Active Learning with Fast Model Updates and Class-Balanced Selection for Imbalanced Datasets

Zhixin Huang[1,*,†], Yujiang He[1,†], Marek Herde[1], Denis Huseljic[1] and Bernhard Sick[1]

[1]University of Kassel, Germany

Abstract

Active learning strategies aim to minimize the number of queried samples for model training. However, two challenges in pool-based deep active learning on imbalanced datasets are observed in experiments: (1) the declining performance of active learning strategies due to imbalanced class distribution; (2) the lack of sample diversity in acquisition batches due to the absence of timely model updates. This paper proposes the AL-FaMoUS, a general solution combining fast model updates and class-balanced minibatch selection to the active learning process. Furthermore, an implementation of the AL-FaMoUS, which selects one single sample in each acquisition minibatch, is experimentally evaluated on four image and three time-series imbalanced datasets. The results demonstrate that the implemented AL-FaMoUS outperforms the other adopted AL strategies, including uncertainty sampling and BALD solely combined with either the fast model update or the class balance selection strategy, in terms of Macro F1 score.

Keywords

Active Learning, Imbalanced Dataset, Fast Model Updates, Class-Balanced Selection

1. Introduction

Active learning (AL) describes a machine learning paradigm to select the most informative and representative samples for annotation. It aims to optimize the performance of the model while reducing the overhead of annotating by querying as few samples as possible with high quality. In the general AL setup, there is a large unlabeled dataset $\mathcal{D}_u$ and a total annotation budget B, which indicates the number of samples to be annotated. Assuming the AL algorithm queries the oracle for labels of b samples at each cycle, a complete active learning process is done within $\lfloor B/b \rfloor$ cycles. Each cycle can be briefly described as follows: (1) calculating the utility scores $\boldsymbol{u}$ of the unlabeled samples using a specific acquisition function; (2) ranking the utility scores and selecting the top b samples to annotate; (3) adding the b labeled samples to the labeled dataset $\mathcal{D}_l$, based on which the model is retrained; (4) utilizing the retrained model for the next cycle until the budget runs out. AL selection strategies can be distinguished by the selected acquisition function.

IAL@ECML-PKDD'23: 7th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 22nd, 2023, Torino, Italy

*Corresponding author.

†These authors contributed equally.

zhixin.huang@uni-kassel.de (Z. Huang); yujiang.he@uni-kassel.de (Y. He); marek.herde@uni-kassel.de (M. Herde); dhuseljic@uni-kassel.de (D. Huseljic); bsick@uni-kassel.de (B. Sick)

0000-0002-6124-0206 (Z. Huang); 0000-0001-5817-5805 (Y. He); 0000-0003-4908-122X (M. Herde); 0000-0001-6207-1494 (D. Huseljic); 0000-0001-9467-656X (B. Sick)

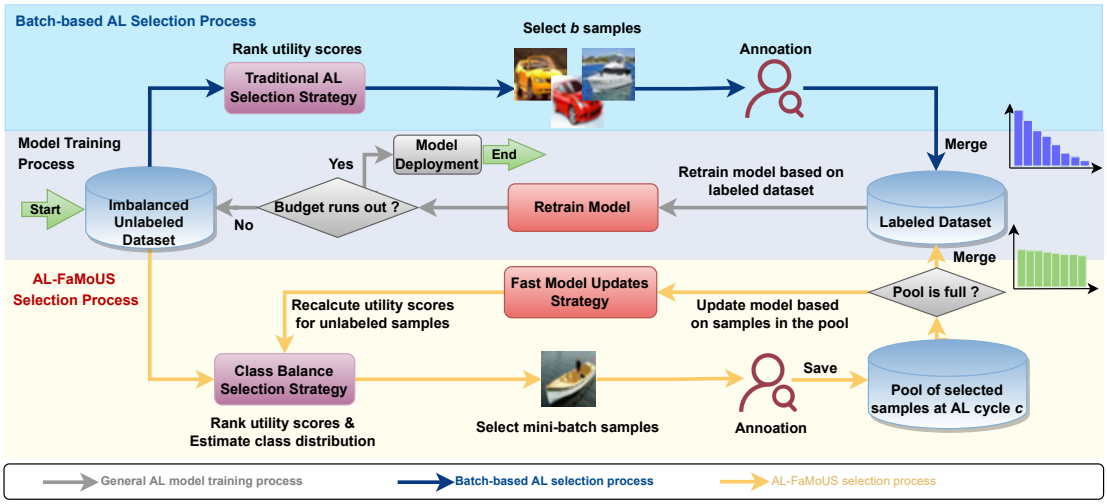


Figure 1: Comparison between the batch-based AL process (in blue) and the proposed AL-FaMoUS process (in yellow). The former only ranks the utility scores and queries the b samples at once in each cycle, and the latter queries the b_{mini} samples m times until the pool $\mathscr{D}^{\oplus}$ used to store the mini-batch samples is full. After each query, the model is fast updated based on the extended pool $\mathscr{D}^{\oplus}$ and then used to recalculate the utility scores of the remaining unlabeled samples. The class-balanced selection strategy ranks the updated utility scores and estimates the class distribution to query the next mini-batch of samples.

After reviewing related research, we find two main challenges that may emerge as AL is applied to real-world scenarios. The first is the declining performance of AL strategies on imbalanced datasets. AL is generally studied on close to uniform datasets where a similar amount of samples per class is available [1]. Most existing AL algorithms focus on ranking the utility of samples across all classes, which proves effective in balanced class scenarios. However, in the presence of imbalanced classes, not only the contribution of single samples differs but also the proportion of each class is various [2]. Imbalance in $\mathscr{D}_l$ and $\mathscr{D}_u$ poses drawbacks regarding model training and sample selection, respectively. In the former, during the optimization process, each training set highly likely contains a higher proportion of major-class samples selected from imbalanced $\mathscr{D}_l$, leading to a tendency for the optimization to reduce the loss of the model in terms of major-class samples [3, 4]. In the latter, the model tends to disregard minor-class samples due to their lack of representation [5], resulting in a preference for selecting major-class samples from $\mathscr{D}_u$ in query [6]. This worsens the imbalance in $\mathscr{D}_l$. The existing literature has proposed various methods to solve the imbalanced problem [7, 8, 9, 10, 11, 12], whose core idea is to query more labels of informative minor-class samples to keep the labeled data balanced. These methods are successfully applied to the batch-based selection in the deep active learning (DAL) scenario, that is, b is much greater than 1. But the batch selection in DAL leads to the next challenge, i.e., a lack of sample diversity within a batch.

