

RESEARCH ARTICLE

Enhanced Active Learning Through Exclusion of Semi-Informative Sets

KYUNGWOOK MIN¹, SEUNGUK YU¹, (Graduate Student Member, IEEE),
YOONJI LEE¹, (Graduate Student Member, IEEE), AND YOUNGBIN KIM^{1,2}, (Member, IEEE)

¹Department of Artificial Intelligence, Chung-Ang University, Dongjak, Seoul 06974, South Korea

²Graduate School of Imaging Science, Multimedia and Film, Chung-Ang University, Dongjak, Seoul 06974, South Korea

Corresponding author: Youngbin Kim (ybkim85@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship, in 2025, Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) [Artificial Intelligence Graduate School Program (Chung-Ang University)] under Grant RS-2021-II211341; and in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant RS-2025-00556246.

ABSTRACT In this study, we explore methods for minimizing labeling costs, which is a critical issue in the active learning process. We focus on efficiently selecting the most essential datasets from an entire pool of unlabeled datasets. Alongside previous methodologies that sample datasets based on model uncertainty and feature map diversity, we propose a novel approach aimed at budget optimization. Notably, we introduce the concept of *semi-informative sets*, which allows for the effective exclusion of data that contribute relatively little to learning. This approach aims to maximize budget efficiency, leading to two main strategies: The first strategy involves refining the prediction probabilities of the model by applying a label-smoothing technique to improve the loss function that is directly associated with the prediction accuracy. The second strategy centers on the effective selection of high-confidence data through a more precise threshold setting. To achieve this, we introduce a learnable threshold that is adjusted based on the reliability of the already learned data. These approaches present a novel direction for active learning that considers budget constraints and potentially contributes to the development of practical active learning systems for real-world applications.

INDEX TERMS Active learning, semi-informative sets, budget-constrained.

I. INTRODUCTION

Recent advances in artificial intelligence and machine learning are transforming various sectors of society. These methods are driven by the development of effective learning algorithms and the availability of diverse datasets, which together largely determine model performance. However, labeling of these large datasets is costly and time-consuming. This challenge is particularly pronounced in scenarios in which labeling is expensive or data are scarce, such as diagnosing specific diseases in magnetic resonance imaging (MRI) [1], [2], which requires specialized human resources, or collecting data in complex environments, such as autonomous driving [3], [4], [5]. To address these issues, active learning has emerged as a promising approach for efficiently handling the given datasets [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang¹.

Unlike traditional passive learning, active learning involves the learning algorithm actively selecting and optimizing the training datasets. This method prioritizes the quality of the dataset over the quantity with the aim of maximizing the learning efficiency with minimal dataset. This is particularly beneficial when constructing datasets from scratch and in areas in which labeling is expensive and data collection is difficult. A study by OpenAI showed that employing active learning for dataset construction made the process 2.6 times more efficient [7]. In situations where quantitative data are available, active learning enhances efficiency by avoiding indiscriminate use of the dataset and facilitating a more strategic exploration process. Furthermore, recent research has focused on data integration and processing using IoT and AI, highlighting the increasing importance of efficient data handling in real-world applications [30], [31].

Despite the rapid advancements in this technique using various approaches, there remains a notable gap in the previous research, namely efficient budgeting under limited financial

constraints. While it is important to maximize performance by significantly increasing model size and expanding the dataset, however, in real-world scenarios, budget limitations significantly influence the learning strategies and model performance [8]. In this paper, we focus on both theoretical and practical methodologies for developing efficient strategies under budget constraints. We propose a method that complements existing active learning approaches, aiming to maximize model performance through the efficient use of data within a constrained budget. Our goal is to provide concrete strategies that maximize the effectiveness and cost-efficiency of active learning in real-world applications.

The primary objective of our study is to enhance the efficiency of active learning through budget optimization. Active learning is known for maximizing the data's qualitative value while minimizing labeling costs [6]; however, budget constraints remain a critical factor in practical applications. While most existing research has focused on methodologies to boost the active learning performance, the realistic issue of limited budgets has often been overlooked. Therefore, we explore novel approaches for optimizing active learning within the constraints of real-world budgets.

This study explores the concept of cost-effective active learning [8], focusing on reducing labeling costs while enhancing existing active learning techniques. We introduce the concept of *semi-informative sets* to optimize the use of the labeling budget more efficiently. The proposed *semi-informative sets* help to identify and exclude datasets that are relatively unnecessary for model training from existing labeled datasets, thereby allowing labeling resources to focus on more valuable data.

We propose two main strategies for defining *semi-informative sets*. The first strategy introduces a loss function with a label-smoothing technique to achieve high confidence, thereby increasing the size of the *semi-informative sets*. Consequently, the second strategy involves a system to determine the threshold for learnable confidence. Using these methodologies, we employ *semi-informative sets* that are suitable for a constrained budget and can enable active learning in practical applications. In particular, the approach utilizing *semi-informative sets* is designed to be applicable to existing active learning methods. Experimental results showed that applying our proposed approach improved performance in almost all cases and that integrating *semi-informative sets* with different baselines can achieve high performance without extensive training. This enables efficient learning by strategically exploring the dataset within a limited budget, avoiding unnecessary resource expenditure on the model. Consequently, when applying the proposed method in practical applications, the number of training cycles can be flexibly adjusted based on the situation, allowing for more efficient use of data and better budget management.

The contributions of this study are as follows:

- We present practical approaches for applying active learning in budget-constrained scenarios, strategically exploring the dataset to maximize training efficiency.

- We introduce the concept of *semi-informative sets* to enhance efficient exploration under a limited budget by employing the two strategies: loss function with label-smoothing and learnable threshold.
- Our approach can be easily integrated into existing active learning methodologies, with performance improvements expected from the application of *semi-informative sets*.

The remainder of this paper is organized as follows. Section II provides an overview of active learning and the related studies on active learning strategies. Next, Section III describes the theoretical foundations and implementation of the proposed approach. Then, Section IV outlines the experimental setup, and Section V presents the results and analyses that validate the effectiveness of the proposed approach. Section VI discusses potential risk factors that might impact the experimental results, while Section VII highlights the limitations of this study and proposes directions for future work. Finally, Section VIII summarizes our work and discusses avenues for future research.

II. RELATED WORK

A. ACTIVE LEARNING

Active learning plays an important role in the construction of datasets for machine learning tasks. The goal of active learning is the selection of training datasets based on the uncertainty and diversity. This process reduces the labeling costs and improves model performance by focusing on the most informative datasets, rather than labeling the entire unlabeled datasets. In addition, the most valuable data points for training the model were labeled accordingly using various metrics. For example, if the metric is the uncertainty of the model, the data points that the model considered as uncertain are prioritized for labeling because they are likely to provide additional information, improving the learning process of the model. However, using diversity as a metric aids in capturing the entire data distribution, which helps the model to generalize better and minimizes bias as new datasets are provided. This results in a more balanced data set, especially increasing the model accuracy and reliability.

