative queries. Additionally, we compared the micro-F1 score and computation time of the random sampling and clustering-based sampling during the refined unlabeled pool sampling. As shown in Figure 10b, random sampling performs almost identically to clustering-based sampling, suggesting it can serve as a replacement during the refined unlabeled pool selection. Moreover, the querying time with random sampling decreases by 40% . These findings demonstrate that our method is effective, efficient, and robust.

5 CONCLUSION

In this paper, we proposed an innovative MLAL query strategy, CRAB, which takes into account inherent label relationships within a Bayesian framework. By updating the correlation matrices with the annotated data, our model is competent to query more representative samples in the initial stage and achieves a more accurate score for evaluating the informativeness of instances. Additionally, with the utilization of beta scoring rules, our model maintains consistently robust performance on imbalanced datasets. Leveraging pseudo labels and correlation-aware sampling, our strategy eliminates the need for additional training modules, and our model demonstrates significant performance improvements in MLAL on four benchmark datasets. Future research could explore the correlation at the instance space and investigate additional relationships between the data features and label distributions.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable and helpful comments. This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, and U Rajendra Acharya. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021. ISSN 1566-2535.

Maxime Arens, Lucile Callebert, Mohand Boughanem, and José G Moreno. Rebalancing label distribution while eliminating inherent waiting time in multi label active learning applied to transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13621–13632, 2024.

Felix Buchert, Nassir Navab, and Seong Tae Kim. Toward label-efficient neural network training: Diversity-based sampling in semi-supervised active learning. *IEEE Access*, 11:5193–5205, 2023.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification:

Structure and applications. *Working draft, November*, 3: 13, 2005.

Ilias Chalkidis and Anders Søgaard. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, 2022.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex-a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, 2021.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017.

Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015.

Shuyue Chen, Ran Wang, Jian Lu, and Xizhao Wang. Stable matching-based two-way selection in multi-label active learning with imbalanced data. *Information Sciences*, 610:281–299, 2022.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7949–7962, 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Kailun Gong and Tingting Zhai. An online active multi-label classification algorithm based on a hybrid label query strategy. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 463–468. IEEE, 2021.

Yan Gu, Jicong Duan, Hualong Yu, Xibei Yang, and Shang Gao. Plvi-ce: a multi-label active learning algorithm with simultaneously considering uncertainty and diversity. *Applied Intelligence*, 53(22):27844–27864, 2023.

Xue Han, Qing Wang, Yitong Wang, Jiahui Wang, Chao Deng, and Junlan Feng. Feature mixing-based active learning for multi-label text classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10551–10555. IEEE, 2024.

Patrick Hemmer, Niklas Kühl, and Jakob Schöffer. Deal: Deep evidential active learning for image classification. *Deep Learning Applications, Volume 3*, pages 171–192, 2022. ISSN 9811633568.

Jiayu Huang, Nazbanoo Farpour, Bingjian J Yang, Muralidhar Mupparapu, Fleming Lure, Jing Li, Hao Yan, and Frank C Setzer. Uncertainty-based active learning by bayesian u-net for multi-label cone-beam ct segmentation. *Journal of Endodontics*, 50(2):220–228, 2024. ISSN 0099-2399.

Jun Huang, Guorong Li, Shuhui Wang, Zhe Xue, and Qingming Huang. Multi-label classification by exploiting local positive and negative pairwise label correlation. *Neurocomputing*, 257:164–174, 2017.

Rui Huang and Liuyue Kang. Local positive and negative label correlation analysis with label awareness for multi-label classification. *International Journal of Machine Learning and Cybernetics*, 12(9):2659–2672, 2021.

Sheng-Jun Huang and Zhi-Hua Zhou. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th international conference on data mining*, pages 1079–1084. IEEE, 2013.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren. Active learning with complementary sampling for instructing class-biased multi-label text emotion classification. *IEEE Transactions on Affective Computing*, 14(1):523–536, 2020.

Yova Kementchedjhieva and Ilias Chalkidis. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5828–5843, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.

360. URL https://aclanthology.org/2023.findings-acl.360.

SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3944–3953, 2023.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Xin Li and Yuhong Guo. Active learning with multi-label svm classification. In *IjCAI*, volume 13, pages 1479–1485. Citeseer, 2013.

