International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-023-00429-8

SPECIAL ISSUE PAPER

Analyzing the potential of active learning for document image classification

Saifullah Saifullah[1,2] · **Stefan Agne**[1,3] · **Andreas Dengel**[1,2] · **Sheraz Ahmed**[1,3]

Received: 14 November 2022 / Accepted: 23 March 2023
© The Author(s) 2023

Abstract
Deep learning has been extensively researched in the field of document analysis and has shown excellent performance across a wide range of document-related tasks. As a result, a great deal of emphasis is now being placed on its practical deployment and integration into modern industrial document processing pipelines. It is well known, however, that deep learning models are data-hungry and often require huge volumes of annotated data in order to achieve competitive performances. And since data annotation is a costly and labor-intensive process, it remains one of the major hurdles to their practical deployment. This study investigates the possibility of using active learning to reduce the costs of data annotation in the context of document image classification, which is one of the core components of modern document processing pipelines. The results of this study demonstrate that by utilizing active learning (AL), deep document classification models can achieve competitive performances to the models trained on fully annotated datasets and, in some cases, even surpass them by annotating only 15–40% of the total training dataset. Furthermore, this study demonstrates that modern AL strategies significantly outperform random querying, and in many cases achieve comparable performance to the models trained on fully annotated datasets even in the presence of practical deployment issues such as data imbalance, and annotation noise, and thus, offer tremendous benefits in real-world deployment of deep document classification models. The code to reproduce our experiments is publicly available at https://github.com/saifullah3396/doc_al.

Keywords Document image classification · Document analysis · Active learning · Deep active learning

1 Introduction

Document analysis is a field of research that deals with automating the process of reading, analyzing, and understanding business documents. Modern businesses rely heavily on business documents to communicate details of their internal and external transactions, which is critical to their efficiency and productivity. As large volumes of documents are produced on a daily basis, there is an urgent need today to automate the processing of these documents to facilitate tasks such as search, retrieval, and information extraction. However, automated processing of documents can be particularly challenging for a number of reasons, including high levels of data complexity [1], large inter-class similarity and intra-class variance [2], and corruption of scanned document data with various types of distortions [3].

To address the aforementioned challenges, deep learning has been extensively explored in the field and has proven to be exceptionally effective in a wide range of document analysis tasks such as document image classification [4, 5], layout analysis [5], OCR [6], etc. However, deep learning presents some unique challenges of its own. One major disadvantage of deep learning-based approaches is that their performance is heavily dependent on the availability of large amounts of annotated training data. While most real-world tasks have a vast amount of data available that could

✉ Saifullah Saifullah
saifullah.saifullah@dfki.de

Stefan Agne
stefan.agne@dfki.de

Andreas Dengel
andreas.dengel@dfki.de

Sheraz Ahmed
sheraz.ahmed@dfki.de

[1] German Research Center for Artificial Intelligence, 67663 Kaiserslautern, Germany

[2] RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany

[3] DeepReader GmbH, 67663 Kaiserslautern, Germany

Published online: 25 April 2023

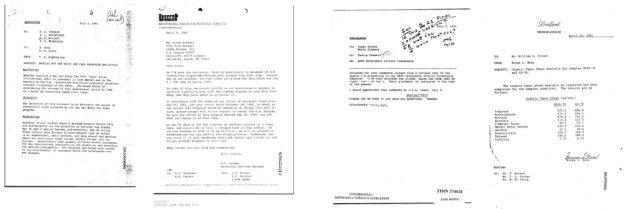


Fig. 1 Two samples of different classes with high similarity (top left) and two samples of the same class with high variance (top right) are shown

be annotated, which is also true for document processing tasks, data annotation is generally a labor-intensive task and can become extremely costly for certain tasks that require domain experts' knowledge to annotate. Furthermore, modern document processing pipelines are often continuously under development to accommodate new and evolving tasks and data requirements. Consequently, data annotation often becomes a routine process within these pipelines, resulting in an increased labor cost.

Active learning (AL) is a relatively new and emerging research topic that directly addresses the above mentioned challenges of data annotation costs [7]. The goal of active learning is to maximize the performance of deep learning models while minimizing the costs associated with annotation. Generally, AL involves training a machine learning model on a small labeled dataset and then using it to extract the most informative samples from an unlabeled pool of data samples. The newly extracted samples are then sent to the Oracle for labeling and incorporated back into the labeled dataset. Lastly, the model is retrained on the updated labeled dataset and the process is repeated. Active learning has recently made remarkable advancements [8], particularly in the fields of image classification [9–11] and semantic segmentation [12], where it has shown to significantly reduce annotation costs without adversely affecting model performance. However, despite the fact that AL can provide substantial benefits in terms of reducing annotation costs associated with document analysis tasks, only limited literature has been published in this direction. This paper investigates AL for reducing data annotation costs specifically in the context of document image classification, which is one of the core elements of modern document processing pipelines.

The task of document image classification poses a significant challenge due to the high intra-class variance and inter-class similarity [2, 13]. An example of this is shown in Fig. 1. Not only does this make it difficult for deep learning models to distinguish between different document classes, but also for humans to annotate them, which in turn increases the possibility of high annotation noise in document classification datasets. Therefore, this paper explores not only the effectiveness of existing AL approaches for reducing annotation costs while maintaining high performance, but also their effectiveness in countering labeling noise in document image classification. Data imbalance is another prevalent issue in real-world document datasets, and therefore, this paper additionally investigates the performance of AL under different scenarios of data bias and imbalance. In summary, this paper offers two main contributions:

1. This work shows for the first time that active learning can be used to significantly reduce data annotation costs while achieving competitive performance on document image classification benchmarks.
2. This work investigates the potential of different AL strategies in countering annotation noise and data imbalance and presents a thorough comparative analysis of their effectiveness in such scenarios.

2 Related work

2.1 Active learning

Active learning has seen tremendous research growth in the past few years, with a wide range of approaches proposed in this area [8]. Current active learning methods primarily fall into two categories: membership query-synthesis [14, 15] and pool-based approaches [11, 12, 16]. Query-synthesis active learning methods not only look for informative samples in the unlabeled dataset but also generate their own informative samples using generative models. In contrast, the pool-based methods rely mainly on different sampling techniques in order to query the most informative samples from an unlabeled dataset. This work focuses primarily on investigating pool-based active learning strategies for classifying document images, and therefore, previous work in this area is reviewed in greater depth.

Several approaches to pool-based active learning have been proposed in the past, which can generally be divided into three main categories: uncertainty-based approaches [16, 17], representation-based approaches [11], and enhanced hybrid approaches [12, 18]. Uncertainty-based approaches aim to identify and select those samples from the unlabeled dataset on which the trained model exhibits the greatest degree of uncertainty. These approaches have been proposed in both Bayesian and non-Bayesian frameworks. In non-Bayesian realm, various uncertainty measures are directly employed, such as entropy [19], distance from decision boundaries [17], and expected risk minimization. By contrast, Bayesian approaches estimate uncertainty using Gaussian processes. A study by Gal et al. [20] showed that neural networks with Dropout [21] applied before each weight layer approximate a probabilistic deep Gaussian pro-

cess, and used Dropout to estimate uncertainty in predictions. In a slightly different direction, some works have also proposed model ensembles to compute uncertainty [22].

Representation-based approaches focus primarily on querying samples that increase the diversity of the batch being queried. One popular representation-based method is KMeans Sampling [8], which generates the sample clusters from the unlabeled dataset using KMeans Clustering and then selects samples in proportion to their squared distances from the nearest centroid. CoreSets [11] is another popular representational-based approach that relies on reducing the distance between the queried samples and the labeled samples in feature space and has shown promising results in large-scale image classification applications.

Several enhanced or hybrid approaches have also been proposed in recent years. CEAL [18] is a hybrid active learning approach that may be used in conjunction with any existing active learning query method. CEAL first uses the underlying AL strategy to extract the samples from the unlabeled dataset and then extracts additional samples by assigning pseudo-labels to those samples that are confidently predicted by the model. Enhanced adversarial approaches such as DeepFool Active Learning (DFAL) [23] and Adversarial Basic Interactive Method (AdvBIM) [24] have also recently become popular, which seek adversarial examples in unlabeled datasets to increase the diversity of the samples being queried. Sinha et al. [12] proposed a hybrid adversarial approach that combines variational auto-encoders with an adversarial discriminator to increase batch diversity. Similarly, Shui et al. [25] recently proposed the Wasserstein Adversarial Active Learning (WAAL) approach which trains an independent discriminator model to search for diverse unlabeled samples. Loss Prediction Loss (LPL) [26] is another recent hybrid approach that trains a separate network in parallel with the target model to predict the loss of inputs with respect to the target model and then queries the samples that result in the highest predicted loss. In a different direction, Ash et al. [27] proposed the Batch Active Learning by Diverse Gradient Embeddings (BADGE) approach that finds a tradeoff between uncertainty and diversity by computing gradient embeddings for unlabeled samples and clustering them with the KMeans++ algorithm. Ash et al. [28] also recently proposed Batch Active Learning via Information MaTrices (BAIT) as an improvement over BADGE, which uses gradient embeddings in combination with Fisher information to determine the optimal tradeoff between uncertainty and diversity.

2.2 Document image classification

The topic of document image classification has been extensively researched in the past. Early work in this area concentrated primarily on exploiting the structural similarity in documents [29], feature matching [30], or applying classical approaches such as K-nearest neighbors [31] or hidden Markov models [32] to distinguish between classes of documents.

Recent advances in deep learning have led to the development of numerous image-based and multi-modal approaches for the classification of document images [4, 5, 13]. A major contribution to this field was made by Kang et al. [33], who demonstrated that even a shallow neural network with just four layers could achieve dramatic performance improvements over traditional approaches. In the following years, the works of Harley et al. [2] and Afzal et al. [13] explored the potential of convolutional neural networks (CNNs) in combination with transfer learning and demonstrated exceptional performance improvements on popular benchmark datasets. Several CNN-based approaches have been proposed since then that have explored different directions such as transfer learning [13], parallel training [4], multi-view stacking [34], and inherent interpretability [35] in the context of document image classification. In recent studies, self-supervised pretraining has also been investigated for document classification both in the image domain [36, 37] and in the multi-modal domain [1, 5, 38] in order to leverage large-scale document datasets for training without incurring additional annotation costs. In such approaches, however, data annotation is still required in order to fine-tune the models on the downstream tasks.

3 Methods

This section describes our active learning setup and the different query strategies that were investigated in this study.

3.1 Active learning setup

Let $\mathcal{D}_\mathrm{L} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ denote the labeled training dataset in a standard supervised learning setting, where x_i denotes a data sample and y_i denotes the corresponding class label, and let $\mathcal{D}_\mathrm{U} = \{x_1, x_2, \ldots, x_m\}$ denote a larger pool of unlabeled samples such that $n << m$, then the goal of active learning (AL) is to iteratively select b most informative samples from the unlabeled dataset ($x_\mathrm{U} \sim \mathcal{D}_\mathrm{U}$) using a query function f, such that when they are annotated and aggregated back the labeled dataset $\mathcal{D}_\mathrm{L}$, the overall classification performance of the machine learning model $\mathcal{M}$ trained on the updated labeled dataset $\mathcal{D}_\mathrm{L}$ is maximized.

In this study, the standard batch active learning (BAL) approach was used, which has previously been shown to be effective in training convolutional neural networks (CNNs) for image classification tasks [11]. In a standard supervised setting of BAL, each AL cycle (see Fig. 2) begins with training the deep learning model ($\mathcal{M}$) on the labeled dataset $\mathcal{D}_\mathrm{L}$.