1) UNCERTAINTY-BASED ACTIVE LEARNING

Uncertainty-based active learning is a traditional method that identifies the data points that the model is most uncertain about predicting in an unlabeled dataset. This approach is based on the assumption that the data points that are most confusing to the model are the most useful for training. Several studies have been proposed for measuring uncertainty, with the most common metric based on predictive probability. Other notable measures include the Least Confidence [9], which selects data points for which the model has the least confidence; Margin Confidence [10], which selects data points with the smallest probability difference between the two most likely classes in a classification model; Bayesian Active Learning by Disagreement (BALD) [11] and Expected Predictive Information Gain (EPIG) [12], which leverage

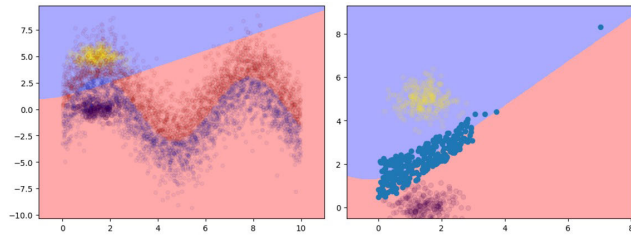


FIGURE 1. Problems that can arise during uncertainty-based active learning. Even with sufficient data, when labeled data are biased, performing uncertainty-based active learning may lead to overfitting. (Left): Scatter plots of unlabeled data points in red and blue and labeled data points in yellow and purple, and the decision boundaries of the classifier trained on the initial dataset, background in blue and red. (Right): Data points in cyan are selected based on the uncertainty of the classifier regarding the unlabeled dataset shown on the left.

the predictive uncertainty of Bayesian models. In addition, Learning Loss [13] determines the useful data by evaluating the expected loss that the model would incur for particular data points. These approaches focus on improving areas in which the model currently struggles during the training process. By prioritizing data points with high uncertainty, the model can quickly ascend the learning curve and improve the overall learning efficiency. However, as shown in Fig. 1, if the model focuses solely on misinterpreted patterns, it may become biased by overlooking other important data features. This can lead to overfitting, which reduces the generalization capacity of the model, and results in poor performance.

2) DIVERSITY-BASED ACTIVE LEARNING

This approach focuses on maximizing data diversity. It aims to select data points that represent the overall data distribution, thereby improving the overall efficiency and performance of the model training. Representative methods include clustering-based active learning [14], [32], [33], in which important data points are selected from each cluster to capture the different characteristics of the entire dataset. The effectiveness of diversity-based selection is highly dependent on the quality of the learned feature representations, and recent work has shown that transfer learning can improve discriminative feature extraction in data-limited scenarios [34]. Such approaches demonstrate that better feature embeddings can enhance clustering quality and downstream selection performance.

Another method is Core-set [15], in which small but important points are selected from the dataset. In addition, because diversity is primarily achieved by clustering data points in the feature space, certain methods ensure diversity using probabilistic models in the feature space [16]. This approach is particularly useful for large-scale datasets such as ImageNet [36], where full uncertainty-based querying can be computationally prohibitive and distributional coverage becomes more critical than boundary refinement. However, performance may vary depending on the quality of feature extraction, and precise decision boundaries may still be difficult to determine.

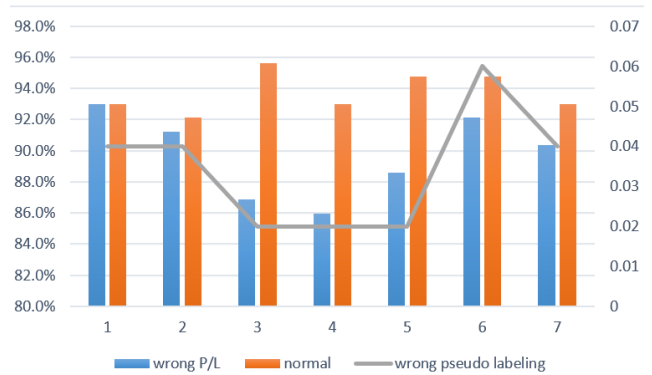


FIGURE 2. Performance differences per epoch due to incorrect pseudo labeling (P/L) in cost-effective active learning. The blue and orange bars represent the metrics for incorrect pseudo labeling and standard active learning, respectively, while the gray line indicates the proportion of incorrectly assigned labels in pseudo labeling. The case with incorrect pseudo labeling showed relatively unstable performance differences as the epochs progressed.

3) HYBRID-BASED ACTIVE LEARNING

Hybrid-based active learning, which combines uncertainty and diversity into a composite strategy, is a more recently researched approach. BatchBald [17] utilized both uncertainty and gradient information to enhance the dataset diversity. The gradient lower bounds of the model are used to select the most uncertain and diverse data points. This method is particularly useful for batch training in deep neural networks. In addition, a study that proposed the Exploration-P methodology [18] highlighted a limitation in calculating the instance similarity based solely on feature vectors. It suggests that the accuracy of the diversity measurements can be enhanced by learning the similarity in conjunction with a classification model. Unlike traditional methods that query data points near decision boundaries, this method employs an efficient query strategy that selects data based on both uncertainty and diversity, even with a limited number of labels. Such hybrid strategies have shown strong empirical performance on medium-scale image classification benchmarks such as CIFAR-10 [35], where balancing decision-boundary refinement with representative sampling improves learning efficiency under limited annotation budgets.

4) BUDGET-CONSCIOUS ACTIVE LEARNING

In recent years, various methods have been explored to enhance performance with minimal budget. These approaches aim to extract representative data points for each label by computing the typicality; TypiClust forms clusters based on the most typical data points in a dataset [19]. This approach focuses on determining the most representative data points within each cluster by considering the different data characteristics. ProbCover selects data points that cover the entire distribution of a dataset as uniformly as possible [20]. This approach is based on the concept of probabilistic coverage, which aims to select data points covering the entire data space as broadly as possible. The authors showed