Liqun Liu, Funan Mu, Pengyu Li, Xin Mu, Jing Tang, Xingsheng Ai, Ran Fu, Lifeng Wang, and Xing Zhou. Neuralclassifier: an open-source neural hierarchical multi-label text classification toolkit. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–92, 2019.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879, 2016.

Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9274–9283, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs. In *Empirical Methods in Natural Language Processing 2020*, pages 2935–2943. Association for Computational Linguistics (ACL), 2020.

Dwarikanath Mahapatra, Behzad Bozorgtabar, Zongyuan Ge, Mauricio Reyes, and Jean-Philippe Thiran. Combining graph transformers based multi-label active learning and informative data augmentation for chest xray classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21378–21386, 2024.

Xue-Yang Min, Kun Qian, Ben-Wen Zhang, Guojie Song, and Fan Min. Multi-label active learning through serial–parallel neural networks. *Knowledge-Based Systems*, 251: 109226, 2022.

Lars Möllenbrok and Begüm Demir. Active learning guided fine-tuning for enhancing self-supervised based multi-label classification of remote sensing images. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4986–4989. IEEE, 2023.

Lars Möllenbrok, Gencer Sumbul, and Begüm Demir. Deep active learning for multi-label classification of remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2023.

Felipe Kenji Nakano, Ricardo Cerri, and Celine Vens. Active learning for hierarchical multi-label classification. *Data Mining and Knowledge Discovery*, 34(5):1496–1530, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Carlos Perales-González, Mariano Carbonero-Ruz, Javier Perez-Rodriguez, David Becerra-Alonso, and Francisco Fernández-Navarro. Negative correlation learning in the extreme learning machine framework. *Neural Computing and Applications*, 32:13805–13823, 2020.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. Sampling bias in deep active classification: An empirical study. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4058–4068, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1417. URL https://aclanthology.org/D19-1417.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. ISSN 0360-0300.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, volume 1, page 5. Citeseer, 2001.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*, 2019.

Weishi Shi, Dayou Yu, and Qi Yu. A gaussian process-bayesian bernoulli mixture model for multi-label active learning. *Advances in Neural Information Processing Systems*, 34:27542–27554, 2021.

Raquel Simao, Marília Barandas, David Belo, and Hugo Gamboa. Study of uncertainty quantification using multi-label ecg in deep learning models. In *BIOSIGNALS*, pages 252–259, 2023.

Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. In *2020 25th International conference on pattern recognition (ICPR)*, pages 1220–1227. IEEE, 2020. ISBN 1728188083.

Guoliang Su, Zhangquan Wu, Yujia Ye, Maoxing Chen, and Jun Zhou. Cost-efficient multi-instance multi-label active learning via correlation of features. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 410–414. IEEE, 2023.

Wei Tan, Ngoc Dang Nguyen, Lan Du, and Wray Buntine. Harnessing the power of beta scoring in deep active learning for multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15240–15248, 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *31st Annual Conference on Neural Information Processing Systems (NIPS)*, volume 30 of *Advances in Neural Information Processing Systems*, 2017. URL <GotoISI>://WOS:000452649406008.

Min Wang, Tingting Feng, Zhaohui Shan, and Fan Min. Attribute and label distribution driven multi-label active learning. *Applied Intelligence*, 52(10):11131–11146, 2022.

Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8068–8078, 2022.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, 2009.

Yang Yang, Yuxuan Zhang, Xin Song, and Yi Xu. Not all out-of-distribution data are harmful to open-set active learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xiangliang Zhang. Cmal: Cost-effective multi-label active learning by querying subexamples. *IEEE Transactions on Knowledge and Data Engineering*, 34(5):2091–2105, 2020.

Ye Zhang and Byron C Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263, 2017.

Yuanjian Zhang, Tianna Zhao, Duoqian Miao, and Witold Pedrycz. Granular multilabel batch active learning with pairwise label correlation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(5):3079–3091, 2021.

Rui Zheng, Shulin Zhang, Lei Liu, Yuhao Luo, and Mingzhai Sun. Uncertainty in bayesian deep label distribution learning. *Applied Soft Computing*, 101:107046, 2021.

Yu Zhu, Jinghao Lin, Shibi He, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. Addressing the item cold-start problem by attribute-driven active learning. *IEEE Transactions on Knowledge and Data Engineering*, 32 (4):631–644, 2019.