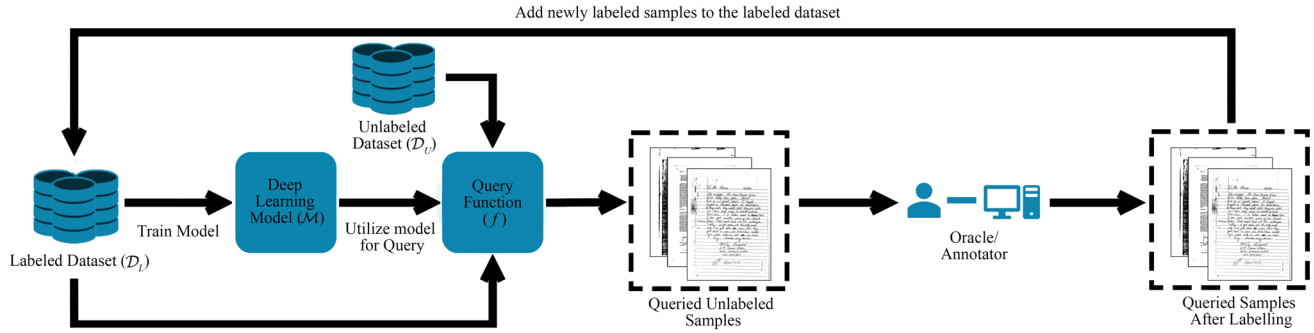


Fig. 2 An overview of the active learning cycle. The model $\mathcal{M}$ is first trained on the labeled dataset $\mathcal{D}_{\mathrm{L}}$ which is then utilized to query samples from the unlabeled dataset $\mathcal{D}_{\mathrm{U}}$. Samples are labeled by the oracle and aggregated back into the labeled set, and the process is repeated

The trained model $\mathcal{M}$ is then utilized in combination with a predefined query function f of choice to query a batch of samples of size b from the unlabeled dataset $\mathcal{D}_{\mathrm{U}}$. Newly selected samples are then sent to the Oracle for annotation and aggregated back into the labeled dataset $\mathcal{D}_{\mathrm{L}}$ for the next round of training. This cycle is repeated until either the total annotation budget $\mathcal{B}$ has been exhausted or a predefined termination condition has been met. In this work, a fixed number of active learning rounds was used as a termination condition.

The query function f is the most important component of AL, which describes the criteria by which samples are selected from the unlabeled dataset $\mathcal{D}_{\mathrm{U}}$. In order to maximize the machine learning model's performance with minimal annotation costs, it is necessary to select the most informative samples from $\mathcal{D}_{\mathrm{U}}$ during each AL round. A variety of query functions f have been proposed in the past, each defining the informativeness of a sample according to a different criterion. In this study, several existing pool-based query approaches were explored, including uncertainty-based approaches, representation-based approaches, and enhanced hybrid approaches.

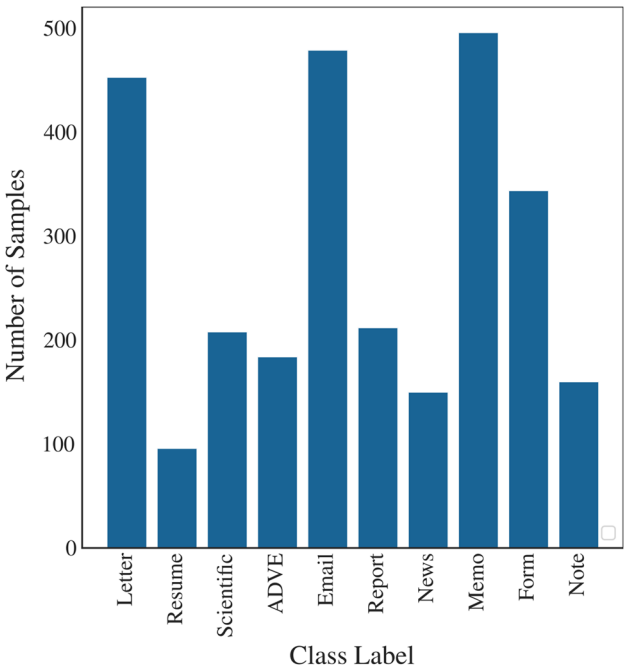


Fig. 3 The distribution of classes in the Tobacco3482 training set

3.1.1 Uncertainty-based approaches

Several uncertainty-based query functions have been investigated in this paper from both the Bayesian and Non-Bayesian realms. Non-Bayesian sampling techniques include Margin Sampling [8], Least Confidence Sampling [8], and Entropy Sampling [19]. In the Bayesian setting, the techniques Bayesian Active Learning Disagreement (BALD) [16], Margin Sampling, Least Confidence Sampling, and Entropy Sampling were explored in combination with Monte Carlo Dropout [20].

3.1.2 Representation-based approaches

In this domain, two approaches, namely CoreSets [11] and KMeans Sampling [8], were investigated. The CoreSets approach was implemented utilizing the KCenterGreedy algorithm, as originally proposed. However, due to the high dimensionality of the output embeddings of the model, it was not computationally feasible to apply the KCenterGreedy algorithm directly to the output of the model. To address this issue, principal component analysis (PCA) was used to reduce the dimensionality of the output embedding of the model before applying the querying algorithm.

3.1.3 Enhanced/Hybrid approaches

A number of enhanced adversarial [23, 24] and hybrid [12, 18] approaches were investigated in this work, including Cost-Effective Active Learning (CEAL) [18], the Adversarial Basic Interactive Method (AdvBIM) [24], WAAL [25], LPL [26], BADGE [27], and BAIT [28]. CEAL [18] was

used in combination with the Entropy uncertainty measure approach in this work.

4 Experiments and results

This section describes and presents the results of the experiments conducted in this paper.

4.1 Datasets

To evaluate the effectiveness and feasibility of different AL techniques, two publicly available document benchmark datasets were utilized, namely RVL-CDIP [2] and Tobacco3482, both of which have been extensively utilized for benchmarking document image classification in the past [2, 4, 13]. RVL-CDIP [2] is a large-scale dataset containing 400K labeled document images from 16 document classes, divided into training, testing, and validation sets of 320K, 40K, and 40K, respectively. Tobacco3482, in contrast, is a smaller dataset consisting of only 3482 images divided into 10 different document categories. It is important to note, however, that unlike RVL-CDIP, Tobacco3482 has an imbalanced class distribution as shown in Fig. 3. This makes it useful for investigating the performance of AL algorithms in the presence of class imbalance. In this work, the Tobacco3482 dataset was divided into training, test, and validation sets of 2504, 700, and 278 in size, respectively.

4.2 Implementation details

In order to simulate a realistic AL scenario, a small percentage of the original training set of the respective datasets was randomly sampled to create the initial labeled dataset $\mathcal{D}_{\mathrm{L}}$. This percentage was set at 10% for RVL-CDIP and at 5% for the Tobacco3482 dataset. The remaining samples in the training set were used to create the unlabeled pool $\mathcal{D}_{\mathrm{U}}$ from which the samples were extracted for annotation by the Oracle. In each AL cycle, the batch size for querying was set to 2.5% of the full training set, which is equivalent to 8000 for RVL-CDIP, and 63 for Tobacco3482 dataset and the AL cycle was repeated for each experiment until a total of 40% of the original dataset was annotated.

All the experiments were conducted using the standard ResNet-50 model [39] pretrained on the ImageNet-22k dataset [40], which has previously been demonstrated to perform exceptionally well on the aforementioned document datasets [13]. As has been done in previous works [4, 13], the input images for the model were down-scaled to a resolution of 224×224, converted to RGB color space, and normalized with the ImageNet mean of $(0.485, 0.456, 0.406)$ and standard deviation of $(0.229, 0.224, 0.225)$. Training was conducted using the standard Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, which was gradually reduced over the training cycle using the Cosine Decay Learning Rate Scheduler. In each AL cycle, 40 training epochs were used, a number previously determined to be sufficient for this task [4, 13]. For the RVL-CDIP dataset, a batch size of 256 was used, while for the Tobacco3482 dataset, a batch size of 64 was employed.

For the Tobacco3482 dataset, additionally, two different training settings, namely, $\text{Tobacco3482}_{ImageNet}$ and $\text{Tobacco3482}_{RVL-CDIP}$ were investigated. In $\text{Tobacco3482}_{ImageNet}$ setting, the models pretrained on ImageNet-22k were used. Whereas in $\text{Tobacco3482}_{RVL-CDIP}$ setting, the models pretrained on the RVL-CDIP dataset were utilized in order to assess the effectiveness of active learning in combination with document-specific pretraining. It is important to note that for pretraining the models on the RVL-CDIP dataset, we also used the ImageNet-22k pretrained weights for model initialization. In addition, results for all the experiments on Tobacco3482 dataset were presented with mean and standard deviation over 5 runs.

Some techniques were excluded from investigation for the larger dataset RVL-CDIP due to prohibitive computational requirements. With KMeans Sampling [8], the high CPU computational costs were faced, whereas, for BADGE [27], and BAIT [28], the memory requirements scaled proportionally with the size of the dataset, which rendered it impossible to apply these techniques to the RVL-CDIP dataset. Moreover, WAAL uses a two-stage training strategy for implementing discriminative learning and trains the networks on both the labeled and unlabeled datasets in parallel. This results in huge training costs when the unlabeled pool $\mathcal{D}_{\mathrm{U}}$ is large as compared to other active learning strategies. As a result, to apply WAAL on RVL-CDIP dataset, only a batch percentage of 5.0% was investigated so that the total computational costs of the active learning process could be reduced.

4.3 Performance evaluation

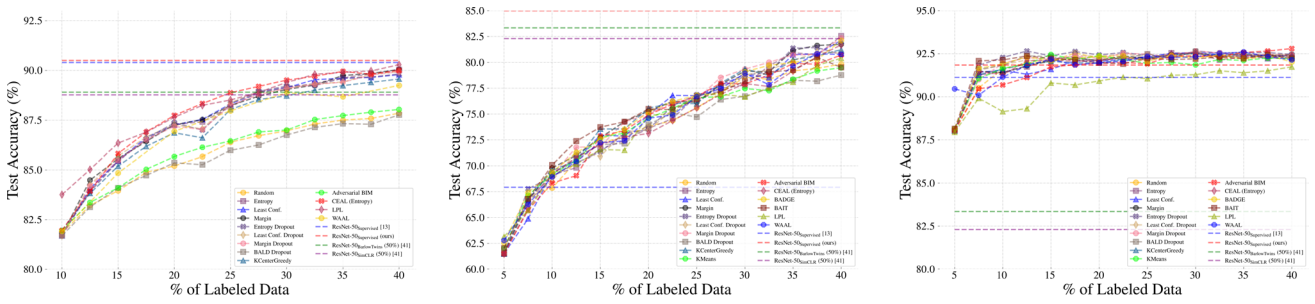
This section presents the performance results of different active learning algorithms on the two document datasets. For each AL method, Table 1 presents both the average accuracy achieved on 40% of original training datasets and the area under the budget curve (AUBC) which is useful in comparing the overall performance of an AL method under varying budgets. Figure 4 illustrates the budget–accuracy curves for each AL strategy, under different dataset settings, which indicate the accuracy achieved by the model after each AL round until a total of 40% of the training dataset was annotated.

For comparison of the active learning performance with standard supervised training on fully annotated dataset, the accuracy achieved by the model with 100% annotated dataset is also presented in the table as mentioned by ResNet-

Table 1 ResNet-50 performance under supervised learning with fully annotated training datasets (top) and active learning with 40% annotated training datasets (bottom)

	Model ($\mathcal{M}$)/Query strategy (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
		Acc	AUBC	Acc	AUBC	Acc	AUBC
Supervised learning (Ann. Data = 100%)	ResNet-50$_{Supervised}$ (Afzal et al. [13])	90.40	–	67.93	–	91.13	–
	ResNet-50$_{Supervised}$ (ours)	90.50	–	85.00	–	91.85	–
Self-supervised learning (Ann. Data = 50%)	ResNet-50$_{BarlowTwins}$ (Siddiqui et al. [36])	88.90	–	–	–	83.35	–
	ResNet-50$_{SimCLR}$ (Siddiqui et al. [36])	88.77	–	–	–	82.30	–
Active learning with ResNet-50 (Ann. Data = 40%)	Random Sampling	87.85	0.2575	79.83 ± 0.57	0.2593	92.34 ± 0.52	0.3218
Unc. non-Bayesian	Entropy [19]	90.04	0.2629	**82.54 ± 0.85**	0.2608	92.34 ± 0.31	0.3219
	Least Conf. [8]	89.79	0.2629	82.11 ± 0.80	0.2619	92.51 ± 0.13	0.3218
	Margin [8]	90.05	0.2630	81.74 ± 1.09	0.2631	92.43 ± 0.10	0.3220
Unc. Bayesian	Entropy (Dropout) [16]	89.82	0.2628	80.77 ± 0.45	0.2616	92.46 ± 0.21	**0.3230**
	Least Conf. (Dropout) [16]	89.93	0.2629	80.89 ± 1.07	0.2610	92.46 ± 0.12	0.3224
	Margin (Dropout) [16]	89.93	0.2627	82.34 ± 0.99	**0.2631**	92.40 ± 0.12	0.3223
	BALD (Dropout) [16]	87.76	0.2570	78.77 ± 0.50	0.2579	92.29 ± 0.45	0.3223
Repr	KCenterGreedy [11]	89.57	0.2618	81.23 ± 1.90	0.2625	92.37 ± 0.16	0.3219
	KMeans [8]	*	*	79.49 ± 1.17	0.2598	92.20 ± 0.24	0.3216
Enh./ Hybrid	AdvBIM [24]	88.03	0.2582	80.71 ± 1.41	0.2606	**92.80 ± 0.44**	0.3211
	CEAL (Entropy) [18]	90.00	0.2639	81.69 ± 0.61	0.2594	92.40 ± 0.19	0.3220
	BADGE [27]	*	*	82.11 ± 0.90	0.2627	92.03 ± 0.19	0.3222
	BAIT [28]	*	*	79.51 ± 1.37	0.2624	92.23 ± 0.16	0.3222
	LPL [26]	**90.29**	**0.2644**	80.40 ± 1.31	0.2589	91.74 ± 0.44	0.3174
	WAAL [25]	89.24	0.2614	80.77 ± 0.96	0.2607	92.14 ± 0.20	0.3217

For each task, the highest accuracy and AUBC are bolded
*Computationally unfeasible for large datasets



(a) RVL-CDIP (b) Tobacco3482$_{\text{ImageNet}}$ (c) Tobacco3482$_{\text{RVL-CDIP}}$

Fig. 4 Accuracy–budget curves for the different active learning strategies on RVL-CDIP and Tobacco3482 datasets

50$_{Supervised}$. The ResNet-50$_{Supervised}$ performance reported by Afzal et al. [13] for Tobacco3482 settings differs from ours in that they only used 100 randomly selected samples per class (a total of 1000 samples) for training, while we utilized the entire training dataset. The performance difference on both the Tobacco3482$_{ImageNet}$ and Tobacco3482$_{RVL-CDIP}$ settings is evident as a result of this difference in approach.