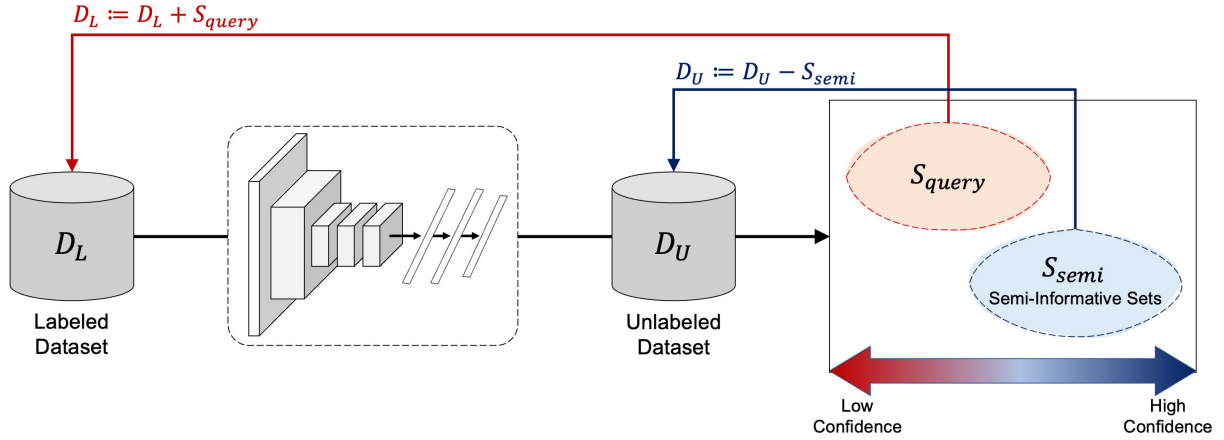


FIGURE 3. Structure of the proposed approach utilizing *semi-informative sets*. It involves training a model using a labeled dataset to obtain *semi-informative sets* with relatively high confidence from the unlabeled dataset. Data points deemed to have sufficient confidence are excluded from subsequent processes, while the same procedure is repeated for datasets identified as requiring additional labeling.

that determining typical data points helps to improve the performance when the budget is small. Conversely, non-typical active learning strategies and random sampling improve the performance more than sampling typical data points as the budget increases. This explains why the probability of selecting data that accounts for uncertainty, diversity, and typicality increases as the budget grows.

B. COST-EFFICIENT ACTIVE LEARNING

Cost-effective active learning follows the core principles of active learning, with a specific focus on reducing labeling costs. This methodology uses pseudo-labeling [21] for high-confidence unlabeled datasets in semi-supervised learning. An important part of this process is the measurement of confidence and uncertainty in the model. The confidence and uncertainty can be calculated from the class-wise prediction probabilities or Bayesian approaches [22]. Data points with high confidence that are calculated through these measures, provide additional information in the learning process via pseudo-labeling, contributing to reduced labeling costs.

However, incorrect pseudo-labeling can negatively affect the model performance [23]. Fig. 2 presents the test accuracy of cost-effective active learning with incorrect pseudo labeling applied to the Breast Cancer dataset, compared to the accuracy without using the mislabeled dataset from the incorrect pseudo labeling. The results showed that cost-effective active learning with incorrect pseudo-labeling underperforms compared to general active learning. This outcome was remarkable, even though the mislabeling rate remained between just 0.02 and 0.06. Therefore, developing methods that reduce the risk of incorrect pseudo-labeling is essential. This study introduces a novel approach grounded in cost-effective active learning, which incorporates budget constraints for high-confidence data. Our goal is to manage data labeling costs while optimizing model performance.

III. METHOD

We introduce *semi-informative sets* for efficient data extraction within a constrained budget, addressing the limitations of pseudo-labeling in cost-effective active learning. The proposed architecture with the introduced *semi-informative sets* S_{semi} is illustrated in Fig. 3. It consists of label-like information about data points, determined by the confidence of the model from the remaining unlabeled dataset.

The process involves training a model using the labeled dataset D_L , calculating the confidence levels of the unlabeled dataset D_U with the trained model, and assessing uncertainty or variance to prioritize the data points that require labeling. It sorts the data points in D_U based on confidence levels, designates datasets consisting of high-confidence data points as S_{semi} according to a threshold, and excludes S_{semi} from D_U . Concurrently, it identifies the query set S_{query} for labeling requests based on the sampling priority within the set budget. As a result, we enhance data efficiency by excluding *semi-informative sets* from the unlabeled dataset that already exhibit high confidence. The model training process is then applied only to those data points in S_{query} that require additional labeling.

A. DEFINING SEMI-INFORMATIVE SETS

$$M_B \leq Budget \leq |D_U|, \quad (1)$$

$$1 \leq M_B \leq |D_U|, \quad D_U := D_U - S_{semi}, \quad (2)$$

$$\text{minimize}(D_U) = \text{maximize}(S_{semi}), \quad (3)$$

First, for efficient budgeting, the budget can range between M_B , which is the smallest value of the given budget, and the size of unlabeled dataset $|D_U|$ (Equation (1)); and accordingly, M_B can have a value between 1 and $|D_U|$. However, because D_U is subject to the process of excluding *semi-informative sets* S_{semi} (Equation (2)), the size of the S_{semi} should be increased from the fixed D_U to ensure the minimum value of the budget (Equation (3)).

To optimize budget usage within this situation, we exclude the *semi-informative sets* S_{semi} with high confidence from the unlabeled dataset D_U , which represents the area for exploration during budget allocation. This approach imposes constraints on the remaining space, allowing the model to focus its exploration more effectively. Maximizing the size of the *semi-informative sets* to exclude from the unlabeled dataset is a key objective for enhancing budget efficiency. By increasing the size of S_{semi} , we can concentrate on more valuable data points within a limited budget, thereby minimizing waste by eliminating the selection of data points that are considered to provide a sufficient amount of information.

B. MAXIMIZING PREDICTION CONFIDENCE

We present strategies for enhancing the prediction confidence to maximize the size of *semi-informative sets* S_{semi} . In classification tasks, the predictive confidence can be derived from the probabilities assigned to each class. High prediction confidence occurs when the probability assigned to the correct class significantly exceeds the probabilities assigned to the other classes. Therefore, improving prediction reliability involves not only increasing the probability of the correct class but also improving the overall probability distribution by lowering the probabilities assigned to the incorrect classes.

Enhancing predictive probabilities can be categorized into two ways: model performance and training methods. Model performance, which encompasses factors beyond the data, such as the model type and architecture, is often not considered in active learning. In the context of classification model training, the commonly used cross-entropy loss calculates the loss based on the binary distribution of true label information from one-hot encoding and the predictive probability for the correct class [24]. In this setup, the model learns solely from correct answers, lacking explicit cues to differentiate between incorrect ones.