Traditional active learning approaches [13, 14, 15], characterized by simpler models and smaller datasets, commonly employ a single-sample-based selection strategy ($b = 1$). However, in DAL context, with the complexity of deep neural networks (DNNs) and the scale of datasets

increasing significantly, the computational cost of retraining the model from an initialization state at each AL cycle becomes substantial. As a result, samples are typically selected in batches for annotation to reduce the number of model retraining iterations [16]. However, it can postpone the retraining of models [17]. An out-of-date model that only learned limited learning patterns can hardly distinguish various novel patterns and thus may select the top b samples containing the same learning pattern at one cycle. In [17] and our experiments in section 4, it was observed that the samples tend to be homogeneous in the same acquisition batch due to the absence of timely model updates. This problem is called a loss of sample diversity in this paper.

Two bullet points of this paper can be summarized. First, to address the above challenges, we propose a **A**ctive **L**earning Process with **Fa**st **Mo**del **U**pdates and a Class-Balanced Minibatch **S**election strategy, referred to as **AL-FaMoUS**, as illustrated in Fig. 1. Compared to the batch-based AL process, which ranks the unlabeled samples only by the utility scores, the AL-FaMoUS exploits specific class-balanced selection strategies to additionally evaluate the importance of these samples from the perspectives of class distributions. In AL-FaMoUS, the budget per each cycle b is further divided into m mini-batches, i.e., AL-FaMoUS selects $b_{\text{mini}} = \lfloor b/m \rfloor$ samples from D_u for annotation at each time and then adds them to a pool $\mathscr{D}_i^{\oplus}$, where $i \in \{0, 1, \dots, m\}$ and $\mathscr{D}_0^{\oplus} = \emptyset$. Next, the model is fast updated based on the pool $\mathscr{D}_i^{\oplus}$ to avoid an overestimation of the utility scores of the samples containing a known learning pattern [17]. The class-balanced selection strategy calculates and ranks the utility scores based on the updated model and then queries the next b_{mini} samples according to the estimated class distribution. When the pool size equals the budget b, the fast update cycle ends, and the samples in the pool are transferred to the labeled dataset. The model will then be completely retrained based on the labeled dataset. AL-FaMoUS is a general-solution-oriented process, which does not have any restriction on the setting of the acquisition function, the neural network's architecture, the class selection strategy, and the model update strategy. All configurations are completely dependent on the users' requirements and application conditions.

Second, to evaluate the performance ability of the proposal, a implementation of the AL-FaMoUS combined with a Bayesian fast update strategy and a class-balance selection strategy is experimentally evaluated on 7 imbalanced datasets. The results demonstrate the better performance of the implemented AL-FaMoUS than other adopted AL strategies in various applications.

2. Related Work

We can briefly categorize learning strategies on imbalanced datasets into two types: (1) increasing the amount of minor-class samples and (2) weighting the major classes and minor classes. The first type can be done with the help of oversampling [7, 8] or creating synthetic minority class samples by generative models [12]. However, in most cases, generative models are hardly trainable with limited minor-class samples. Other than that, the oversampling strategies trend to artificially imitate the known learning patterns for class balance in the dataset but can lead to a loss of sample diversity, eventually making models overfit the training samples [2]. The second type [9, 10, 11] increases the weight of the minor-class samples in loss during training but makes the information of major-class samples partially ignored [6].

These learning strategies guide the idea of AL in the direction of annotating the minor-class samples through oracle to provide a more balanced class distribution of the training data for model training, such as in [18, 19]. However, various limitations and challenges are still present. For example, Lin et al. proposed an active-learning-based search engine assisting oracle in annotating highly informative samples [20]. But this work is specifically applicable to language datasets. Lei et al. proposed a method to rank the priority of annotation for minor-class samples, which is mainly applicable to binary classification problems [21]. Similar research on binary classification was done by [22]. Novel acquisition functions were proposed in [23, 2] to select samples on imbalanced datasets based on Bayesian Active Learning by Disagreement (BALD, [24]) and uncertainty sampling, respectively. However, both works are not straightforward to adapt to other active learning strategies. Aggarwal et al. also proposed a novel acquisition function, which must require a pre-trained model based on an independent dataset [1]. Additionally, none of these works argue the imbalance problem and the loss of sample diversity existing in applications of time series analysis.

Model updating research investigates how to update trained models continually to adapt to novelties emerging in data streams. Continual learning is a method that enables deep neural networks to learn tasks sequentially while alleviating the forgetting problem, such as [25, 26, 27, 28]. However, these methods are updating strategies for deep neural networks based on a large training dataset, which indicates a high dependency on data collection and computation overhead. As one of the implementations of the AL-FaMoUS, we optimized an original batch-based class balancing selection strategy [6] and adopted a fast Bayesian update method [17] based on last-layer Laplace approximations (last-layer LA [29]) via Spectral Normalized Neural Gaussian Process (SNGP [30]). It updates the approximate weight distribution of the last layer instead of retraining the whole neural network.

3. AL-FaMoUS

This section introduces one of the implementations of AL-FaMoUS, where b_{mini} is set to one, i.e., the model queries a single sample for the orcale at each time by considering the class balance. After each query, the model will be updated quickly using the fast Bayesian update method. The simplified method is referred to as single-sample-based AL-FaMoUS in this paper and is formalized in Algorithm 1.