In addition to standard supervised learning approaches, we also compared the results of our experiments with self-supervised learning approaches. Siddiqui et al. [36] recently examined two state-of-the-art self-supervised approaches in the context of document classification, namely BarlowTwins [41] and SimCLR [42]. The research demonstrated how self-supervised pretraining can assist in reducing annotation costs on both large and small datasets. As they also used the ResNet-50 model for their analysis, our results can be directly compared with theirs. The results presented in this paper are based on an experiment in which Siddiqui et al. [36] first pretrained a ResNet-50 model on the RVL-CDIP dataset using the SimCLR and BarlowTwins self-supervised approaches and then fine-tuned the model on the RVL-CDIP and Tobacco3482 datasets with only 50% of the training data annotated. Model performance accuracies resulting from these experiments are shown in Table 1 denoted by ResNet-50$_{BarlowTwins}$ and ResNet-50$_{SimCLR}$.

4.3.1 Results on RVL-CDIP

As can be seen from Table 1, many of the AL techniques investigated in this work showed significantly better performance than the Random Sampling baseline. Furthermore, they were able to achieve a performance comparable to the model trained on fully annotated dataset (ResNet-50$_{Supervised}$) by using only 40% of the labeled training dataset. The enhanced LPL approach showed a significantly better performance compared to others both in terms of accuracy and AUBC which is also clearly visible in Fig. 4a. Uncertainty-based techniques such as Entropy and Margin also showed competitive performance despite their simplicity compared to enhanced approaches. One interesting observation is that while CEAL (Entropy) showed similar accuracy to Entropy, its AUBC was much higher in comparison. From Fig. 4a, it can be seen that CEAL (Entropy) results in an overall consistently higher accuracy over varying budgets as compared to Entropy. Some techniques such as BALD and AdvBIM performed even worse than the Random Sampling baseline in this case. KCenterGreedy also appeared to perform similarly to top uncertainty-based methods in terms of accuracy; however, its AUBC remained significantly lower. WAAL despite being a state-of-the-art (SotA) enhanced hybrid approach also performed poorly compared to other simpler approaches. These performance differences can also be observed in Fig. 4a, where the accuracy–budget curves of AdvBIM and BALD show a very similar behavior to the Random Sampling baseline, whereas the accuracy–budget curves of KCenterGreedy and WAAL stayed consistently lower than their competitors.

4.3.2 Results on Tobacco3482$_{ImageNet}$

A slightly different trend was seen in this setting, where Entropy showed the highest mean accuracy of 82.54%, which is ≈2.7% higher than the Random Sampling baseline. Other techniques that performed comparatively well were Margin (Dropout) and BADGE. One noticeable observation in this scenario is that many of the techniques showed really high variance due to the small dataset size and large class imbalance. The variation could also explains why the Margin (Dropout) approach had the highest overall AUBC, regardless of its lower accuracy compared to Entropy. BALD, KMeans, and BAIT performed even worse than the Random Sampling baseline in this scenario which is surprising as BAIT has been shown to perform much better than other approaches in natural image classification domain [28]. One interesting observation from Fig. 4b is that BAIT performed better than other approaches in the first few rounds (up to 15% labeled dataset); however, its performance degraded suddenly afterward. The SotA enhanced approaches LPL and

WAAL also performed poorly compared to other approaches in this scenario. A possible explanation is that these techniques are greatly sensitive to data imbalance. For example, for LPL, data imbalance may result in bias in loss prediction, resulting in poor performance as a consequence. Interestingly, these observations are similar to those reported by [8]. Overall, it can be noted that AL strategies did not perform as well in this scenario as they did on RVL-CDIP. However, these results are not surprising, since AL methods have previously been shown to suffer from biased sampling in case of data imbalance [8, 43].

4.3.3 Results on Tobacco3482$_{RVL-CDIP}$

It is evident from Table 1 that the document-specific pretraining resulted in a significant improvement in the performance of AL algorithms. It is interesting to note that even the Random Sampling baseline outperformed the ResNet-50$_{Supervised}$ model trained on the full dataset, proving that active learning can be an effective training process even with just Random querying in some scenarios. Uncertainty-based techniques performed overall better than other approaches (except AdvBIM) in this setting; however, the differences between their performance was minor. Figure 4c also illustrates an interesting observation that the majority of AL strategies under this scenario outperformed the ResNet-50$_{Supervised}$ model even with just 10–15% of the training data annotated. Another noteworthy observation from Fig. 4c is that AdvBIM started out the worst in initial rounds but performed significantly better than other approaches achieving a mean accuracy of 92.80% at 40% labeled training dataset. Similar to the Tobacco3482$_{ImageNet}$ case, many enhanced approaches including BADGE, BAIT, LPL and WAAL performed poorly in this scenario, sometimes performing even worse than the Random Sampling baseline.

4.3.4 Comparison with self-supervised approaches

From Table 1, it can be observed that the AL approaches consistently outperformed the self-supervised approaches ResNet-50$_{BarlowTwins}$ and ResNet-50$_{SimCLR}$ despite overall 10% fewer annotated samples. In addition, it can be seen from Fig. 4a that the self-supervised performance on RVL-CDIP is outperformed by a number of AL strategies with only 25–30% of the training data labeled. It should also be noted that while AL strategies showed only modest improvements over the self-supervised approaches on the RVL-CDIP dataset, they demonstrated much more significant improvements on Tobacco3482. It can be seen from Fig. 4c that, when using standard RVL-CDIP pretraining, fine-tuning on the Tobacco3482 dataset, even with 5% annotated data, was more effective than fine-tuning from self-supervised pretraining with 50% annotated data. This suggests that task-specific (classification) pretraining on a larger dataset (RVL-CDIP) can provide significant performance boosts on smaller datasets in comparison with self-supervised pretraining. This does, however, mean that the larger pretraining dataset must be fully annotated itself, which may not always be possible, in which case, self-supervised pretraining may be a more viable option. While the above is true, it can also be noticed from Fig 4b that even without RVL-CDIP pretraining for the Tobacco3482 dataset, AL strategies with only 40% of the training dataset were able to achieve performances comparable to those achieved by the self-supervised pretraining approaches. This indicates that active learning can be as effective in reducing annotation costs as self-supervised pretraining even if there is no document-specific training.

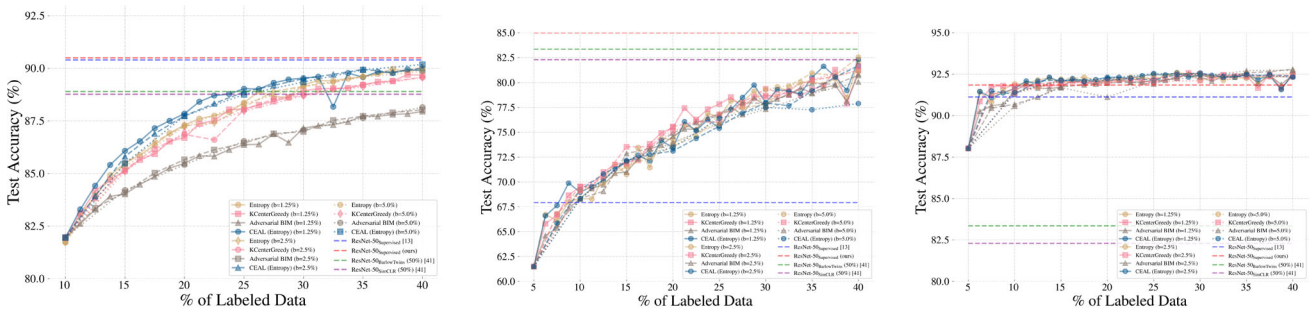
4.4 Query time analysis

The time it takes to perform a querying operation is an imperative consideration when selecting an active learning strategy. Since Random Sampling is also an effective AL method, ideally the querying methods should be as time-efficient as that. Table 2 provides the mean querying time (Avg. t_{query}) spent by each AL strategy and the total training time (Rel. t_{train}) taken by each strategy per AL round relative to the Random Sampling baseline. As can be seen from the table, non-Bayesian uncertainty-based approaches were the fastest in terms of both Avg. t_{query} and Rel. t_{train}. The Bayesian uncertainty-based methods, on the other hand, were significantly slower which is expected since these methods use multiple forward passes with Dropout to compute uncertainty. CEAL (Entropy) also showed competitive computational times since it used Entropy as the underlying querying method. However, its training time was comparatively much higher as it adds additional training samples to the labeled dataset $\mathcal{D}_L$ per round. In order to overcome the large memory requirements of the KCenterGreedy algorithm for large datasets, batch processing was used in our work. As a consequence, its querying time increased proportionally with the size of the dataset. This is evident from Table 2, where its querying time is similar to uncertainty-based techniques on the small Tobacco3482 dataset but increases considerably on RVL-CDIP. Although BADGE was found to perform as well as non-Bayesian uncertainty-based approaches, BAIT showed nearly ten times more computation time than BADGE. LPL despite training additional models showed competitive training and querying time. The training time for WAAL, on the other hand, scaled considerably with the size of the dataset, taking about 18× more time for training the model on the RVL-CDIP dataset compared to Random Sampling baseline. This is the reason why WAAL was only investigated with a batch percentage of 5.0% on RVL-CDIP dataset.

Table 2 For each AL strategy, the average time spent on querying samples from the unlabeled pool $\mathcal{D}_U$ and the total training time relative to the Random Sampling baseline are shown

	Strategy (f)	RVL-CDIP		Tobacco3482	
		Avg. t_{query} (s)	Rel. t_{train}	Avg. t_{query} (s)	Rel. t_{train}
	Random Sampling	0.0	1.0	0.0	1.0
Unc. non-Bayesian	Entropy [19]	539	1.03	6	0.98
	Least Conf. [8]	532	1.01	6	1.00
	Margin [8]	430	1.02	6	0.92
Unc. Bayesian	Entropy (Dropout) [16]	5592	1.39	61	1.27
	Least Conf. (Dropout) [16]	5535	1.36	63	1.31
	Margin (Dropout) [16]	5553	1.36	65	1.39
	BALD (Dropout) [16]	5527	1.35	61	1.25
Repr.	KMeans [8]	*	*	9	1.06
	KCenterGreedy [11]	1713	1.81	8	1.11
Enh./Hybrid	AdvBIM [24]	4302	1.28	41	1.14
	CEAL (Entropy) [18]	740	1.82	11	0.90
	BADGE [27]	*	*	15	1.08
	BAIT [28]	*	*	145	1.92
	LPL [26]	614	1.23	6	2.22
	WAAL [25]	362	17.99	9	3.26

*Computationally unfeasible for large datasets



(a) RVL-CDIP (b) Tobacco3482$_{\text{ImageNet}}$ (c) Tobacco3482$_{\text{RVL-CDIP}}$

Fig. 5 Accuracy–budget curves for the different active learning strategies on RVL-CDIP and Tobacco3482 datasets under varying query batch sizes