To address this limitation, we utilize label smoothing [25], which modifies the loss function to penalize incorrect predictions, thereby refining the overall probability distribution. This approach enhances the model's ability to differentiate between correct and incorrect predictions, ultimately improving prediction confidence.

$$Loss_{CE} = -\sum_{c=1}^N \log(p_c) y_c, \quad (4)$$

$$y'_c = y_c(1 - \alpha) + \frac{\alpha}{N}, \quad (5)$$

$$Loss_{ours} = -\sum_{c=1}^N \log(p_c) y'_c, \quad (6)$$

When the conventional cross-entropy formula is defined as shown in Equation (4), label smoothing addresses overconfidence by assigning a small, uniform probability to the other labels, represented as y'_c . The softmax probability corresponding to the given label is denoted as p_c and α is a hyperparameter. By integrating this adjustment into the original loss function, we were able to leverage even

Algorithm 1 Training With *semi-informative sets*

```

1: Split labeled, unlabeled dataset  $D_L, D_U$ 
2: Set maximum cycle iteration  $cycle_{max}$ 
3: Set image classification model  $M$ 
4: Set threshold parameter  $\beta$ 
5: Initialize current cycle iteration  $cycle \leftarrow 0$ 
6: while  $cycle < cycle_{max}$  do
7:   Train  $M$  with  $D_L$ 
8:   Initialize high-confidence sets  $S_{query} \leftarrow \{\}$ 
9:   Initialize semi-informative sets  $S_{semi} \leftarrow \{\}$ 
10:  Set learnable threshold  $\tau$  in Equation (9)
11:  for each data point in  $D_U$  do
12:    Set current confidence  $Conf = M(\text{data point})$ 
13:    if  $Conf < \tau$  then
14:      Store current data point in  $S_{semi}$ 
15:    else
16:      Store current data point in  $S_{query}$ 
17:  Update  $D_L += S_{query}$ 
18:  Update  $D_U -= S_{semi}$ 
19:  Update  $cycle += 1$ 

```

uncertain information in the process of selecting *semi-informative sets*.

C. LEARNABLE CONFIDENCE THRESHOLD

Previous studies have proposed a confidence threshold δ for pseudo-labeling in cost-effective active learning, as shown in Equations (7) and (8). Initially, they set a threshold δ_0 and allowed it to be adjusted linearly by dr in small increments for each iteration of active learning. However, it depends on a static threshold that does not change with the model's training progress, as it is updated solely based on the variable t , representing the number of active learning iterations, along with fixed constants δ_0 and dr .

$$\delta = \delta_0 \quad \text{when } t = 0, \quad (7)$$

$$\delta = \delta - dr * t \quad \text{when } t > 0, \quad (8)$$

To address this limitation, we propose a threshold that goes beyond the concept of linearly varying thresholds based solely on hyperparameter specifications, instead utilizing some of the results from the trained model, as shown in Equation (9). This process involves training with the labeled data D_L . Once training is complete, the model computes confidence by combining uncertainty information $F_{uncertainty}(D_L)$ and the accuracy of the labeled data Acc_{train} . This approach dynamically sets the threshold τ to define the dataset S_{query} for training, based on insights gained during model training.

$$\tau = \beta * Acc_{train} * F_{uncertainty}(D_L), \quad (9)$$

Leveraging both information to learn the threshold enables the model to calibrate its confidence in a data-dependent manner, reflecting both distributional ambiguity and empirical reliability. Furthermore, we set a hyperparameter β that

enables the model to adjust the confidence level of the training data, which was set to 0.7 in our experiments. This enhancement establishes a clear threshold τ , informed by the learning process, ensuring the acquisition of more reliable datasets. Therefore, our proposed approach excludes *semi-informative sets* S_{semi} by using a threshold τ based on the current state of the dataset, allowing the model to focus more on low-confidence data.

We present the training process utilizing the *semi-informative* in Algorithm 1 through pseudo code. The model undergoes iterative training up to a predefined maximum number of cycles. In each cycle, the model trained on the labeled dataset D_L is applied to the unlabeled dataset D_U , assessing the confidence of its data points. Based on the learnable threshold, D_U is divided into S_{query} or S_{semi} , and this classification directly impacts the subsequent training cycles by prioritizing data points based on their confidence. By repeating this process, we progressively filter and select data points that offers greater value for training.

IV. EXPERIMENTAL SETTINGS

A. DATASET AND ENVIRONMENTS

Since image classification is the primary task of active learning in this study, we employed the ResNet architecture to train the labeled dataset. We utilized a dataset of short videos extracted from movie trailers to assess the effectiveness of our proposed approach in capturing semantic information across frames and identifying nuanced differences. In addition, we evaluated our approach using the MNIST dataset, which included a random imbalance in its construction. This dataset contains images of handwritten digits created by individuals, intended for use in classification tasks. Specifically, the imbalance was designed so that classes 0 to 4 contained data sizes ten times smaller than those of classes 5 to 9. This evaluation aimed to determine whether our approach of excluding *semi-informative sets* remains effective in the context of imbalanced datasets.

The experiments were conducted using the PyTorch framework on a single T4 GPU. The overall experiments consisted of three independent trials, each comprising seven training cycles. We averaged the test accuracy across the three trials for each cycle, applying a budget constraint of 50 trials per active learning cycle. To introduce diversity in applying model uncertainty, we used a subset comprising one-fifth of the full training dataset. The model was trained for 200 epochs, starting with an initial learning rate of 0.1, which was adjusted at predefined milestones. Gradient propagation from the loss prediction module was halted after 120 epochs. Additional parameters included a momentum of 0.9 and a weight decay of $5e-4$. Furthermore, the parameter α for label smoothing was set to 0.95.

B. BASELINES

The proposed approach utilizing *semi-informative sets* can be applied independently of the model selection. Therefore,

we selected various active learning baseline methods and compared the differences before and after applying our approach to each model. The baseline methods used are as follows.

- BALD [11] is a Bayesian approach that selects data points with the highest model uncertainty, based on the disagreement among different predictions sampled from the Bayesian posterior.
- Entropy-based [26] method is based on calculating the entropy of each data points to identify those with high entropy values which means a high degree of uncertainty.
- Learning Loss [13] proposes a loss prediction module and incorporates into a deep neural network to predict the loss of unlabeled data points, enabling the model to select the most informative samples.
- Least Confidence [9] utilizes model confidence to select data points which shows lowest confidence in its predictions, reflecting that the model is least certain about its predictions.
- Margin Confidence [10] identifies data points where the difference in predicted probabilities between the two most likely classes is smallest in classification and aims to target the most ambiguous cases.
- Ratio Confidence [27] utilizes ratio rather than probability, to select data points where the ratio between the scores of the two most likely predicted classes is the smallest.

C. EVALUATION METRICS

The evaluation compared test accuracy between traditional active learning methodologies and our proposed approach, which excluded *semi-informative sets*. We closely monitored test accuracy across iterations to assess the rate of performance enhancement. Additionally, we investigated the impact of class imbalance, which is an atypical data scenario, on the iterative improvement of test accuracy.