Let there be an unlabeled dataset $\mathscr{D}_u$ with N samples of K categories and a labeled dataset $\mathscr{D}_l$. Using Bayes' theorem, we can estimate a posterior distribution $p(\boldsymbol{\omega}|\mathscr{D}_l)$ over weights $\boldsymbol{\omega}$ of a Bayesian neural network (BNN) with the given labeled dataset $\mathscr{D}_l$. At the beginning of each AL cycle, a sample pool $\mathscr{D}^{\oplus}$ is created to store the newly annotated sample. We define $\mathscr{D}_i^{\oplus} = \mathscr{D}_{i-1}^{\oplus} \cup \{(\boldsymbol{x}_i, y_i)\}, i = 1, 2, \dots, b$, where $\mathscr{D}_0^{\oplus}$ is an empty set. The unlabeled sample $\boldsymbol{x}_i \in \mathscr{D}_u$ is queried by considering both the ranked utility score and the estimated class distribution, and y_i refers to the corresponding label given by the oracle. The model is fast updated based on the $\mathscr{D}_i^{\oplus}$ after each query, as proposed in subsection 3.2. At the end of each AL cycle, the b samples in the pool $\mathscr{D}^{\oplus}$ are transferred to $\mathscr{D}_l$. The initial size of $\mathscr{D}_l$ is denoted as b_0. At the cost of a low computation overhead for multiple fast updates, the AL method can guarantee the balance and the diversity of samples in the labeled dataset.

Algorithm 1 Single-sample-based AL-FaMoUS

Input: Unlabeled dateset $\mathscr{D}_u$, Budget per Cycle b, Initial labeled dataset $\mathscr{D}_l$, Total Budget B
Init: $c = 0$, $\mathscr{D}^{\oplus} = \mathscr{D}_0^{\oplus} = \emptyset$
while $c \leq \lfloor \frac{B}{b} \rfloor$ **do**
  (Re)train a BNN based on $\mathscr{D}_l$, obtain $q(\boldsymbol{\omega}_l|\mathscr{D}_l)$ by Eq. 6
  $c \leftarrow c + 1$
  $\mathscr{D}_0 = \mathscr{D}_l \cup \mathscr{D}_0^{\oplus}$
  $i = 1$
  while $i \leq b$ **do**
    Compute $\boldsymbol{P}$ by $p(y \mid \boldsymbol{x}, \mathscr{D}_{i-1})$ for $\boldsymbol{x} \in \mathscr{D}_u$
    Compute $\boldsymbol{\alpha}(c)$ from Eq. 1
    Calculate utility scores $\boldsymbol{u}$ for $\boldsymbol{x} \in \mathscr{D}_u \setminus \mathscr{D}_{i-1}^{\oplus}$
    Solve Eq. 3 to obtain $\boldsymbol{z}$ where $\|\boldsymbol{z}\|_1 = 1$
    Obtain label y_i for $\boldsymbol{x}_i$, where $z_i = 1$
    $\mathscr{D}_i^{\oplus} = \mathscr{D}_{i-1}^{\oplus} \cup \{(\boldsymbol{x}_i, y_i)\}$
    Fast Bayesian update by Eq. 7 to obtain updated distribution of $\boldsymbol{\omega}_l$.
    $i \leftarrow i + 1$
  end while
  $\mathscr{D}_l \leftarrow \mathscr{D}_l \cup \mathscr{D}_b^{\oplus}$, $\mathscr{D}_u \leftarrow \mathscr{D}_u \setminus \mathscr{D}_b^{\oplus}$, $\mathscr{D}^{\oplus} = \emptyset$
end while

3.1. Class Balance Selection

We define a matrix $\boldsymbol{P}$ with N rows and K columns. The component $p_{n,k}$ represents the model's softmax probabilistic output with respect to the unlabeled sample $\boldsymbol{x}_n \in \mathscr{D}_u$ belonging to the category k. Besides, we estimate a vector $\boldsymbol{\alpha}(c)$ at an AL cycle c, which describes the difference in the number of samples with respect to each category between the labeled dataset's distribution (imbalanced) and the desired distribution (balanced), formatted as follows:

$$\boldsymbol{\alpha}(c) = [\alpha_1^c, \alpha_2^c, \dots, \alpha_K^c]^{\mathrm{T}}, \tag{1}$$

α_k^c refers to the number of samples belonging to category k that should be annotated at the AL cycle c, and can be estimated by

$$\alpha_k^c = \max\left(\frac{b_0 + (c-1)b + |\mathscr{D}^{\oplus}|}{K} - n_k, 0\right). \tag{2}$$

$|\mathscr{D}^{\oplus}|$ denotes the number of samples that have been annotated and stored in the pool, $c = 1, 2, \dots, \lfloor B/b \rfloor$ denotes the index of the current cycle, and n_k is the number of samples belonging to category k that were annotated in the previous cycles.

Next, we define a vector $\boldsymbol{z}$, where each binary variable $z_n \in \{0, 1\}$ indicates whether the corresponding sample $\boldsymbol{x}_n$ is selected for annotation or not. $\|\boldsymbol{z}\|_1$ equals to one in the setup of single-sample-based AL selection strategy, where $\|\cdot\|_1$ refers to the L1-Norm. By maximizing the $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u}$, i.e., the summed utility scores of the selected samples, the optimal selection results can

be obtained. Furthermore, in order to guarantee the class balance, we add a regularization term $\|\boldsymbol{\alpha}(c) - \boldsymbol{P}^{\mathrm{T}}\boldsymbol{z}\|_1$ to the optimization goal. This regularization is defined as the distance between the desired and the estimated class distribution at AL cycle c [6]. The former refers to the required samples for each class at AL cycle c, while the latter refers to the distribution of selected samples in $\mathscr{D}_u$ with the pseudo labels that are given according to the softmax probability matrix. In this way, the overall optimization function is written as:

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} \left\{ -\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u} + \lambda \|\boldsymbol{\alpha}(c) - \boldsymbol{P}^{\mathrm{T}}\boldsymbol{z}\|_1 \right\}, \tag{3}$$

where λ is the regularization parameter that controls the contribution of class balance in selection.