4.5 Effect of varying batch size

One critical hyperparameter for active learning training scenarios is the batch size. Previous studies have shown that smaller batches result in better AL performance since fewer redundant samples are queried [12]. We examined this effect for the document domain using batch percentages of $\%b = 1.25\%$, $\%b = 2.5\%$, and $\%b = 5.0\%$ on a subset of techniques as presented in Table 3. Figure 5 also illustrates the accuracy–budget curves for four different techniques under varying batch sizes. Only four techniques are shown here for visual clarity. Although there were minor differences in accuracy and AUBC across different batch sizes, no major differences were observed that could be directly correlated with increasing or decreasing batch sizes for RVL-CDIP. For Tobacco3482 settings, some minor differences were seen for the AUBC across difference batches which can also be observed in Fig. 5. For example, the effect of different batch sizes on CEAL (Entropy) and AdvBIM approaches is clearly visible with $b = 5.0\%$ resulting in worse performance in comparison. On the other hand, increasing the batch size also resulted in decreased variance across different runs for the Tobacco3482$_{ImageNet}$ case as is evident from the mean standard deviations across different batch sizes. To confirm our findings, we also conducted a single-factor ANOVA test [44] to compare the AUBCs of all methods across the three batch size groups. For RVL-CDIP, no statistically significant difference was found. (Alpha was greater than 0.05.) However, a significant difference was found (alpha was greater than 0.05) for the Tobacco3482 settings in which it was observed that the AUBC of the AL strategies was generally lower for $b = 5.0\%$ than for other batch sizes. This is also evident

Table 3 ResNet-50 performance under supervised learning with fully annotated training datasets (top) and active learning with 40% annotated training datasets and with varying batch sizes (bottom)

	Model/query function (f)		RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
			Acc	AUBC	Acc	AUBC	Acc	AUBC
Supervised learning (Ann. Data = 100%)	ResNet-50$_{Supervised}$ (Afzal et al. [13])		90.40	–	67.93	–	91.13	–
	ResNet-50$_{Supervised}$ (ours)		90.50	–	85.00	–	91.85	–
Self-supervised learning (Ann. Data = 50%)	ResNet-50$_{BarlowTwins}$ (Siddiqui et al. [36])		88.90	–	–	–	83.35	–
	ResNet-50$_{SimCLR}$ (Siddiqui et al. [36])		88.77	–	–	–	82.30	–
Active learning with ResNet-50 (Ann. Data = 40%)								
Batch percentage ($\%b = 1.25\%$)	Random Sampling		87.82	0.2574	80.00 ± 0.52	0.2581	91.83 ± 0.19	0.3208
	Entropy [19]		90.00	0.2630	82.06 ± 1.28	0.2613	92.46 ± 0.31	0.3221
	Entropy (Dropout) [16]		89.92	0.2631	81.43 ± 1.16	0.2608	92.29 ± 0.23	0.3225
	KCenterGreedy [11]		89.68	0.2622	81.23 ± 1.40	0.2624	92.40 ± 0.12	0.3219
	KMeans [8]		*	*	78.51 ± 0.77	0.2578	92.23 ± 0.22	0.3221
	AdvBIM [24]		87.95	0.2579	80.86 ± 0.85	0.2599	92.46 ± 0.23	0.3215
	CEAL (Entropy) [18]		89.88	0.2641	82.31 ± 0.94	0.2620	92.54 ± 0.23	0.3223
	BADGE [27]		*	*	81.34 ± 1.68	0.2623	92.26 ± 0.23	0.3222
	BAIT [28]		*	*	79.86 ± 0.74	0.2615	91.91 ± 0.24	0.3219
	LPL [26]		90.20	0.2645	79.94 ± 1.67	0.2603	91.40 ± 0.98	0.3186
	WAAL [25]		–	–	80.40 ± 0.72	0.2610	91.94 ± 0.34	0.3215
		Mean=	89.35	0.2618	80.72 ± 1.07	0.2607	92.16 ± 0.30	0.3216
Batch percentage ($\%b = 2.5\%$)	Random Sampling		87.85	0.2575	79.83 ± 0.57	0.2593	92.34 ± 0.52	0.3218
	Entropy [19]		90.04	0.2629	82.54 ± 0.85	0.2608	92.34 ± 0.31	0.3219
	Entropy (Dropout) [16]		89.82	0.2628	80.77 ± 0.45	0.2616	92.46 ± 0.21	0.3230
	KCenterGreedy [11]		89.57	0.2618	81.23 ± 1.90	0.2625	92.37 ± 0.16	0.3219
	KMeans [8]		*	*	79.49 ± 1.17	0.2598	92.20 ± 0.24	0.3216
	AdvBIM [24]		88.03	0.2582	80.71 ± 1.41	0.2606	92.80 ± 0.44	0.3211
	CEAL (Entropy) [18]		90.00	0.2639	81.69 ± 0.61	0.2594	92.40 ± 0.19	0.3220
	BADGE [27]		*	*	82.11 ± 0.90	0.2627	92.03 ± 0.19	0.3222
	BAIT [28]		*	*	79.51 ± 1.37	0.2624	92.23 ± 0.16	0.3222
	LPL [26]		90.29	0.2644	80.40 ± 1.31	0.2589	91.74 ± 0.44	0.3174
	WAAL [25]		–	–	80.77 ± 0.96	0.2607	92.14 ± 0.20	0.3217
		Mean=	89.37	0.2616	80.82 ± 1.05	0.2608	92.28 ± 0.28	0.3215

Table 3 continued

	Model/query function (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
		Acc	AUBC	Acc	AUBC	Acc	AUBC
Batch percentage($\%b = 5\%$)	Random Sampling	87.64	0.2570	79.80 ± 0.66	0.2584	92.51 ± 0.19	0.3208
	Entropy [19]	89.85	0.2627	81.11 ± 0.68	0.2606	92.57 ± 0.17	0.3216
	Entropy (Dropout) [16]	89.91	0.2626	81.97 ± 0.40	0.2585	92.40 ± 0.31	0.3217
	KCenterGreedy [11]	89.57	0.2618	81.63 ± 0.70	0.2616	92.49 ± 0.30	0.3211
	KMeans [8]	*	*	79.51 ± 0.90	0.2595	92.03 ± 0.44	0.3209
	AdvBIM [24]	88.15	0.2580	80.06 ± 0.33	0.2586	92.74 ± 0.19	0.3204
	CEAL (Entropy) [18]	90.18	0.2636	77.89 ± 1.81	0.2574	92.31 ± 0.19	0.3217
	BADGE [27]	*	*	80.46 ± 0.92	0.2605	92.51 ± 0.16	0.3222
	BAIT [28]	*	*	80.11 ± 0.42	0.2621	91.86 ± 0.23	0.3217
	LPL [26]	90.22	0.2643	80.46 ± 1.76	0.2584	92.06 ± 0.41	0.3169
	WAAL [25]	–	–	80.51 ± 0.85	0.2608	92.06 ± 0.30	0.3209
	Mean=	89.36	0.2614	80.32 ± 0.86	0.2597	92.32 ± 0.26	0.3209

*Computationally unfeasible for large datasets

from the mean AUBC values across different batch sizes in Table 3.

4.6 Bias in initial labeled dataset

4.6.1 Experimental setup

This section describes a set of experiments that examined the effects of bias in the initial labeled dataset $\mathcal{D}_{L0}$ on the overall performance of the AL query strategies. It is often the case that data bias occurs in real-world deployment scenarios when there is a relative shortage of labeled data for one class in comparison with another. Due to this bias, the model trained in the first round may not accurately represent the underlying distribution of data. We simulate the bias by excluding samples from the initial labeled datasets of m randomly selected classes. This experiment was carried out for two cases, $m = 2$ and $m = 4$, and the results are presented in Table 4. The accuracy–budget curves for this scenario are illustrated in Fig. 6.

4.6.2 Results on RVL-CDIP

In the case of the RVL-CDIP dataset with 16 classes, the effect of bias was negligible since most techniques exhibited a similar performance trend as in the case of no bias reported previously in Table 1. As can be seen in Fig. 6a, d, the models initially performed poorly in comparison due to a lack of data for some classes; however, their performance quickly improved in subsequent cycles. Overall, no significant performance drop was observed when m was increased from 2 to 4 as well with LPL, CEAL (Entropy), and uncertainty-based approaches again comparatively performing the best in this scenario.

4.6.3 Results on Tobacco3482$_{ImageNet}$

In Tobacco3482$_{ImageNet}$ setting, AL strategies seemed especially effective in countering the data bias compared to Random Sampling baseline on both $m = 2$ and $m = 4$ cases as they understandably targeted the missing classes in the subsequent AL cycles. Margin and its Dropout variant showed the highest accuracy in this scenario closely followed by KCenterGreedy and Least Confidence. Surprisingly, many strategies showed a higher accuracy with $m = 4$ compared to $m = 2$. This may have been due to the removal of high frequency classes in the initial dataset allowing the model to be less susceptible to overall class imbalance. The classes that were randomly selected for removal for $m = 2$ and $m = 4$ cases for this dataset were {Memo, Letter} and {Memo, Letter, Form, ADVE}. As can be noticed from Fig. 3, for the $m = 4$ case, Email was the only high frequency class left after the sample removal from the initial labeled

Table 4 ResNet-50 performance under supervised learning with fully annotated training datasets (top) and active learning with 40% annotated training datasets and data bias in the initial labeled dataset $\mathcal{D}_{L0}$ (bottom)

	Model/query function (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
		Acc	AUBC	Acc	AUBC	Acc	AUBC
Supervised learning (Ann. Data = 100%)	ResNet-50$_{Supervised}$ (Afzal et al. [13])	90.40	–	67.93	–	91.13	–
	ResNet-50$_{Supervised}$ (ours)	90.50	–	85.00	–	91.85	–
Self-supervised learning (Ann. Data = 50%)	ResNet-50$_{BarlowTwins}$ (Siddiqui et al. [36])	88.90	–	–	–	83.35	–
	ResNet-50$_{SimCLR}$ (Siddiqui et al. [36])	88.77	–	–	–	82.30	–
Active learning with ResNet-50 (Ann. Data = 40%)							
Classes removed ($m = 2$)	Random Sampling	87.72	0.2559	78.80 ± 1.53	0.2506	92.31 ± 0.34	0.3186
	Entropy [19]	**89.98**	0.2608	81.43 ± 1.37	0.2523	92.57 ± 0.23	0.3182
	Least Conf. [8]	89.86	0.2609	81.74 ± 1.07	0.2544	92.54 ± 0.12	0.3184
	Margin [8]	89.91	0.2613	81.51 ± 0.80	**0.2577**	92.46 ± 0.26	0.3186
	Entropy (Dropout) [16]	89.89	0.2606	80.69 ± 1.07	0.2534	92.54 ± 0.12	0.3193
	Least Conf. (Dropout) [16]	89.77	0.2608	81.66 ± 0.88	0.2572	92.46 ± 0.29	0.3184
	Margin (Dropout) [16]	89.78	0.2610	**82.09 ± 1.30**	0.2566	92.40 ± 0.19	0.3189
	BALD (Dropout)[16]	87.92	0.2557	78.60 ± 1.83	0.2502	**92.74 ± 0.53**	**0.3194**
	KCenterGreedy [11]	89.63	0.2593	81.94 ± 1.47	0.2529	92.37 ± 0.33	0.3182
	KMeans [8]	*	*	78.17 ± 0.55	0.2509	92.26 ± 0.12	0.3183
	AdvBIM [24]	88.06	0.2564	80.14 ± 1.42	0.2521	92.43 ± 0.27	0.3174
	CEAL (Entropy) [18]	89.91	0.2613	81.03 ± 1.24	0.2515	92.37 ± 0.33	0.3182
	BADGE [27]	*	*	81.49 ± 1.35	0.2556	92.26 ± 0.16	0.3186
	BAIT [28]	*	*	79.57 ± 0.90	0.2546	92.09 ± 0.22	0.3183
	LPL [26]	89.94	**0.2616**	79.46 ± 1.89	0.2480	91.63 ± 0.51	0.3053
	WAAL [25]	89.03	0.2575	79.60 ± 1.39	0.2549	92.03 ± 0.51	0.3172
Classes removed ($m = 4$)	Random Sampling	87.75	0.2534	79.60 ± 1.43	0.2465	92.54 ± 0.21	0.3154
	Entropy [19]	89.88	0.2587	81.09 ± 1.01	0.2448	92.49 ± 0.24	0.3125
	Least Conf. [8]	89.83	0.2591	81.57 ± 0.97	0.2478	92.54 ± 0.16	0.3130
	Margin [8]	89.92	0.2593	**82.89 ± 0.51**	0.2498	**92.60 ± 0.12**	0.3152
	Entropy (Dropout) [16]	89.91	0.2586	81.03 ± 1.23	0.2463	92.49 ± 0.22	0.3156
	Least Conf. (Dropout) [16]	89.84	**0.2594**	80.60 ± 1.38	0.2472	92.43 ± 0.30	0.3155
	Margin (Dropout) [16]	89.74	0.2592	80.94 ± 0.93	0.2507	92.37 ± 0.28	0.3158