Each evaluation was conducted three times to ensure consistency; therefore, we report the mean test accuracy, calculated as follows:

$$Acc_{test} = \frac{1}{3} \sum_{k=1}^3 \frac{\# \text{ of correct predictions}}{\text{total \# of predictions}}. \quad (10)$$

V. DISCUSSION

A. EXPERIMENTAL RESULTS

We compared the performance of baseline active learning methods with and without the proposed approach using *semi-informative sets*, as shown in Table 1. In almost every case, excluding *semi-informative sets* S_{semi} with high confidence led to a significant performance improvement. This suggests that even within a limited budget, without the need for additional data or larger models, optimized data sampling can be achieved by strategically excluding *semi-informative sets* thereby preventing unnecessary budget expenditure.

TABLE 1. Quantitative experimental results comparing the average accuracy per cycle for each method when utilizing the movie trailers dataset. We compared the baseline methods with our approach that excludes the *semi-informative* sets. In most cases, the methods using our proposed approach exhibited superior performances. The performance of the best model is highlighted in bold and the models that showed the greatest performance improvement with our approach are underlined in each iteration.

Method	Metric	1	2	3	4	5	6	7
BALD	Original	45.62	68.10	86.25	87.43	90.20	93.12	94.98
	w/o S_{semi}	60.45	74.19	83.96	85.97	88.59	93.21	95.31
	ΔAcc_{test}	+14.83	+6.09	-2.29	-1.45	-1.61	+0.09	+0.33
Entropy-based	Original	52.42	61.09	76.70	85.54	89.81	93.15	95.50
	w/o S_{semi}	58.54	68.73	82.65	89.98	93.13	94.81	95.80
	ΔAcc_{test}	+6.12	+7.64	+5.95	+4.44	+3.32	+1.66	+0.30
Learning Loss	Original	46.78	53.44	61.57	67.29	68.24	72.93	74.79
	w/o S_{semi}	55.23	67.31	75.94	77.69	81.40	83.83	85.26
	ΔAcc_{test}	+8.46	<u>+13.87</u>	<u>+14.37</u>	<u>+10.41</u>	<u>+13.16</u>	<u>+10.90</u>	<u>+10.47</u>
Least Confidence	Original	47.77	73.26	85.32	92.11	94.22	95.80	96.83
	w/o S_{semi}	56.29	76.18	86.94	92.26	93.80	95.58	96.15
	ΔAcc_{test}	+8.52	+2.92	+1.62	+0.16	-0.43	-0.21	-0.68
Margin Confidence	Original	41.13	66.95	85.43	90.73	93.33	95.39	96.23
	w/o S_{semi}	54.91	75.25	88.01	92.63	94.23	95.55	96.60
	ΔAcc_{test}	+13.78	+8.29	+2.58	+1.90	+0.90	+0.16	+0.37
Ratio Confidence	Original	44.64	66.16	83.48	88.92	92.75	94.63	96.09
	w/o S_{semi}	60.86	76.32	88.05	91.56	94.14	95.58	96.41
	ΔAcc_{test}	<u>+16.22</u>	+10.16	+4.57	+2.64	+1.40	+0.95	+0.32

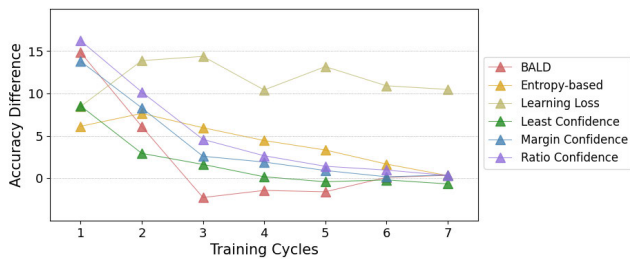


FIGURE 4. The performance differences observed as each iteration progressed after excluding the *semi-informative* sets S_{semi} for each model when utilizing the movie trailers dataset. Except for BALD, all methods in every iteration showed performance improvements when our proposed approach was applied. Notably, some methods demonstrated significant performance gains, with gaps exceeding 15 points in the early iterations.

The performance differences of each method, after applying our proposed approach across all methods and training cycles, can be observed in Fig. 4. By removing *semi-informative* sets S_{semi} during training, we observed a significant performance increase, especially in the early cycles. Notably, in cycle 1, applying our approach to Ratio Confidence and BALD improved performance by 16.22 and 14.83, respectively. This suggests that even in the very early stages of training, strategically selecting data by excluding *semi-informative* sets S_{semi} can significantly influence performance, allowing for substantial gains with fewer training cycles.

Among all methodologies, Learning Loss consistently showed a ΔAcc_{test} of nearly 10 points when our approach was applied, indicating a greater effect across all training cycles. The model showed favorable results with our proposed approach, confirming that excluding *semi-informative* sets

S_{semi} was beneficial for model training across all cycles. While other baseline methods directly address the uncertainty of specific data points by utilizing prediction probabilities, Learning Loss performs relatively indirect active learning through loss values, thereby yielding greater advantages compared to other baseline methods.

As the model was trained through to the final cycle, Margin Confidence showed the highest overall performance, reaching a score of 96.60 when our approach was applied. On the other hand, Learning Loss had the lowest absolute performance in the final cycle. However, even for this model, we observed significant performance improvements in every training cycle when applying our approach. In conclusion, regardless of the model, our method of excluding *semi-informative* sets S_{semi} led to positive performance gains, with the final results varying depending on the characteristics of the baseline methods.

B. TRAINING THE UNBALANCED DATASET

In contrast to the previous experiments, we intentionally created a situation with label imbalance to assess the performance of our approach using *semi-informative* sets S_{semi} compared to traditional baseline active learning methods. The results are presented in Table 2. In most cases, we observed improvements in performance when excluding *semi-informative* sets based on high confidence. This suggests that even when training with imbalanced label distributions, strategically excluding *semi-informative* sets S_{semi} allows for optimized learning, unaffected by data imbalance issues.

The performance differences of each method when applying our proposed approach to the unbalanced MNIST dataset across all cycles can be seen in Fig. 5. In Fig. 4, no label

TABLE 2. Quantitative experimental results comparing the average accuracy per cycle for each method when utilizing the unbalanced MNIST dataset. We compared the baseline methods with our approach that excludes the *semi-informative sets*. In most cases, the methods using our proposed approach exhibited superior performances. The performance of the best model is highlighted in bold and the models that showed the greatest performance improvement with our approach are underlined in each iteration.