3.2. Fast Bayesian Update

The posterior distribution of the BNN based on all labeled data can be expressed as $p(\boldsymbol{\omega}|\mathscr{D}_l \cup \mathscr{D}_i^{\oplus})$ with $i = 0, 1, \ldots, b$. For the sake of brevity, we use the abbreviation $\mathscr{D}_i = \mathscr{D}_l \cup \mathscr{D}_i^{\oplus}$ here. As explained above, the $p(\boldsymbol{\omega}|\mathscr{D}_0)$ refers to the initial BNN available at the beginning of each AL cycle, namely $\mathscr{D}_0 = \mathscr{D}_l$ because $\mathscr{D}_0^{\oplus}$ is an empty set. With the given posterior distribution, the predictive distribution for an unlabeled sample $\boldsymbol{x}_i$ can be specified via marginalization:

$$p(y_i \mid \boldsymbol{x}_i, \mathscr{D}_i) = \mathop{\mathbb{E}}_{\boldsymbol{\omega}|\mathscr{D}_i} \Big[p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\omega}) \Big] = \int p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathscr{D}_i) \, \mathrm{d}\boldsymbol{\omega}. \tag{4}$$

The distribution $p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\omega})$ denotes the probabilistic output of a neural network with weights $\boldsymbol{\omega}$. Instead of retraining the BNN from randomly initialized weights, we adjust the distributional parameters to update the posterior distribution once a new sample $\boldsymbol{x}_i$ is annotated. That is, the initial BNN's posterior distribution $p(\boldsymbol{\omega} \mid \mathscr{D}_l)$ serves as a prior distribution used to estimate the updated BNN's posterior distribution, which can be written as follows:

$$p(\boldsymbol{\omega} \mid \mathscr{D}_i) = \frac{p(\mathscr{D}_i^{\oplus} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathscr{D}_l)}{p(\mathscr{D}_i^{\oplus} \mid \mathscr{D}_l)} \propto p(\mathscr{D}_i^{\oplus} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathscr{D}_l) \propto \prod_{(\boldsymbol{x},y) \in \mathscr{D}_i^{\oplus}} p(y \mid \boldsymbol{x}, \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathscr{D}_l), \tag{5}$$

where $\mathscr{D}_i^{\oplus} = \mathscr{D}_{i-1}^{\oplus} \cup \{(\boldsymbol{x}_i, y_i)\}$ and $\mathscr{D}_i = \mathscr{D}_l \cup \mathscr{D}_i^{\oplus}$. Here, $\mathscr{D}_i^{\oplus}$ and $\mathscr{D}_l$ are assumed to be independently distributed.

As a specific BNN, we employ the SNGP [30] with last-layer LA [29] as the fast Bayesian update method [17]. SNGP is composed of last-layer LA with spectral normalization [31] and random Fourier features (RFF) [32]. SNGP learns hidden features through spectral normalization as the output of the penultimate layer. By applying an RFF mapping to these outputs, we obtain a D-dimensional representation, denoted as $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^D$. The last-layer LA is then performed on $\boldsymbol{\phi}(\boldsymbol{x})$ using an approximate multivariate normal distribution over the weights of the last layer $\boldsymbol{\omega}_l \in \mathbb{R}^D$ as follows:

$$q(\boldsymbol{\omega}_l \mid \mathscr{D}_l) = \mathcal{N}(\boldsymbol{\omega}_l \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \propto q(\boldsymbol{\omega}_l) \prod_{(\boldsymbol{x},y) \in \mathscr{D}_l} p(y \mid \boldsymbol{x}, \boldsymbol{\omega}_l). \tag{6}$$

Here, $q(\boldsymbol{\omega}_l)$ represents the prior, and $\hat{\boldsymbol{\mu}} \in \mathbb{R}^D$ and $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{D \times D}$ denote the mean vector and variance matrix of the approximate distribution over $\boldsymbol{\omega}_l$, respectively.

The last layer is fast updated based on $\mathscr{D}_i^{\oplus}$ by adjusting the parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ once the new sample pair $\{(\boldsymbol{x}_i, y_i)\}$ is annotated. We update the approximated distribution of the weight in the last layer using the method based on the Gauss-Newton algorithm, which was proposed in [17]. The method can be expressed as follows:

$$q(\boldsymbol{\omega}_l \mid \mathscr{D}_i) \propto q(\boldsymbol{\omega}_l \mid \mathscr{D}_l) \prod_{(\boldsymbol{x},y)\in\mathscr{D}_i^{\oplus}} p(y \mid \boldsymbol{x}, \boldsymbol{\omega}_l) \approx \mathcal{N}(\boldsymbol{\omega}_l \mid \hat{\boldsymbol{\mu}}_i^{\text{upd}}, \hat{\boldsymbol{\Sigma}}_i^{\text{upd}}), \tag{7}$$

where $\hat{\boldsymbol{\mu}}_i^{\text{upd}}$ and $\hat{\boldsymbol{\Sigma}}_i^{\text{upd}}$ denote the updated mean weight vector and covariance matrix, respectively.

According to Eq. 4 and Eq. 7, the fast updated model can remake predictions of samples in $\mathscr{D}_u$ via mean-field approximation [33] on the updated normal distribution:

$$p(y \mid \boldsymbol{x}, \mathscr{D}_i) \approx \text{softmax}\left(\frac{\boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\, \hat{\boldsymbol{\mu}}_i^{\text{upd}}}{\sqrt{1 + \pi/8 \cdot \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\, \hat{\boldsymbol{\Sigma}}_i^{\text{upd}}\, \boldsymbol{\phi}(\boldsymbol{x})}}\right). \tag{8}$$

4. Experiment

This section evaluates the proposed single-sample-based AL-FaMoUS across four image and three time-series datasets. For a comprehensive evaluation, we employed various models and utilized the AL strategies of Least Confidence [14] based on uncertainty sampling (US) and BALD as baselines. The experimental results demonstrate that the single-sample-based AL-FaMoUS outperforms the baselines in terms of F1-score.