Table 4 continued

Model/query function (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
	Acc	AUBC	Acc	AUBC	Acc	AUBC
BALD (Dropout) [16]	87.86	0.2533	78.54 ± 1.45	0.2472	92.31 ± 0.33	0.3160
KCenterGreedy [11]	89.47	0.2571	81.97 ± 1.40	0.2479	92.51 ± 0.30	0.3160
KMeans [8]	*	*	78.43 ± 0.84	0.2500	92.43 ± 0.30	**0.3162**
AdvBIM [24]	88.19	0.2547	80.31 ± 0.82	0.2469	92.46 ± 0.37	0.3079
CEAL (Entropy) [18]	89.77	0.2593	80.69 ± 0.98	0.2442	92.40 ± 0.37	0.3125
BADGE [27]	*	*	81.03 ± 0.36	**0.2524**	92.14 ± 0.27	0.3157
BAIT [28]	*	*	79.94 ± 0.39	0.2488	92.40 ± 0.33	0.3151
LPL [26]	**90.04**	0.2582	77.43 ± 2.96	0.2381	91.31 ± 0.90	0.3007
WAAL [25]	89.28	0.2547	80.83 ± 0.84	0.2481	92.14 ± 0.34	0.3144

For each task, the highest accuracy and AUBC are bolded
*Computationally unfeasible for large datasets

dataset. Another interesting observation can be made from Fig. 6e where it is visible that KMeans Sampling, BAIT, and BADGE both performed better than other techniques in the few initial AL cycles; however, their performance degraded after approximately 20% data was annotated. This behavior was also previously seen on the Tobacco3482$_{ImageNet}$ setting described in Sect. 4.3. A possible explanation for this behavior is that BAIT and BADGE are highly vulnerable to class imbalance in the data. WAAL and LPL showed a trend similar to the case of no bias discussed in Sect. 4.3 performing even worse than the Random Sampling baseline in some cases.

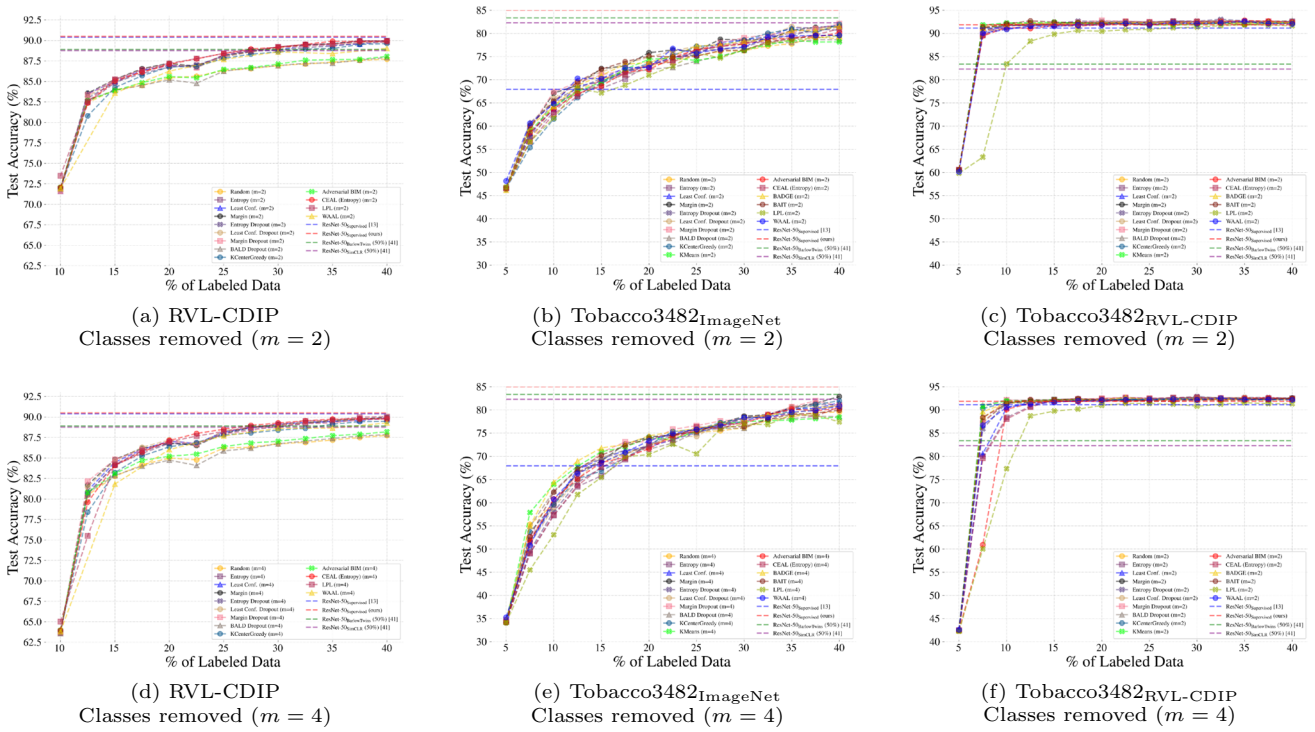
4.6.4 Results on Tobacco3482$_{RVL-CDIP}$

For the Tobacco3482$_{RVL-CDIP}$ scenario, the difference in accuracy among different AL strategies including Random Sampling baseline at 40% annotated data was quite negligible and all of them surpassed the performance of the ResNet-50$_{Supervised}$ model with just 15%-20% data annotated. Uncertainty-based approaches, however, scored comparatively higher at 40% annotated data. From Fig. 6c, f, it can also be observed that the enhanced approaches BADGE and BAIT and the diversity-based techniques such as KCenterGreedy and KMeans were able to handle the bias much better than others achieving a higher performance in first few cycles in comparison with others. On the other hand, enhanced approaches AdvBIM and LPL seemed especially vulnerable to the bias, taking a number of iterations to reach the same level of performance as other methods.

4.6.5 Removing highest and lowest frequency classes

In this section, we present the results of another experiment in which, rather than removing the classes at random, we removed the top 4 highest frequency classes and the top 4 lowest frequency classes to determine its overall effect on the different AL strategies. This experiment was only performed for the Tobacco3482$_{ImageNet}$ and Tobacco3482$_{RVL-CDIP}$ settings as only those settings have class imbalance. The top 4 highest frequency classes that were removed from the dataset include Letter, Email, Memo, and Form. In contrast, the top 4 lowest frequency classes that were removed include Resume, News, Note, and ADVE.

In this experiment, the results are presented only as accuracy–budget curves, as shown in Fig. 7. A few interesting conclusions can be drawn from the figure. To begin with, it can be observed that only removing the highest and lowest frequency classes had a significant effect on the performance of representation-based AL strategies such as KMeans, BADGE, and BAIT. This effect was especially pronounced in the Tobacco3482$_{ImageNet}$ case. In both cases of the Tobacco3482$_{ImageNet}$ scenario, the removal of the high-



(a) RVL-CDIP
Classes removed ($m = 2$)

(b) Tobacco3482$_{\text{ImageNet}}$
Classes removed ($m = 2$)

(c) Tobacco3482$_{\text{RVL-CDIP}}$
Classes removed ($m = 2$)

(d) RVL-CDIP
Classes removed ($m = 4$)

(e) Tobacco3482$_{\text{ImageNet}}$
Classes removed ($m = 4$)

(f) Tobacco3482$_{\text{RVL-CDIP}}$
Classes removed ($m = 4$)

Fig. 6 Accuracy–budget curves for the different active learning strategies on RVL-CDIP and Tobacco3482 datasets under biased initial labeled dataset

est frequency classes (Fig. 7a) and the removal of the lowest frequency classes (Fig. 7c), these techniques performed even worse than the Random Sampling baseline. Surprisingly, KCenterGreedy approach while it is also a representation-based approach stayed unaffected in this experiment. Overall, we observed no significant effect on performance of other approaches in this scenario, with the exception of CEAL (Entropy), which showed some instability in last few AL rounds when the top four classes with the lowest frequency were removed.

4.7 Robustness to annotation noise

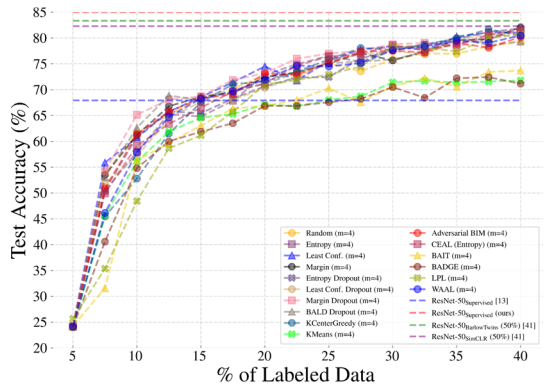
4.7.1 Experimental setup

The problem of annotation noise is very common in real-world deployment scenarios of machine learning models. Past studies have shown that even with human annotators, the amount of mislabeled samples in the training dataset can reach up to 10% of its size [45]. This section describes a set of experiments in which a realistic annotation noise scenario was created for document data by randomly switching labels of those document classes that exhibit similarity with each other based on predetermined weights $W \in \mathbb{R}^{N \times N}$.

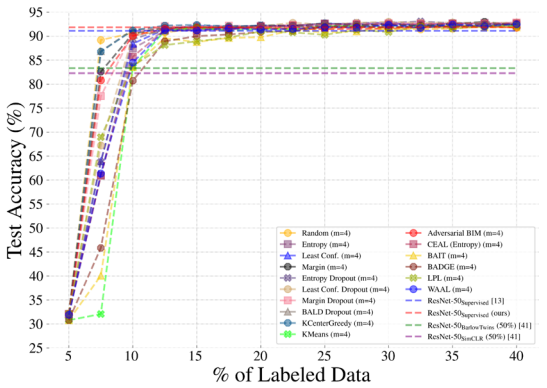
This process is detailed in Algorithm 1. As shown, for each class l, we determine the probabilities of drawing labels for similar classes based on the predetermined weights $W \in \mathbb{R}^{N \times N}$. Then, based on the class probabilities $p \in \mathbb{R}^{1 \times N}$, we draw n random labels and assign them to the existing samples of the class l, where n is the total number of samples to be updated. This process is repeated for each class and the training set is updated. Note that we used an in-place update to the dataset in this scenario, so the noise added to the dataset for classes that were mutually similar could be spread over both classes. For example, if two classes Letter and Memo were similar to each other, Algorithm 1 resulted in switching a total of 10% of the samples between them. Additionally, it allowed switching samples between classes that were only indirectly similar. For example, if Letter was similar to Memo and Memo was similar to Presentation, then some of the samples from Letter class were also converted to Presentation class.

Algorithm 1 Noise Labeling Strategy

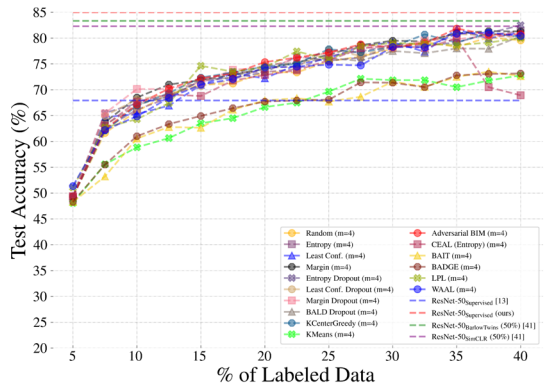
1: **Input**: Training set with image-label pairs $S = \{(x, y)\}_{k=1}^{K}$, $y \in \{1, \ldots, N\}$, class similarity weight matrix $W \in \mathbb{R}^{N \times N}$, and annotation noise percentage ϵ
2: **for** $l = 1$ to N **do**
3: Let $p = \frac{W_l}{\sum_{j \in N} W_{lj}} \in \mathbb{R}^{1 \times N}$ and p_i be the probability to draw the ith class label
4: Randomly draw $n = \epsilon \mid S_{y=l} \mid$ labels $L = \{i : i \in 1, \ldots, N\}_{k=1}^{n}$ with probability of each label defined by p_i and assign them to the sample set $S_{y=l}$
5: **end for**



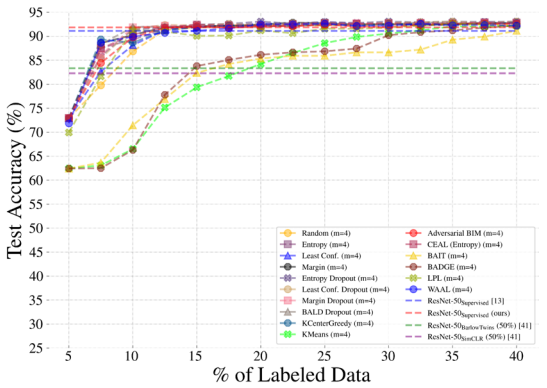
(a) $\text{Tobacco3482}_{\text{ImageNet}}$
Top 4 highest frequency classes removed