Method	Metric	1	2	3	4	5	6	7
BALD	Original	37.80	46.52	62.58	67.90	72.86	77.11	81.54
	w/o S_{semi}	31.61	43.69	62.70	71.02	79.34	87.26	89.94
	ΔAcc_{test}	-6.19	-2.83	+0.12	+3.12	+6.48	+10.15	+8.40
Entropy-based	Original	37.28	42.18	44.75	51.55	65.48	78.01	83.84
	w/o S_{semi}	35.67	47.60	66.94	75.54	80.53	86.62	90.92
	ΔAcc_{test}	-1.61	+5.42	+22.19	+23.99	+15.05	+8.61	+7.08
Least Confidence	Original	36.43	50.46	73.85	80.93	84.20	89.08	91.95
	w/o S_{semi}	34.86	47.37	65.22	76.99	82.01	88.22	92.15
	ΔAcc_{test}	-1.57	-3.09	-8.63	-3.94	-2.19	-0.86	+0.20
Margin Confidence	Original	40.16	56.71	76.11	82.63	85.65	91.07	93.48
	w/o S_{semi}	37.53	51.65	73.54	81.99	86.67	91.68	93.91
	ΔAcc_{test}	-2.63	-5.06	-2.57	-0.64	+1.02	+0.61	+0.43
Ratio Confidence	Original	34.51	48.69	70.91	80.43	85.41	90.05	92.92
	w/o S_{semi}	37.17	53.10	69.03	77.88	84.40	89.28	92.29
	ΔAcc_{test}	+2.66	+4.41	-1.88	-2.55	-1.01	-0.77	-0.63

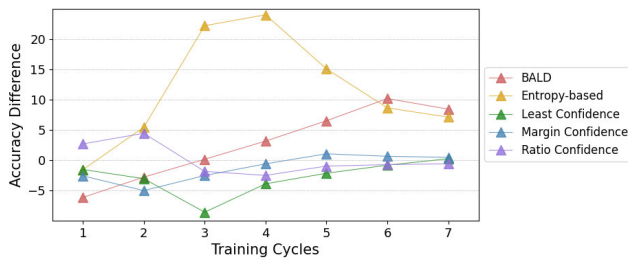


FIGURE 5. The performance differences observed as each iteration progressed after excluding the *semi-informative sets* S_{semi} for each model when utilizing the unlabeled MNIST dataset. Except for BALD, all methods in every iteration showed performance improvements when our proposed approach was applied. Notably, some methods demonstrated significant performance gains, with gaps exceeding 15 points in the early iterations.

imbalance was applied to the dataset, resulting in a significant improvement in performance during the early learning stages when using *semi-informative sets* S_{semi} for active learning. However, when we introduced label imbalance, using ten times fewer data for half of the classes compared to the others, we occasionally observed negative improvements, as seen in the case of Least Confidence.

Although the extreme label imbalance did not lead to noticeable performance improvements during the early learning stages, as seen in the previous results, we observed that our approach using *semi-informative sets* S_{semi} showed advantages as training progressed, particularly in more traditional methods such as BALD and Entropy-based approaches. Notably, in the case of Entropy-based methods, the performance improvement reached up to 23 points, demonstrating the significant benefits of applying *semi-informative sets* S_{semi} to the unbalanced dataset.

As training progressed through the final cycle of the training, we observed the best overall performance with

Margin Confidence, reaching a score of 93.91 when applying our approach with the *semi-informative sets* S_{semi} . This model consistently outperformed other baseline methods in all cycles, indicating that Margin Confidence demonstrated relative robustness in handling the label-imbalanced dataset. In conclusion, regardless of the model used, our approach of excluding *semi-informative sets* S_{semi} led to consistent performance improvements, with the extent of improvement varying depending on the characteristics of the baseline methods.

C. MAXIMIZING THE SEMI-INFORMATIVE SETS

The primary goal of the approach proposed in this study is to maximize the size of the *semi-informative sets* S_{semi} , thereby reducing the size of the unlabeled dataset D_U . To achieve this, we quantitatively assessed the size of the *semi-informative sets* S_{semi} as the training cycles progressed with the unbalanced MNIST dataset, as illustrated in Fig. 6. The values represented as the bar graph indicate the size of the *semi-informative sets* S_{semi} measured at each cycle, while the line graph depicts the test set accuracy for each corresponding cycle.

In the case of BALD, it secured the largest number of *semi-informative sets* S_{semi} compared to other active learning baseline methods, resulting in the greatest reduction in the size of the unlabeled dataset D_U . However, when examining the corresponding test set accuracy, it was found to have the lowest values among the models used in the experiment. In fact, for other models, including those with the highest absolute test set accuracy, performance was generally good when they acquired between 1,000 and 2,000 *semi-informative sets* S_{semi} in each cycle. These results indicate that simply increasing the size of the *semi-informative sets* S_{semi} does not guarantee an absolute improvement in performance, regardless of the method used.

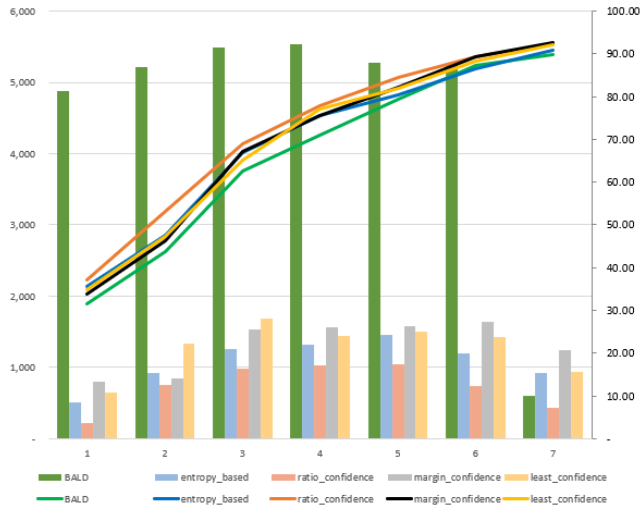


FIGURE 6. The changes in the size of the *semi-informative sets* and performance trends as training cycles for each method are shown. In all methodologies, the size of the *semi-informative sets* initially increases and then gradually decreases. As for performance, it consistently improves with repeated training cycles.

Additionally, in all methodologies, the size of the *semi-informative sets* S_{semi} initially increased slightly during the early stages of training, but began to gradually decrease after the 4th to 5th cycles. This observation suggests that, in the early phases of training, the model likely lacked sufficient knowledge to secure a relatively high number of reliable *semi-informative sets* S_{semi} . As active learning progressed and the model parameters were updated, it became increasingly possible to acquire more reliable datasets. Thus, the gradual increase in size can be attributed to this progression up to a certain number of cycles. Toward the end of the training, it appears that each method reached a stable performance level, encountering a limit in acquiring additional *semi-informative sets* S_{semi} with high confidence.

D. EFFICIENCY OF THE SEMI-INFORMATIVE SETS

Our proposed methodology involves computations such as label smoothing, uncertainty estimation, and threshold adjustment. However, as high-confidence data are gradually excluded, the volume of data that must actually be trained on decreases over time. Consequently, with each iteration of active learning, the overall training cost is reduced. Compared to Bayesian active learning methods such as BALD or approaches requiring an additional small module, our approach can be considered more cost-efficient. In BALD, the model estimates a probabilistic distribution over its parameters and calculates predictive uncertainty through multiple MC dropout samplings. These repeated forward passes significantly increase the computational burden when dealing with long sequence lengths or large-scale datasets. By contrast, our method employs a learnable threshold to select the training data, thus forgoing the need for repeated inferences and model updates.