4.1. Experiment Setup

4.1.1. Imbalanced Datasets

The experiments were conducted on four image classification datasets (MNIST [34], LETTER [35], FMNIST [36], and CIFAR10 [37]) and three time-series classification datasets (ECG5000 [38], CROP [39], and Electric Devices [40]). The original datasets are balanced, except for ECG5000 and Electric Devices. As in [41] we randomly selected 50% of the classes in these datasets and removed 90% of the samples belonging to these labels to make originally balanced datasets imbalanced.

4.1.2. Strategies

We used random sampling (RS), US, and BALD as the baselines. As common active learning strategies, the comparison among the three baselines can reflect whether US and BALD could effectively improve the classification performance in most datasets by selecting more informative samples. Furthermore, based on the acquisition functions of US and BALD, we employed and evaluated the fast Bayesian update (FBU), batch-based class balance selection methods (CB), i.e. $\|\boldsymbol{z}\|_1 = b$, and the proposed single-sample-based AL-FaMoUS, respectively. An example regarding the expected behavior for different selection strategies is illustrated in Fig. 2.

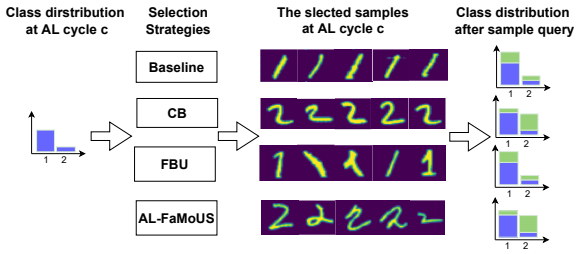
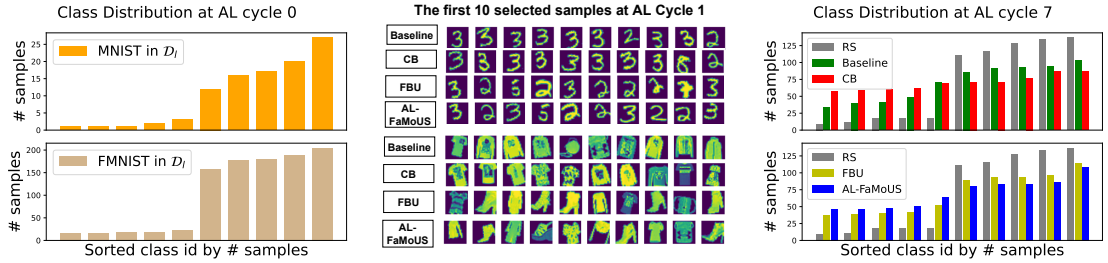


Figure 2: An example to explain the expected different behavior of the adopted selection methods on an imbalanced dataset of binary classification. Compared to Baseline, CB could effectively improve the proportion of the minor class in the labeled dataset by choosing more minor-class samples. But these selected samples contain a homogeneous learning pattern, the same as Baseline. The samples selected by FBU contain diverse learning patterns but worsen the imbalance problem. AL-FaMoUS could outperform them from both perspectives.



(a) Initial imbalanced $\mathcal{D}_l$ (b) Selected samples at AL cycle 1 (c) $\mathcal{D}_l$ of MNIST at AL cycle 7

Figure 3: Demonstration of different observations for various selection strategies applied to MNIST and FMNIST. The initially labeled dataset is imbalanced. Compared with Baseline and CB, FBU and AL-FaMoUS prefer to select the samples containing various learning patterns in a batch. After several AL cycles, the class distribution of $\mathcal{D}_l$ is shown in 3c. Compared with Baseline and FBU, CB and AL-FaMoUS can make $\mathcal{D}_l$ more balanced.

4.1.3. Implementation

The setup parameters are different for each dataset. Up to 2% of the samples in each imbalanced dataset were randomly selected to initialize the labeled dataset $\mathcal{D}_l$. Therefore, these initial $\mathcal{D}_l$ are also imbalanced. The examples of MNIST and FMNIST are shown in 3a. The budget b per AL cycle was in the range between 32 and 300.

We used ResNet-6 [42] as the backbone of SNGP for MNIST and FMNIST, and ResNet-18 for CIFAR10. The SNGP with a fully connected network was used for LETTER. These models were trained using the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and an initial learning rate of 0.1. Furthermore, we used temporal convolutional network (TCN, [43]) as the backbone of SNGP for the datasets ECG, CROP, and Electric Devices. The TCNs were trained using the Adam optimizer [44] with an initial learning rate of 0.001.

Similar to [6], we used the model trained on initial $\mathcal{D}_l$ to set the regularization parameter λ on each dataset. Fig.4 gives an example regarding the impact of λ on the regularization term.

As λ increases, the regularization term in Eq. 3 approaches convergence and does not decrease further. We choose the smallest λ close to convergence from three options $\{0.05, 0.1, 0.5\}$. CB and AL-FaMoUS used the identical and constant λ on the same dataset. See more details about hyperparameter selection in Appendix A.

4.1.4. Evaluation Metrics

We evaluated the performance of all strategies on the test data under Macro F1 score [45] and the class balance index (CBI). Macro F1 score highlights the performance of these strategies on the minor classes. CBI reflects the final class balance of the algorithm's cumulative selection samples.n_k is the number of samples belonging to the category k that were annotated in the previous $\lfloor B/b \rfloor$ cycles. s_k represents the proportion of samples with category k in all annotated samples, that is expressed as $s_k = {}^{n_k}/_{\sum_{i=1}^{K} n_i}$, where $\sum_{k=1}^{K} s_k = 1$. CBI is expressed as the inverse of the standard deviation of the s_k array:

$$CBI = \frac{1}{\sigma([s_1, s_2, \dots, s_K])}. \tag{9}$$

When the proportion of each category of annotated sample is more similar, the higher the CBI, i.e., the algorithm can select more minority samples in the unbalanced data.