(b) $\text{Tobacco3482}_{\text{RVL-CDIP}}$
Top 4 highest frequency classes removed

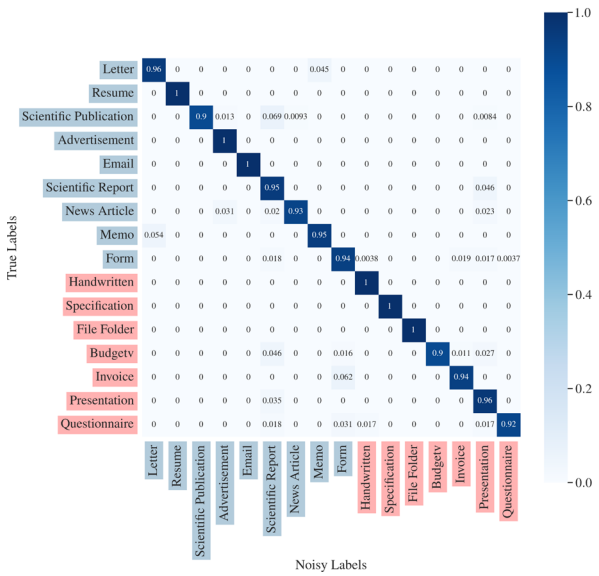


(c) $\text{Tobacco3482}_{\text{ImageNet}}$
Top 4 lowest frequency classes removed

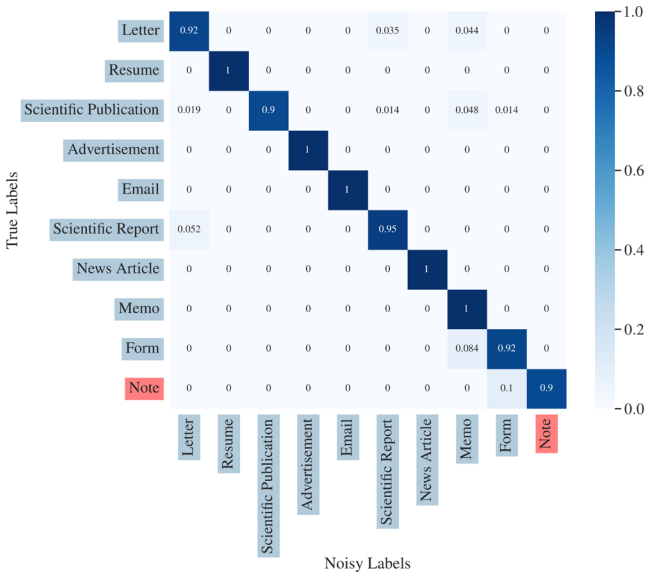


(d) $\text{Tobacco3482}_{\text{RVL-CDIP}}$
Top 4 lowest frequency classes removed

Fig. 7 Accuracy–budget curves for the different active learning strategies on Tobacco3482 datasets with top 4 highest frequency classes removed (left) and top 4 lowest frequency classes removed (right)



(a) RVL-CDIP



(b) $\text{Tobacco3482}_{\text{ImageNet}}$

Fig. 8 Confusion matrices that show the degree of annotation noise added per class for the RVL-CDIP and Tobacco3482 datasets under noise percentage of $\epsilon = 10\%$. As illustrated, some classes were not subjected to annotation noise due to their distinct differences from others

Table 5 ResNet-50 performance under supervised learning with fully annotated training datasets (top) and active learning with 40% annotated training datasets and varying degrees of annotation noise (bottom)

	Model/query function (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
		Acc	AUBC	Acc	AUBC	Acc	AUBC
Supervised learning (Ann. Data = 100%)	ResNet-50 ($\epsilon = 10\%$)	88.62	–	83.28	–	92.00	–
	ResNet-50 ($\epsilon = 20\%$)	86.77	–	80.43	–	91.00	–
Active learning with ResNet-50 (Ann. Data = 40%)							
Noise percentage ($\epsilon = 10\%$)	Random Sampling	86.37	0.2529	79.00 ± 1.67	0.2516	91.91 ± 0.48	0.3188
	Entropy [19]	**89.00**	0.2591	**80.97 ± 0.94**	0.2543	92.23 ± 0.26	0.3207
	Least Conf. [8]	88.84	0.2590	79.54 ± 1.00	0.2540	92.17 ± 0.33	0.3208
	Margin [8]	88.57	0.2590	80.46 ± 0.90	0.2553	**92.43 ± 0.23**	**0.3212**
	Entropy (Dropout) [16]	88.77	0.2592	79.71 ± 1.23	0.2542	91.97 ± 0.29	0.3206
	Least Conf. (Dropout) [16]	88.71	0.2590	79.80 ± 1.32	0.2557	91.80 ± 0.26	0.3202
	Margin (Dropout) [16]	88.72	0.2590	80.03 ± 0.36	**0.2570**	91.80 ± 0.33	0.3209
	BALD (Dropout) [16]	86.14	0.2526	78.43 ± 0.59	0.2556	92.14 ± 0.52	0.3212
	KCenterGreedy [11]	88.30	0.2580	80.63 ± 1.12	**0.2570**	92.09 ± 0.19	0.3211
	KMeans [8]	*	*	78.17 ± 1.00	0.2499	91.60 ± 0.21	0.3190
	AdvBIM [24]	87.31	0.2548	80.20 ± 0.33	0.2551	91.83 ± 0.06	0.3200
	CEAL (Entropy) [18]	88.83	0.2595	78.09 ± 1.33	0.2528	92.14 ± 0.25	0.3207
	BADGE [27]	*	*	78.26 ± 1.24	0.2540	91.54 ± 0.19	0.3199
	BAIT [28]	*	*	79.09 ± 0.82	0.2539	91.60 ± 0.21	0.3197
	LPL [26]	88.70	**0.2603**	77.43 ± 2.79	0.2521	91.83 ± 0.79	0.3183
	WAAL [25]	88.11	0.2575	79.80 ± 1.09	0.2552	92.34 ± 0.13	0.3205
Noise percentage ($\epsilon = 20\%$)	Random Sampling	84.19	0.2473	77.23 ± 1.06	0.2473	91.14 ± 0.42	0.3174
	Entropy [19]	87.30	0.2542	**79.14 ± 0.83**	0.2502	91.51 ± 0.26	0.3175
	Least Conf. [8]	87.11	0.2536	78.54 ± 0.74	0.2492	91.57 ± 0.17	0.3180
	Margin [8]	87.30	0.2539	78.60 ± 1.27	**0.2526**	91.83 ± 0.23	0.3180
	Entropy (Dropout) [16]	87.28	0.2540	78.40 ± 0.79	0.2482	91.60 ± 0.19	0.3177
	Least Conf. (Dropout) [16]	87.27	0.2538	78.40 ± 0.79	0.2482	91.57 ± 0.29	0.3180
	Margin (Dropout) [16]	87.05	0.2536	77.74 ± 0.59	0.2478	91.74 ± 0.26	0.3175
	BALD (Dropout) [16]	84.51	0.2467	78.46 ± 1.06	0.2496	90.97 ± 0.19	0.3174
	KCenterGreedy [11]	86.77	0.2527	77.37 ± 0.54	0.2463	91.74 ± 0.27	0.3183
	KMeans [8]	*	*	76.03 ± 1.59	0.2466	91.60 ± 0.21	0.3190

Table 5 continued

Model/query function (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
	Acc	AUBC	Acc	AUBC	Acc	AUBC
AdvBIM [24]	86.83	0.2521	77.31 ± 0.69	0.2505	91.31 ± 0.36	0.3156
CEAL (Entropy) [18]	**88.19**	**0.2572**	75.83 ± 3.42	0.2471	91.51 ± 0.19	0.3174
BADGE [27]	*	*	75.74 ± 1.39	0.2462	91.54 ± 0.19	**0.3199**
BAIT [28]	*	*	75.86 ± 0.35	0.2447	91.60 ± 0.21	0.3197
LPL [26]	87.18	0.2550	73.69 ± 1.24	0.2407	91.83 ± 0.79	0.3183
WAAL [25]	86.50	0.2523	77.71 ± 1.17	0.2507	**91.86 ± 0.20**	0.3179

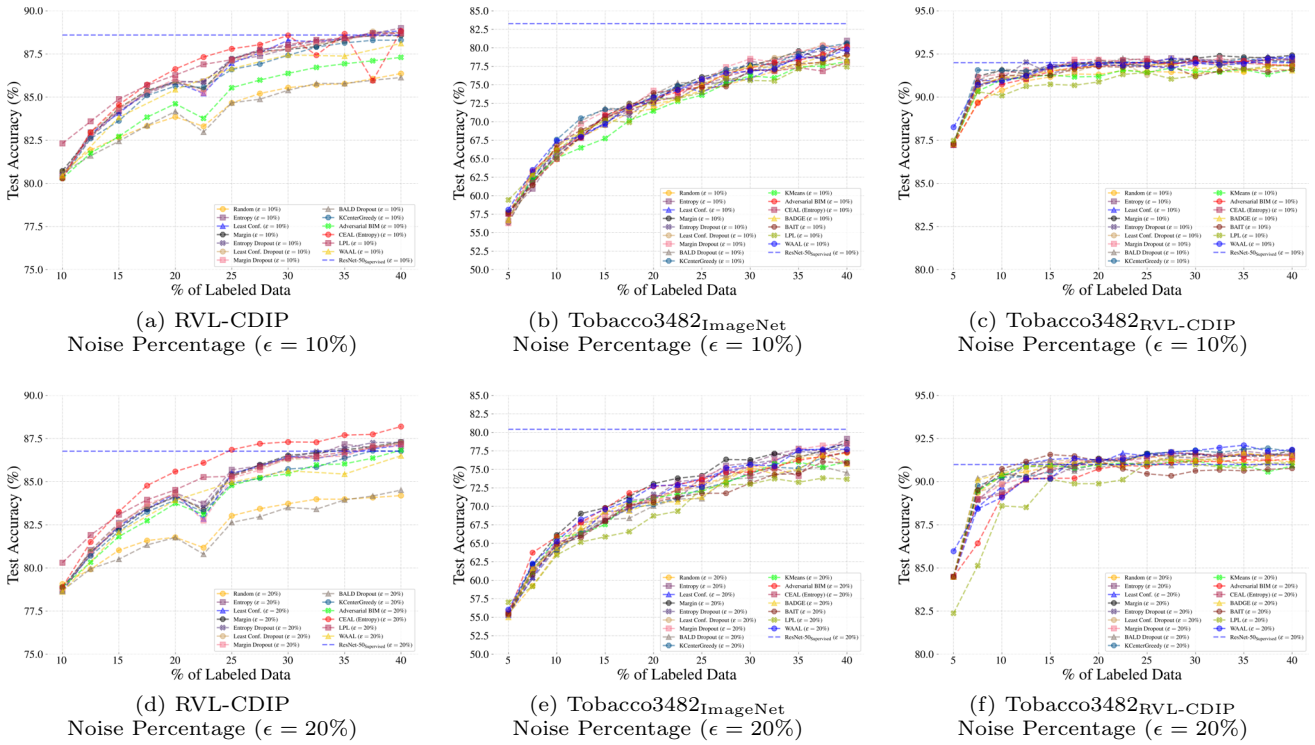
For each task, the highest accuracy and AUBC are bolded
*Computationally unfeasible for large datasets

Since document classes usually have very high intra-class variance and inter-class similarity, it was difficult to visually determine which classes should be considered similar in our scenario. Therefore, the weights $W \in \mathbb{R}^{N \times N}$ for the similarity between classes were heuristically determined by inspecting the classes between which the fully trained ResNet-50$_{Supervised}$ model had shown the highest confusion. For example, the confusion matrix of the ResNet-50$_{Supervised}$ model generated on the test set of RVL-CDIP dataset is depicted in Fig. 10a, which directly shows which document classes were confused with each other the most. We use this confusion to directly define the similarity. For example, to generate the class similarity weight for each class, we simply applied a threshold followed by normalization to the off-diagonal entries of the confusion matrix. The resulting similarity weights are presented in Fig. 10b, from which it can be seen that the classes Letter and Memo were determined to be mutually similar both with weights of 1.0. In many cases, it was also possible for one class to be similar to another, but not vice versa. For example, the class Scientific Report was found to be similar to the class News Article with a weight of 0.4 but the opposite was not true.