While Bayesian approaches may excel in fine-grained uncertainty estimation for complex data distributions, using *semi-informative sets* is more practical in terms of labeling efficiency and computational cost, especially for large-scale image data or in resource-constrained environments. On the other hand, active learning based on learning loss attaches a small additional module to predict the potential loss of each sample, thereby labeling only those data points likely to yield high loss. Although this eliminates the repeated inference cost associated with MC dropout in Bayesian methods, it requires training this module alongside the main model, increasing the parameter count and potentially the computational load. Moreover, the module's initial predictions may be unstable, necessitating multiple iterations, and if its training goes awry, inaccurate sample selection can degrade overall performance.

In contrast, our approach does not introduce any auxiliary prediction modules and relies only on quantities already computed during standard training, resulting in negligible additional parameters and minimal computational overhead. This makes our method more computationally efficient and scalable compared to both Bayesian inference with repeated sampling and learning-loss-based active learning frameworks. Compared to our approach, which requires more delicate threshold adjustments, learning loss can appear more intuitive but carries additional risk. In light of these comparisons with various active learning baselines, our method demonstrates robust labeling performance and offers the advantage of progressively reduced computational cost by excluding high-confidence data in subsequent iterations.

VI. THREATS TO VALIDITY

In conducting active learning with *semi-informative sets* S_{semi} , we made several key decisions regarding experimental components. Since these choices can directly influence the experimental outcomes, we consider them as potential threats to validity and have thoroughly analyzed how their selection could impact the results.

A. INTERNAL VALIDITY

We utilized the ResNet architecture for training the labeled dataset D_L to perform image classification task. Although many other models have been used for image classification tasks, our experiments primarily focused on applying the proposed *semi-informative sets* S_{semi} to ResNet. The choice of model could potentially affect the performance of active learning, and thus, results may vary depending on the model used. Nevertheless, ResNet has been consistently employed in image classification research due to its relatively simple architecture, and since image-related models typically rely on convolution layers like ResNet, we expect that applying *semi-informative sets* S_{semi} to other models would also yield positive results.

Unlike conventional methods, we selected high-confidence data points using a threshold that was not determined linearly by hyperparameters but was made learnable based on the model's learning progress with the labeled dataset D_L .

This allowed us to fully utilize both accuracy and uncertainty information from the training data. However, rather than conducting a separate analysis to determine which of these two factors had a greater impact on active learning, we focused on leveraging both simultaneously. By adjusting the hyperparameters to assign more weight to the factor, accuracy or uncertainty information, that contributes more positively to active learning, we believe that the performance of the proposed approach using the *semi-informative sets* S_{semi} could be further maximized.

B. EXTERNAL VALIDITY

We selected image classification as the primary task for conducting active learning with baseline methods. This choice led to a dependency on specific models and datasets, which were incorporated into our experimental setup. To explore whether the proposed *semi-informative sets* S_{semi} approach for active learning could be applied more broadly to other tasks, future work could extend its application to different domains, such as natural language processing or recommendation systems, beyond computer vision. Since our method focuses on strategically selecting data for training without altering the model architecture, it can be flexibly applied to a wide range of tasks without being constrained by the nature of the task itself.

For our experiments, we selected short videos from movie trailers and the MNIST dataset. While various datasets could be considered for the image classification task, we chose these specific datasets for the following reasons. First, the video dataset extracted from movie trailers was expected to be more challenging than standard datasets, as it can include a wide variety of real-world objects within a single frame. Despite this complexity, applying our proposed *semi-informative sets* S_{semi} for active learning resulted in notable performance improvements even in the early stages of training. Additionally, we used the MNIST dataset to deliberately introduce data imbalance. The MNIST dataset was deemed suitable for this purpose because its class distinction is intuitive based on image shape, and it includes a relatively large number of classes, making it ideal for simulating imbalanced data. Even under conditions where certain classes had ten times fewer samples than others, we found that our *semi-informative sets* S_{semi} for active learning demonstrated robust performance across specific models.

VII. LIMITATIONS AND FUTURE WORK

We present the limitations of our proposed active learning approach using *semi-informative sets*, along with future directions for improvement. In our method, we utilized labeled data points for model training and then employed the trained model to classify the unlabeled dataset into *semi-informative sets* S_{semi} or S_{query} . The classification was based on the concept of confidence, distinguishing high from low values. However, we found that the experimental results can significantly vary depending on the criteria used to

determine this confidence threshold. Therefore, beyond the use of confidence, or uncertainty information, it is necessary to further explore representative methods for identifying *semi-informative sets*.

As mentioned in the earlier section on threats to validity, we simultaneously utilized train accuracy and uncertainty information while employing a learnable threshold. However, we regret not thoroughly identifying which of the two factors had a more substantial impact on active learning. We believe that further analysis, possibly through the involvement of additional hyperparameters, is necessary to determine which factor contributes more positively to learning. Additionally, exploring whether other types of information beyond these two could further enhance active learning using our *semi-informative sets* S_{semi} is another area worth investigating. Furthermore, extending the proposed framework to incorporate additional sources of information, such as feature-space structure or data distribution characteristics, could improve the robustness of semi-informative set identification. We also anticipate that adapting the learnable threshold mechanism to different model architectures and domains would broaden the applicability of our approach and provide a promising direction for future research.

In addition, we plan to extend our empirical evaluation to larger and more diverse benchmarks such as CIFAR-10 and ImageNet to analyze how dataset scale, class complexity, and distributional characteristics influence the effectiveness of our method. Such experiments will help clarify the generalizability of our semi-informative set strategy across datasets with varying structural properties and provide deeper insights into its robustness under different data regimes.

VIII. CONCLUSION

In this study, we addressed the longstanding issue of labeling costs in active learning by proposing an approach that prioritizes data requiring labeling through the use of *semi-informative sets*. By leveraging the predictive probability obtained from labeled data and employing a learnable threshold based on the information used during training, we were able to strategically select *semi-informative sets*. This allowed us to maximize learning efficiency within a given budget while aiming to improve performance on image classification tasks. As a result, we observed a significant improvement in performance even during the early stages of learning when using our *semi-informative sets* for active learning. Moreover, depending on the choice of active learning baseline methods, our approach demonstrated robust performance preservation, particularly in label-imbalanced datasets.