In addition, in order to quantify the diversity of selected samples in a batch, we calculated the Euclidean distance and Kullback–Leibler (KL) divergence between the feature vectors of selected samples as the measure of diversity. Let $\boldsymbol{f}_i$ be the feature vector extracted from sample x_i, and d be the dimension of the feature vector. We calculated the Euclidean distance (ED) for each pair of feature vectors in a batch and calculated the average distance as a measure of diversity. For KL divergence-based diversity, the probability density for each feature vector is obtained by normalizing the features, denoted as ρ_i, where $\rho_{i,j}$ is the probability of the j-th feature on the vector $\boldsymbol{f}_i$. The ρ_{ref} is calculated from all the normalized feature vectors in a batch: $\rho_{ref} = {}^{1}/_{b} \sum_{i}^{b} \rho_i$. Then, the KL divergence between each sample's feature vector probability distribution and the reference distribution is calculated as:

$$KL(\rho_i || \rho_{ref}) = \sum_{j=1}^{d} \rho_{i,j} log\left(\frac{\rho_{i,j}}{\rho_{ref,j}}\right). \tag{10}$$

The KL divergence for each selected sample in a batch is represented as a list $[KL(\rho_1||\rho_{ref}), KL(\rho_2||\rho_{ref}), \dots, KL(\rho_b||\rho_{ref})]$. The average KL divergence of the list is measured as diversity.

For MNIST, FMNIST, and CIFRA10, we use the output of the penultimate layer of the pre-trained Resnet50 as the image feature extractor. For other datasets, we use the original sample directly as features, i.e., $x_i = \boldsymbol{f}_i$, to calculate the diversity.

4.2. Experimental Results

Table 1 presents the experimental results of the nine strategies, including Macro F1 score, CBI, and relative diversity, which were averaged across 10 repetitive experimental trials and the AL

Table 1
The CBI and Macro F1 of all strategies averaged over ten repetitions. The best CBI and F1 score achieved on each dataset are marked in bold.

		LETTER		MNIST		FMNIST		CIFAR10		CROP		E-DEVICE		ECG5000	
		CBI	F1	CBI	F1	CBI	F1	CBI	F1	CBI	F1	CBI	F1	CBI	F1
RS		2.79	.414	2.29	.811	2.29	.716	2.05	.531	2.70	.345	2.42	.616	2.44	.479
US	Baseline	3.05	.512	2.75	.868	3.45	.740	2.81	.598	4.41	.336	2.87	.632	2.95	.538
	CB	3.80	.525	2.88	.864	3.69	.742	3.02	.606	4.86	.365	**3.03**	.632	3.06	.530
	FBU	3.47	.597	2.68	.891	3.69	.742	2.93	.591	4.47	.361	2.85	.639	2.92	.528
	AL-FaMoUS	**4.33**	**.609**	**3.01**	**.895**	**3.74**	**.748**	**3.17**	**.607**	**5.49**	**.376**	**3.03**	**.643**	**3.12**	**.541**
BALD	Baseline	3.47	.500	2.71	.855	3.58	.735	2.61	.597	4.21	.337	2.91	.608	2.75	.528
	CB	3.74	.492	2.81	.855	3.72	.734	2.72	.610	4.51	.351	**3.42**	.618	2.94	.532
	FBU	3.50	.572	2.78	**.886**	3.61	.741	2.58	.593	4.21	.367	2.82	.620	2.77	.528
	AL-FaMoUS	**4.12**	**.591**	**3.13**	**.886**	**3.80**	**.746**	**2.90**	**.617**	**5.00**	**.370**	**3.42**	**.626**	**3.05**	**.537**

Table 2
The mean of relative diversity according to Euclidean distance (ED) and KL diversity compared with the Baseline. A positive relative diversity means the selection algorithm can select more diverse samples in a batch than the Baseline. The best relative diversities achieved on each dataset are marked in bold.

		LETTER		MNIST		FMNIST		CIFAR10		CROP		E-DEVICE		ECG5000	
		ED	KL 10^{-3}	ED	KL 10^{-2}	ED	KL 10^{-2}	ED	KL 10^{-4}	ED	KL 10^{-7}	ED	KL 10^{-6}	ED	KL 10^{-3}
US	CB	0.49	1.71	-0.08	2.60	0.63	0.42	0.93	5.02	0.07	5.31	0.06	-1.31	-0.34	-0.45
	FBU	1.49	2.00	8.24	3.33	2.80	2.04	1.09	9.01	0.10	7.22	0.20	1.13	3.28	7.74
	AL-FaMoUS	**1.58**	**2.70**	**8.85**	**3.63**	**3.43**	**2.11**	**4.26**	**1.52**	**0.13**	**1.06**	**0.21**	**1.13**	**4.78**	**1.63**
BALD	CB	0.27	1.05	1.49	0.49	-2.14	-1.01	0.27	4.01	0.03	3.22	0.07	0.73	-0.06	-0.12
	FBU	1.05	1.80	6.75	2.61	0.52	1.02	0.78	9.03	0.14	0.68	0.28	6.82	2.54	4.12
	AL-FaMoUS	**1.08**	**2.34**	**8.20**	**3.40**	**1.38**	**1.34**	**2.02**	**20.0**	**0.16**	**1.15**	**0.61**	**7.55**	**3.56**	**5.33**