This experiment was conducted with two settings of percentage annotation noise (ϵ) per class, $\epsilon = 10\%$ and $\epsilon = 20\%$. The resulting annotation noise confusion matrices for the two datasets RVL-CDIP and Tobacco3482 after applying Algorithm 1 with percentage noise $\epsilon = 10\%$ per class are illustrated in Fig. 8. The results of this experiment are presented in Table 5, and the accuracy–budget curves for each AL method are illustrated in Fig. 9. For a fair evaluation of the performance of AL methods, the ResNet-50$_{Supervised}$ models in this scenario were also trained on the fully annotated noisy datasets and their performances are reported for each case in the table.

4.7.2 Results on RVL-CDIP

For RVL-CDIP dataset, it can be seen that for noise percentage $\epsilon = 10\%$ the performance of the Random Sampling baseline was severely affected. Most uncertainty-based approaches such as Entropy and Least Confidence, and some enhanced approaches including CEAL (Entropy) and LPL, were still able to counter the effects of noise significantly better in comparison, even surpassing the performance of the ResNet-50$_{Supervised}$ model with just 40% annotated data. While Entropy showed better accuracy, LPL showed a better overall performance with a considerably higher AUBC which can also be observed in Fig. 9a. A slightly different trend was seen for the noise percentage $\epsilon = 20\%$, where the performance of all the techniques was seen to be greatly affected by the noise. However, CEAL (Entropy) was still considerably effective in countering its effects, surpassing both the Random Sampling baseline and the fully trained ResNet-



(a) RVL-CDIP Noise Percentage ($\epsilon = 10\%$) (b) $\text{Tobacco3482}_{\text{ImageNet}}$ Noise Percentage ($\epsilon = 10\%$) (c) $\text{Tobacco3482}_{\text{RVL-CDIP}}$ Noise Percentage ($\epsilon = 10\%$)

(d) RVL-CDIP Noise Percentage ($\epsilon = 20\%$) (e) $\text{Tobacco3482}_{\text{ImageNet}}$ Noise Percentage ($\epsilon = 20\%$) (f) $\text{Tobacco3482}_{\text{RVL-CDIP}}$ Noise Percentage ($\epsilon = 20\%$)

Fig. 9 Accuracy–budget curves for the different active learning strategies on RVL-CDIP and Tobacco3482 datasets in the presence of annotation noise of varying degrees

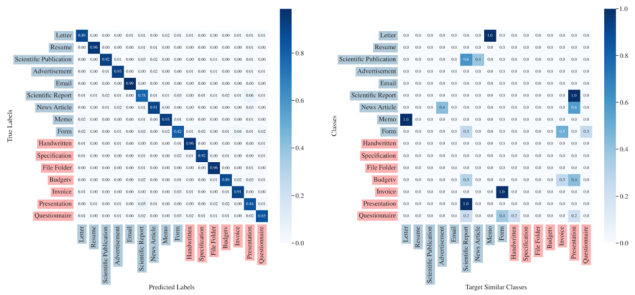
$50_{Supervised}$ model as evident from both Table 5 and Fig. 9d. Besides CEAL (Entropy), the uncertainty-based techniques and the enhanced LPL approach also showed competitive performance in this case, both of which outperformed the fully trained ResNet-$50_{Supervised}$ model at 40% annotated dataset.

4.7.3 Results on Tobacco3482$_{ImageNet}$

In Tobacco3482$_{ImageNet}$ setting, only uncertainty-based approaches such as Entropy and Margin Sampling seemed to consistently perform well for both cases $\epsilon = 10\%$ and $\epsilon = 20\%$. KCenterGreedy and AdvBIM still seemed to perform slightly better than the Random Sampling baseline, but their performance was severely deteriorated for the $\epsilon = 20\%$ case. Most enhanced approaches including CEAL (Entropy), BADGE, LPL, and WAAL were severely affected by the annotation noise and performed significantly worse than even Random Sampling baseline for the $\epsilon = 20\%$ case.

4.7.4 Results on Tobacco3482$_{RVL-CDIP}$

The Tobacco3482$_{RVL-CDIP}$ scenario showed a different trend, where WAAL showed slightly better performance than other techniques on both the $\epsilon = 10\%$ and $\epsilon = 20\%$ cases; however, the overall differences in performance were quite negligible between different approaches. Similar to previous cases, many of the techniques again surpassed the baseline performance even with just 15–25% of the training set queried as evident from Fig. 9c, f.



(a) RVL-CDIP (b) $W_{\text{RVL-CDIP}}$

Fig. 10 Confusion matrix of the ResNet-$50_{Supervised}$ on the test set of RVL-CDIP dataset (left) and the weights generated from it (right) are shown

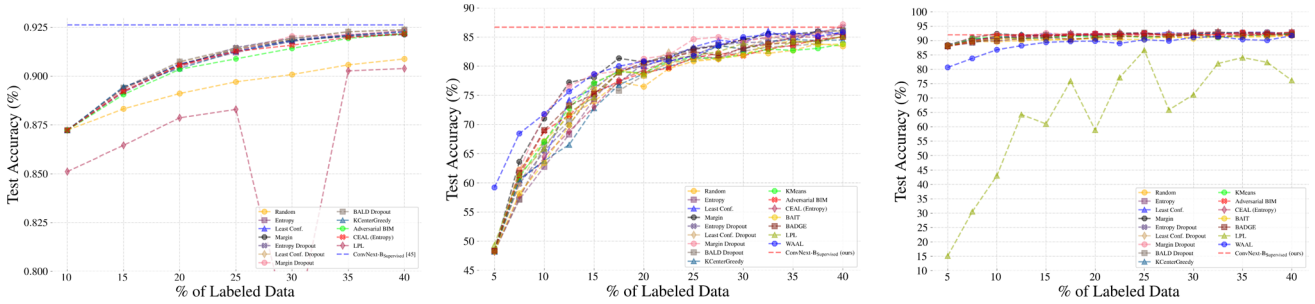
4.8 Generalization to other models

In this section, we investigate whether the AL strategies studied in this work are applicable to other deep networks that can be used for document classification. For this purpose, we apply the AL strategies explored in this paper to the recently introduced ConvNeXt [46] model, specifically

Table 6 ConvNeXt-B performance under supervised learning with fully annotated training datasets (top) and active learning with 40% annotated training datasets (bottom)

	Model ($\mathcal{M}$)/query strategy (f)	RVL-CDIP		Tobacco3482$_{ImageNet}$		Tobacco3482$_{RVL-CDIP}$	
		Acc	AUBC	Acc	AUBC	Acc	AUBC
Supervised learning (Ann. Data = 100%)	ConvNeXt-B$_{Supervised}$	92.63 [35]	–	86.71	–	92.00	–
Active learning with ResNet-50 (Ann. Data = 40%)	Random Sampling	90.88	0.2684	83.43 ± 1.21	0.2668	91.64 ± 0.10	0.3174
Unc. non-Bayesian	Entropy [19]	92.18	0.2723	85.36 ± 0.91	0.2692	92.93 ± 0.10	0.3213
	Least Conf. [8]	92.28	0.2725	85.43 ± 0.20	0.2727	92.29 ± 0.40	0.3218
	Margin [8]	92.13	0.2724	86.14 ± 0.20	0.2772	92.64 ± 0.30	0.3220
Unc. Bayesian	Entropy (Dropout) [16]	92.34	0.2725	86.71 ± 0.20	0.2709	92.21 ± 0.30	0.3210
	Least Conf. (Dropout) [16]	92.36	0.2727	86.07 ± 0.91	0.2725	92.93 ± 0.10	0.3216
	Margin (Dropout) [16]	92.31	0.2726	87.21 ± 0.10	0.2772	92.21 ± 0.30	0.3218
	BALD (Dropout) [16]	92.39	0.2728	86.14 ± 1.21	0.2673	92.50 ± 0.51	0.3207
Repr	KCenterGreedy [11]	92.26	0.2724	84.50 ± 0.51	0.2666	92.14 ± 0.20	0.3214
	KMeans [8]	*	*	83.79 ± 0.30	0.2691	92.29 ± 0.20	0.3195
Enh./Hybrid	AdvBIM [24]	92.18	0.2717	85.00 ± 0.40	0.2691	92.71 ± 0.40	0.3212
	CEAL (Entropy) [18]	92.17	0.2722	85.57 ± 0.81	0.2672	92.71 ± 0.61	0.3214
	BADGE [27]	*	*	85.07 ± 0.91	0.2716	91.93 ± 0.30	0.3172
	BAIT [28]	*	*	83.79 ± 0.10	0.2670	91.93 ± 0.30	0.3186
	LPL [26]	90.39	0.2584	85.14 ± 0.81	0.2699	76.14 ± 7.48	0.2321
	WAAL [25]	*	*	85.86 ± 0.20	0.2789	91.71 ± 0.00	0.3115

For each task, the highest accuracy and AUBC are bolded
*Computationally unfeasible for large datasets



(a) RVL-CDIP (b) Tobacco3482$_{\text{ImageNet}}$ (c) Tobacco3482$_{\text{RVL-CDIP}}$

Fig. 11 Accuracy–budget curves of the ConvNeXt-B model for the different active learning strategies on RVL-CDIP and Tobacco3482 datasets

its ConvNeXt-B variant. To train the model on RVL-CDIP and Tobacco3482 datasets, we used the same training strategy as [35] and trained the model with image resolution of 224 × 224, Adam optimizer, LabelSmoothing, and CutMix and Mixup augmentations. For RVL-CDIP dataset, we compare the results of our experiments with supervised learning performance achieved by [35] on ConvNeXt-B/224 scenario, whereas for the Tobacco3482$_{ImageNet}$ and Tobacco3482$_{RVL-CDIP}$ cases, we separately trained the model on full Tobacco3482 dataset for comparison. For a fair comparison, we used the same training strategies for both active learning and supervised learning. The results of these experiments are presented in Table 6 in which for each AL strategy, both the average accuracy achieved on 40% of original training datasets and the area under the budget curve (AUBC) are given. Figure 11 illustrates the budget–accuracy curves for each AL strategy, under different dataset settings, which indicate the accuracy achieved by the model after each AL round until a total of 40% of the training dataset was annotated.

It can be observed from both Table 6 and Fig. 11a that with just 40% of the training dataset annotated, the active learning approaches were able to achieve performances very close to the ConvNeXt$_{Supervised}$ [35] model. Similarly, for the Tobacco3482$_{ImageNet}$ and Tobacco3482$_{RVL-CDIP}$ cases, it can be seen from both Fig. 11b, c that the some active learning strategies were able to even outperform the fully trained supervised learning models. This suggests that active learning may have helped the model learn better distributions on the imbalanced dataset compared to supervised training. A noteworthy observation in this scenario was the instability of LPL strategy in both RVL-CDIP and Tobacco3482$_{RVL-CDIP}$ scenarios. While LPL worked quite well for the ResNet-50 model, for ConvNeXt-B, we observed quite a lot of instability in performance during training using the exact same LPL configuration. As shown in both Fig. 11a, c, the LPL technique performed very poorly in this case and we found it difficult to tune its hyperparameters to reach any satisfactory results. Similarly, we found it difficult to train ConvNeXt-B with WAAL on the RVL-CDIP dataset with the exact same configuration as in the case of ResNet-50, with the discriminator loss becoming unstable during training. Aside from these two enhanced approaches, we observed that most AL techniques generalized fairly well to this model and even produced exceptional performance on the datasets.

5 Practical implications

This section summarizes the overall results of our study and discusses its practical implications in the context of document classification. In Table 7, we present an overview of the performance of different AL strategies under different settings of datasets and experiments. In the top section, we present the top 3 highest performing AL approaches for each scenario, and in the bottom section, we present the top 3 worst performing AL approaches based on the experiments performed on the ResNet-50 model. The types of each approach are also highlighted with different colors in order to present an overall view of which types of approaches performed the best and the worst.