The proposed approach of using *semi-informative sets* for active learning does not require additional model training or external resources. Instead, it operates by assigning importance within the existing dataset, making it adaptable to any active learning method. Although we observed the advantages of our approach on label-imbalanced datasets, achieving a higher level of performance improvement beyond

the current results proved challenging [28], [29]. Therefore, we are considering further research on handling class imbalance as part of our future work. While our experiments demonstrated the benefits of this approach through several representative active learning methods, we anticipate that future research, involving more diverse active learning methods or related models, will continue to show that our *semi-informative sets* can effectively reduce labeling costs and enhance the efficiency of active learning in various scenarios.

ACKNOWLEDGMENT

(Kyungwook Min and Seunguk Yu are co-first authors.)

REFERENCES

- [1] J. N. Stember and H. Shalu, "Deep reinforcement learning with automated label extraction from clinical reports accurately classifies 3D MRI brain volumes," *J. Digit. Imag.*, vol. 35, no. 5, pp. 1143–1152, Oct. 2022.
- [2] E.-S.-A. El-Dahshan, T. Hosny, and A.-B.-M. Salem, "Hybrid intelligent techniques for MRI brain images classification," *Digit. Signal Process.*, vol. 20, no. 2, pp. 433–441, Mar. 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1051200409001377>
- [3] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The ApolloScape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1067–10676.
- [4] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8753527>
- [5] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, "PandaSet: Advanced sensor suite dataset for autonomous driving," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 3095–3101.
- [6] B. Settles. (2009). *Active Learning Literature Survey*. [Online]. Available: <https://minds.wisconsin.edu/handle/1793/60660>
- [7] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," 2023, *arXiv:2305.20050*.
- [8] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.
- [9] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *Proc. AAAI*, 2005, pp. 746–751. [Online]. Available: <https://people.cs.umass.edu/>
- [10] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.*, 2001, pp. 309–318. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-44816-0_31
- [11] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192. [Online]. Available: <https://proceedings.mlr.press/v70/gal17a/gal17a.pdf>
- [12] F. B. Smith, A. Kirsch, S. Farquhar, Y. Gal, A. S. Foster, and T. Rainforth, "Prediction-oriented Bayesian active learning," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2023, pp. 7331–7348. [Online]. Available: <https://proceedings.mlr.press/v206/bickfordsmith23a/bickfordsmith23a.pdf>
- [13] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 93–102.
- [14] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 79. [Online]. Available: <https://icml.cc/ims/conferences/2004/proceedings/papers/94.pdf>
- [15] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.
- [16] Y. Park, J. Park, D. J. Han, W. Choi, H. Kousar, and J. Moon, "Distribution aware active learning via Gaussian mixtures," in *Proc. ICLR*, 2023. [Online]. Available: <https://openreview.net/pdf?id=YXnggA4iiD>
- [17] A. Kirsch, J. V. Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/95323660ed2124450caa2c246b5ed90-Paper.pdf
- [18] C. Yin, B. Qian, S. Cao, X. Li, J. Wei, Q. Zheng, and I. Davidson, "Deep similarity-based batch mode active learning with exploration-exploitation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 575–584. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8215530>
- [19] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," 2022, *arXiv:2202.02794*.
- [20] A. Dekel, G. Hacohen, D. Weinshall, and O. Yehuda, "Active learning through a covering lens," in *Proc. Adv. Neural Inf. Process. Syst.*, 35, 2022, pp. 22354–22367.
- [21] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn.*, 2013, p. 896. [Online]. Available: https://scholar.google.com/scholar?hl=ko&as_sdt=0%2C5&q=Pseudo-label%3A+The+simple+and+efficient+semi+supervised+learning+method+for+deep+neural+networks&btnG=
- [22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1050–1059. [Online]. Available: <https://proceedings.mlr.press/v48/gal16.pdf>
- [23] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," 2021, *arXiv:2101.06329*.
- [24] D. Ha, M. Kim, and C. Y. Jeong, "Comparative performance analysis of loss functions for deep neural network on imbalanced datasets," "Comparative performance analysis of loss functions for deep neural network on imbalanced datasets," in *Proc. Symp. Korean Inst. Commun. Inf. Sci.*, 2022, pp. 900–901. [Online]. Available: https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE11047745&googleIPSandBox=false&mark=0&ipRange=false&b2cLoginYN=false&aiChatView=B&readTime=5-10&isPDFSizeAllowed=true&accessgl=Y&language=ko_KR&hasTopBanner=true
- [25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf
- [26] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/4563068>
- [27] M. Li and I. K. Sethi, "Confidence-based active learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1251–1261, Aug. 2006. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1642660>
- [28] J. Z. Bengar, J. van de Weijer, L. L. Fuentes, and B. Raducanu, "Class-balanced active learning for image classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3707–3716. [Online]. Available: https://openaccess.thecvf.com/content/WACV2022/papers/Bengar_Class-Balanced_Active_Learning_for_Image_Classification_WACV_2022_paper.pdf
- [29] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," 2020, *arXiv:2002.06815*.
- [30] M. Hussien, K. K. Nguyen, A. Ranjha, M. Krichen, A. Alshammari, and M. Cheriet, "Enabling efficient data integration of industry 5.0 nodes through highly accurate neural CSI feedback," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 813–824, Nov. 2023. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10180075>
- [31] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1622–1658, 3rd Quart., 2021. [Online]. Available: <https://ieeexplore.ieee.org/jelaam/9739/9520094/9415623-aam.pdf>
- [32] M.-S. Yang and K. P. Sinaga, "A feature-reduction multi-view k-means clustering algorithm," *IEEE Access*, vol. 7, pp. 114472–114486, 2019. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8793138>

- [33] Y. Jiang, F.-L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 688–701, Apr. 2015. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6862861>
- [34] R. A. Aljanabi, Z. T. Al-Qaysi, and M. S. Suzani, "Deep transfer learning model for EEG biometric decoding," *Appl. Data Sci. Anal.*, vol. 2024, pp. 4–16, Feb. 2024. [Online]. Available: <https://journals.mesopotamian.press/index.php/ADSA/article/view/391>
- [35] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: <https://www.cs.toronto.edu/>
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255. [Online]. Available: https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf



SEUNGUK YU (Graduate Student Member, IEEE) received the B.S. degree from the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, in 2023. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. His research interests include natural language processing, especially in Korean, data analytics, and machine learning.



YOONJI LEE (Graduate Student Member, IEEE) received the B.S. degree in applied statistics from Chung-Ang University, Seoul, South Korea, in 2024, where she is currently pursuing the M.S. degree with the Graduate School of Artificial Intelligence. Her research interests include deep learning, semi-supervised learning, and computer vision.



YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer with Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data mining and deep learning.



KYUNGWOOK MIN received the B.S. degree in statistic from Hankuk University of Foreign Studies, Yongin, South Korea, in 2020, and the M.S. degree from the Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea, in 2024. His research interests include deep learning and active learning.

...