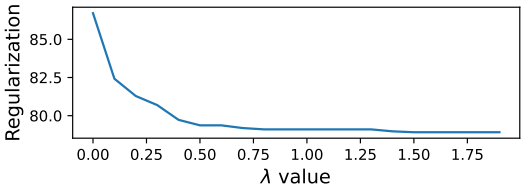
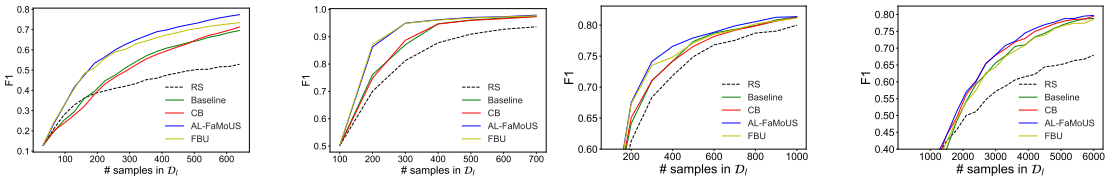


Figure 4: The relationship between the choice of λ and the regularization term in the optimization process of the US-based CB method applied on MNIST.

cycles. Overall, the results indicate that the proposed AL-FaMoUS outperformed the compared AL strategies on all datasets. The findings are summarized as follows:

AL-FaMoUS exhibited better performance. In comparison to other AL strategies, AL-FaMoUS exhibited better performance across all datasets in terms of Macro F1-score. These findings suggest that AL-FaMoUS effectively enhanced the identification of minor classes while concurrently ensuring that the performance of the major classes remains uncompromised. For example, on the LETTER dataset, AL-FaMoUS outperformed US and BALD baseline in terms of Macro F1 score, exhibiting a mean improvement by 9.7% and 9.1%, respectively.

CB did not create a benefit in all datasets. Using CB on relatively simple datasets such as



(a) F1 of BALD on LETTER (b) F1 of BALD on MNIST (c) F1 of BALD on FMNIST (d) F1 of BALD on CIFAR10

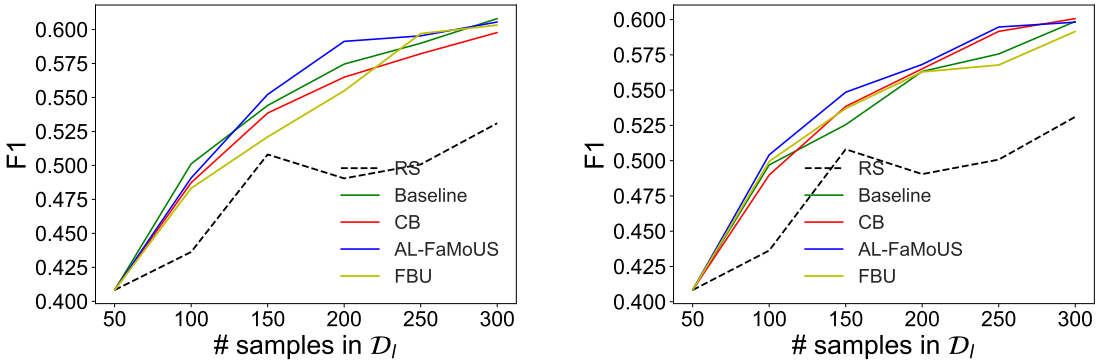
Figure 5: The Marco F1 curves of the RS strategy and the BALD-based AL strategies on the four image imbalanced datasets

MNIST and LETTER does not create a performance gain in all cases. As illustrated in Fig. 5b AL-FaMoUS and FBU obtained comparably excellent results on MNIST compared to CB, whereas CB even had a slightly lower F1 score than the baseline. One reason could be the well-structured dataset with relatively homogeneous learning patterns in the minor class. As shown in Table 1, although the CBI of CB is higher than that of FBU, i.e., CB selects more samples from minority classes. But the diversity of CB selection samples is lower than FBU. Table 4 shows that US-based CB even selected less diverse samples than the Baseline (relative ED diversity is −0.08). Fig. 5a presents a similar conclusion about LETTER.

Fast model updates could enhance sample diversity. The enhancement in performance can be attributed to the presence of sample diversity due to fast model updates. The experimental results presented in Table 4 show that FBU has positive relative diversity scores on all datasets, indicating that FBU can enhance the diversity of sample selection relative to the Baseline. Both FBU and AL-FaMoUS demonstrated favorable results compared to other strategies across different datasets, such as MNIST, LETTER, and Electric-Devices. Specifically, FBU outperformed the BALD baseline by up to 7.8% in terms of Macro F1, as is shown in Fig. 5a.

FBU did not improve the performance on complex datasets. The performance of the FBU method did not surpass that of the baseline approach in complex datasets. Specifically, we observed a decrease 0.7% in F1 score for the US-based FBU method on CIFAR10 compared to the US baseline. This result is consistent with the findings in [17]. In addition, [17] suggested that one promising solution could be an SNGP network consisting of multiple layers with Laplace approximation. Furthermore, Table 1 indicates the AL-FaMoUS strategy based on both acquisition functions of US and BALD achieved a higher mean F1 than the corresponding FBU. As shown in Fig. 5d, when F1 reaches 75%, compared with the AL-FaMoUS, the FBU method requires 12.8% more annotations. It may be explained by the FBU falling into the local optima due to the monotonous major-class samples selected according to the single selection criteria, i.e., the top-ranked utility scores.

Diverse performance on the extremely imbalanced dataset. The performance of different strategies varies on the extremely imbalanced dataset, particularly in the case of ECG5000, where the minor-class samples constitute only 1% of the major-class samples. The Macro F1 measure is particularly sensitive to the performance of the minor classes. Specifically, the F1 score of FBU is significantly lower than that of the US baselines, with a maximum difference of 3%. The CBI in Tab 1 indicates that the reason could be more and more major-class samples



(a) F1 of US on ECG5000

(b) F1 of BALD on ECG5000

Figure 6: The performance of compared methods on ECG5000. The blue line indicates our proposed method.

were selected over AL cycles, which kept making the dataset more imbalanced (CBI of US-based FBU lower than Baseline) and eventually resulted in the model being biased toward the major classes. We also noticed that the performance of CB is also lower than the US-based baseline in Figs. 6a. One possible reason is that in highly imbalanced datasets, model training is biased due to the lack of sample diversity in CB, which can be proved in Tab 4 that relative diversity is negative on ECG5000. The model ignored the minor-class samples completely so that it could not query the desired pseudo-label to perform class balance optimization. In contrast, Fig. 6 presents that AL-FaMoUS outperformed the other strategies regarding F1 score.