From the table, a few interesting observations can be drawn. First, we can observe that the non-Bayesian uncertainty-based approaches were not only computationally efficient, but also consistently produced the best results. On the larger RVL-CDIP dataset, the only two enhanced approaches that performed slightly better than others were LPL and CEAL (Entropy), with CEAL (Entropy) performing slightly better than others in the case of heavier annotation noise. It is worth mentioning, however, that from our results on ConvNeXt-B, we also observed difficulty in extending LPL to other models. In contrast, the three approaches BALD (Dropout), AdvBIM, and WAAL consistently performed worse than the others on this dataset. It is interesting to note that both AdvBIM and WAAL are enhanced approaches that require significantly more computational resources in comparison with the others, yet they failed to produce any satisfactory results in this case. Overall, we can conclude that for large class-balanced

Table 7 An overview of the top three most effective approaches as well as the top three least effective approaches for each of the scenarios investigated in this study. The approach types Unc. Non-Bayesian, Unc. Bayesian, Representation-based, and Enhanced/Hybrid are highlighted in Blue, Seagreen, Yellow, and Orange, respectively

Top 3 Best Performing Approaches for each Active Learning Scenario							
Datasets	Rank	Standard	Query Time	Data Imbalance(m = 2)	Data Imbalance (m = 4)	Noisy Datasets ($\epsilon = 10\%$))	Noisy Datasets $\epsilon = 20\%$)
RVL-CDIP	1	LPL	Margin	Entropy	LPL	Entropy	CEAL (Entropy)
	2	Margin	Entropy	LPL	Margin	Least Conf.	Entropy
	3	Entropy	Least Conf.	CEAL / Margin	Entropy (Dropout)	CEAL (Entropy)	Margin
$\text{Tobacco3482}_{\text{ImageNet}}$	1	Entropy	Entropy	Margin (Dropout)	Margin	Entropy	Entropy
	2	Margin (Dropout)	Least Conf.	KCenterGreedy	KCenterGreedy	KCenterGreedy	Margin
	3	BADGE	Margin	Least Conf.	Least Conf.	Margin	Least Conf.
$\text{Tobacco3482}_{\text{RVL-CDIP}}$	1.	AdvBIM	Entropy	BALD (Dropout)	Margin	Margin	WAAL
	2	Least Conf.	Least Conf.	Entropy	Least Conf.	WAAL	LPL
	3	Least Conf. (Dropout)	Margin	Least Conf. / Entropy (Dropout)	KCenterGreedy	Entropy	Entropy
Top 3 Worst Performing Approaches for each Active Learning Scenario							
Datasets	Rank	Standard	Query Time	Data Imbalance (m = 2)	Data Imbalance (m = 4)	Noisy Datasets ($\epsilon = 10\%$)	Noisy Datasets($\epsilon = 20\%$)
RVL-CDIP	1	BALD (Dropout)	WAAL	BALD (Dropout)	BALD (Dropout)	BALD (Dropout)	BALD (Dropout)
	2	AdvBIM	Entropy (Dropout)	AdvBIM	AdvBIM	AdvBIM	WAAL
	3	WAAL	Least (Dropout)	WAAL	WAAL	WAAL	KCenterGreedy
$\text{Tobacco3482}_{\text{ImageNet}}$	1	BALD (Dropout)	WAAL	KMeans	LPL	LPL	LPL
	2	KMeans	LPL	BALD (Dropout)	KMeans	CEAL (Entropy)	BADGE
	3	BAIT	BAIT	LPL	BALD (Dropout)	KMeans	CEAL (Entropy)
$\text{Tobacco3482}_{\text{RVL-CDIP}}$	1	LPL	WAAL	LPL	LPL	BADGE	BALD (Dropout)
	2	BADGE	LPL	WAAL	WAAL	KMeans	AdvBIM
	3	WAAL	BAIT	BAIT	BADGE	BAIT	Entropy

document datasets, the AL strategies that are most practical in terms of computational performance and computational costs are uncertainty-based approaches such as Entropy and Margin, as well as the enhanced techniques LPL and CEAL (Entropy). While LPL is difficult to train, it can result in slight performance gains. Thus, it is a reasonable choice if adequate training resources are available to tune hyperparameters. On the other hand, CEAL (Entropy) can be particularly useful when dealing with severe labeling noise.

In the $\text{Tobacco3482}_{ImageNet}$ setting, both Bayesian and non-Bayesian uncertainty-based approaches showed the highest performance across different scenarios. While most representation-based approaches were severely affected in the case of bias in the initial labeled dataset, KCenterGreedy fared relatively much better and showed performances similar to uncertainty-based approaches. In contrast, enhanced approaches such as LPL, BAIT, and BADGE were among the worst performing approaches. In addition, BALD (Dropout) and KMeans also performed significantly worse than others in this scenario. Overall, we conclude that for small, imbalanced document datasets without document-specific pretraining, uncertainty-based approaches such as Entropy and Margin as well as the KCenterGreedy (CoreSets) approach are most appropriate to attain both better performance and efficiency.

A similar trend was observed in the $\text{Tobacco3482}_{RVL-CDIP}$ scenario where uncertainty-based approaches outperformed other approaches. However, because most approaches performed at a similar scale in this scenario, it was difficult to identify any significant advantages of some approaches over others. Nevertheless, enhanced approaches such as LPL, BADGE, BAIT, and WAAL were generally among the least effective approaches, with the exception of noisy datasets, in which WAAL and LPL appeared to be more efficient. As a general rule, we again recommend that when using small datasets with document-specific pretraining, the simplest uncertainty-based approaches seem to be the most appropriate option, as they not only provide superior accuracy and computational performance, but also make the training process more convenient.

6 Conclusion

In this paper, we investigated the potential of active learning in reducing annotation costs while enabling machine learning models to perform competitively in document image classification. An analysis of different active learning strategies has revealed that deep learning models can achieve competitive performance with as little as 40% of the training datasets labeled by Oracle with the use of AL, thereby reducing annotation costs by up to 60%. Additionally, domain-specific pretraining was shown to significantly enhance AL performance on small datasets, allowing models to outperform models trained on fully annotated datasets with as little as 15% of the data annotated. We also demon-

strated that, in comparison with self-supervised learning approaches, AL strategies result in better performance on partially annotated datasets. While enhanced approaches such as LPL and CEAL (Entropy) surpassed the simpler uncertainty or representation-based approaches on the RVL-CDIP dataset, their performance was severely degraded under class imbalance on the Tobacco3482 dataset. On the other hand, uncertainty-based approaches such as Entropy and Margin performed more consistently on the Tobacco3482 dataset showing better performance under multiple scenarios of class imbalance, annotation noise, and data bias. Overall, it was observed that class imbalance in the dataset severely affects the performance of various recent SotA techniques such as BADGE, BAIT, and LPL. To address the issues of data imbalance, recently introduced class-balanced active learning approaches [43] may be explored in the future. Another plausible future work could be to explore active learning for multi-modal document analysis tasks, where both image and textual data are utilized in the training process.

Author Contributions SS and SA devised the project and the main conceptual idea. SS developed the technical details of the project, designed the experiments, implemented the work, and analyzed the results. The manuscript was written by SS in consultation with AD and SA, both of whom provided critical feedback and reviews on the manuscript. As the principal supervisor, SA was directly involved in shaping the research, analysis and the manuscript.

Funding This work was supported by the BMBF projects SensAI (BMBF Grant 01IW20007). Open Access funding enabled and organized by Projekt 951 DEAL.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Xu, Y.: *et al.* LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 20, pp. 1192–1200 (2020). arXiv:1912.13318
2. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 991–995 (2015)
3. Siddiqui S.A., Agne, S., Dengel, A., Ahmed, S.: Are deep models robust against real distortions? a case study on document image classification (2022). https://doi.org/10.20944/preprints202202.0058.v1
4. Ferrando, J.: *et al.* Improving accuracy and speeding up document image classification through parallel systems. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12138 LNCS, pp. 387–400 (2020). arXiv:2006.09141
5. Xu, Y. et al.: LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding, pp. 2579–2591 (Association for Computational Linguistics (ACL), 2021). 2012.14740
6. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Patt. Analy. Mach. Intell. **39**(11), 2298–2304 (2015)
7. Settles, B.: Computer Sciences Department Active Learning Literature Survey (2009)
8. Zhan, X., et al.: A comparative survey of deep active learning (2022). https://arxiv.org/abs/2203.13450
9. Li, X., Guo, Y.: Adaptive Active Learning for Image Classification, pp. 859–866 (2013)
10. Yang, Y., et al.: Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vis. **113**, 113–127 (2015). https://doi.org/10.1007/s11263-014-0781-x
11. Sener, O., Savarese, S.: Active Learning for Convolutional Neural Networks: A Core-Set Approach (2018). https://openreview.net/forum?id=H1aIuk-RW
12. Sinha, S., Ebrahimi, S., Darrell, T.: Variational Adversarial Active Learning (2019). https://arxiv.org/abs/1904.00370
13. Afzal, M. Z., Kolsch, A., Ahmed, S., Liwicki, M.: Cutting the Error by half: investigation of very deep cnn and advanced training strategies for document image classification. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Vol. 1, pp. 883–888 (2017). arXiv:1704.03557
14. Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network (2018). https://arxiv.org/abs/1806.05473
15. Mayer, C., Timofte, R.: Adversarial sampling for active learning (2018). https://arxiv.org/abs/1808.06671
16. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian Active Learning with Image Data. *34th International Conference on Machine Learning, ICML 2017*, Vol. 3, pp. 1923–1932 (2017). https://arxiv.org/abs/1703.02910v1. https://doi.org/10.48550/arxiv.1703.02910
17. Cai, W., et al.: Active learning for support vector machines with maximum model change. Undefined 8724 LNAI (PART 1), pp. 211–226 (2014). https://doi.org/10.1007/978-3-662-44848-9_14
18. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Trans. Circuits Syst. Video Technol. **27**(12), 2591–2600 (2016)
19. Shannon, C.E.: A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. **5**(1), 3–55 (2001). https://doi.org/10.1145/584091.584093
20. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning (2015). https://arxiv.org/abs/1506.02142
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014)
22. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation (2017). https://arxiv.org/abs/1706.04737

23. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach (2018). https://arxiv.org/abs/1802.09841
24. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2016). https://arxiv.org/abs/1607.02533
25. Shui, C., Zhou, F., Gagné, C., Wang, B.: Deep active learning: unified and principled method for query and training (2019). https://arxiv.org/abs/1911.09162
26. Yoo, D., Kweon, I.S.: Learning loss for active learning (2019). https://arxiv.org/abs/1905.03677
27. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: deep batch active learning by diverse, uncertain gradient lower bounds (2019). https://doi.org/10.48550/ARXIV.1906.03671
28. Ash, J.T., Goel, S., Krishnamurthy, A., Kakade, S.: Gone fishing: neural active learning with fisher embeddings. Adv. Neural Inf. Process. Syst. **11**, 8927–8939 (2021)
29. Shin, C.K., Doermann, D.S.: Document image retrieval based on layout structural similarity. In: Proceedings of 2006 International Conference on Image Process, Computer Vision and Pattern Recognition, Vol. 2, pp. 606–612 (2016)
30. Kumar, J., Ye, P., Doermann, D.: Structural similarity for document image classification and retrieval. Patt. Recognit. Lett. **43**(1), 119–126 (2014)
31. Baldi, S., Marinai, S., Soda, G.: Using tree-grammars for training set expansion in page classification. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2003-Janua (Icdar), pp. 829–833. (2003)
32. Diligenti, M., Frasconi, P., Gori, M.: Hidden tree Markov models for document image classification. Patt. Anal. Mach. Intell. IEEE Trans. **25**, 519–523 (2003). https://doi.org/10.1109/TPAMI.2003.1190578
33. Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for document image classification. In: Proceedings of the International Conference on Pattern Recognition, pp. 3168–3172 (2014)
34. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In: Proceedings of the International Conference on Pattern Recognition, 2018-Augus, pp. 3180–3185 (2018). arXiv:1801.09321
35. Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: DocX-Classifier: towards an interpretable deep convolutional neural network for document image classification. (2022,9), https://doi.org/10.36227/techrxiv.19310489.v4
36. Siddiqui, S., Dengel, A., Ahmed, S.: Self-supervised representation learning for document image classification. IEEE Access **9**, 164358–164367 (2021)
37. Cosma, A., Ghidoveanu, M., Panaitescu-Liess, M., Popescu, M.: Self-Supervised Representation Learning on Document Images. 2020, https://arxiv.org/abs/2004.10605
38. Powalski, R., et al.: Going full-tilt boogie on document understanding with text-image-layout transformer (2021). https://arxiv.org/abs/2102.09550
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://arxiv.org/abs/1512.03385
40. Deng, J., et al.: Imagenet: a large-scale hierarchical image database, pp. 248–255 (2009)
41. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow Twins: Self-Supervised Learning via Redundancy Reduction. 2021, https://arxiv.org/abs/2103.03230
42. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. 2020, https://arxiv.org/abs/2002.05709
43. Bengar, J.Z., van de Weijer, J., Fuentes, L.L., Raducanu, B.: Class-balanced active learning for image classification (2021). https://arxiv.org/abs/2110.04543
44. Girden, E.: ANOVA: Repeated Measures. Sage (1992)
45. Gupta, G., Sahu, A.K., Lin, W.-Y.: Noisy batch active learning with deterministic annealing (2019). https://arxiv.org/abs/1909.12473
46. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. (2022), https://arxiv.org/abs/2201.03545

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.