Detailed experimental results are given in Appendix B.

5. Conclusion and Future Plans

Our study focused on addressing two primary challenges appearing in AL: (1) the model trained on an imbalanced dataset ignores the significance of the minor-class in sample selection of the AL process and (2) the model that has learned a limited training set and is not timely updated or retrained can not select samples containing diverse, unknown learning patterns, leading to a loss of sample diversity. To tackle both challenges, we proposed the AL-FaMoUS, a general solution combining a class-balanced minibatch selection strategy and fast model updates to the AL process. Moreover, we implemented the single-sample-based AL-FaMoUS and evaluated it on seven public imbalanced datasets using BALD and US as baselines. As a result, the single-sample-based AL-FaMoUS outperformed the other existing AL strategies regarding macro F1 score and selected more diverse samples in the experiments. Besides, the experiments showed that the AL-FaMoUS can be applied to different architectures of BNNs, indicating the adaptability and flexibility of the AL-FaMoUS solution.

In future research, the current experimental setup can be extended from the following perspectives: (1) search the optimal parameters of SNGP at the initial stage of AL; (2) research on the impact of the budget per cycle b on the BNN's performance and the required computation source; (3) verify the performance impact under the different imbalance ratios of the dataset on

AL-FaMoUS; (4) dynamic adjustment of the regularization parameter λ in AL cycles, and (5) evaluate more state-of-the-art AL strategies as baselines, such as BatchBALD [16] or BADGE [46]. In addition, the research directions can also move forward to different practical application scenarios with considering other deep learning domains. Besides the class balance problem, for example, novel and/or anomalous classes can be detected and emphasized at the stage of sample selection by applying novelty/anomaly detection techniques. Also, the fast Bayesian update can be replaced by other updating strategies, such as various continual learning strategies.

Acknowledgments

This work is supported within the Digital-Twin-Solar (03EI6024E) project, funded by BMWi: Deutsches Bundesministerium für Wirtschaft und Energie/German Federal Ministry for Economic Affairs and Energy.

References

[1] U. Aggarwal, A. Popescu, C. Hudelot, Active learning for imbalanced datasets, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1428–1437.

[2] X. Cai, Active learning for imbalanced data: The difficulty and proportions of class matter, Wireless Communications and Mobile Computing 2022 (2022).

[3] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data 6 (2019) 1–54.

[4] A. Bria, C. Marrocco, F. Tortorella, Addressing class imbalance in deep learning for small lesion detection on medical images, Computers in biology and medicine 120 (2020) 103735.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[6] J. Z. Bengar, J. van de Weijer, L. L. Fuentes, B. Raducanu, Class-balanced active learning for image classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1536–1545.

[7] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, 2016, pp. 467–482.

[8] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering 21 (2009) 1263–1284.

[9] N. Japkowicz, The class imbalance problem: Significance and strategies, in: Proc. of the Int'l Conf. on artificial intelligence, volume 56, 2000, pp. 111–117.

[10] Y. Cao, T. Chen, Z. Wang, Y. Shen, Learning to optimize in swarms, Advances in Neural Information Processing Systems 32 (2019).

[11] C. Huang, Y. Li, C. C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5375–5384.

[12] T. Westmeier, D. Botache, M. Bieshaar, B. Sick, Generating synthetic time series for machine-learning-empowered monitoring of electric motor test benches, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2022, pp. 1–10.
[13] H. S. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Workshop on Computational Learning Theory, 1992, pp. 287–294.
[14] B. Settles, Active learning literature survey (2009).
[15] D. Kottke, M. Herde, C. Sandrock, D. Huseljic, G. Krempl, B. Sick, Toward optimal probabilistic active learning using a Bayesian approach, Machine Learning 110 (2021) 1199–1231.
[16] A. Kirsch, J. Van Amersfoort, Y. Gal, BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning, in: Advances in Neural Information Processing Systems, 2019.
[17] M. Herde, Z. Huang, D. Huseljic, D. Kottke, S. Vogt, B. Sick, Fast bayesian updates for deep learning with a use case in active learning, arXiv preprint arXiv:2210.06112 (2022).
[18] Y. Yang, G. Ma, et al., Ensemble-based active learning for class imbalance problem, Journal of Biomedical Science and Engineering 3 (2010) 1022.
[19] W. J. Park, An improved active learning in unbalanced data classification, in: Secure and Trust Computing, Data Management, and Applications: STA 2011 Workshops: IWCS 2011 and STAVE 2011, Loutraki, Greece, June 28-30, 2011. Proceedings 8, Springer, 2011, pp. 84–93.
[20] C. Lin, M. Mausam, D. Weld, Active learning with unbalanced classes and example-generation queries, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 6, 2018, pp. 98–107.
[21] H. Lei, S. Wang, D. Zheng, X. Qu, S. Fan, C. Cui, Improving active learning by data balance to reduce annotation efforts, The Journal of Engineering 2019 (2019) 8650–8653.
[22] I. Sundin, P. Schulam, E. Siivola, A. Vehtari, S. Saria, S. Kaski, Active learning for decision-making from imbalanced observational data, in: International conference on machine learning, PMLR, 2019, pp. 6046–6055.
[23] R. Zhang, A. A. Khan, R. L. Grossman, Y. Chen, Balance: deep bayesian active learning via equivalence class annealing, arXiv preprint arXiv:2112.13737 (2021).
[24] Y. Gal, R. Islam, Z. Ghahramani, Deep Bayesian active learning with image data, in: International Conference on Machine Learning, 2017, pp. 1183–1192.
[25] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences 114 (2017) 3521–3526.
[26] D. Maltoni, V. Lomonaco, Continuous learning in single-incremental-task scenarios, Neural Networks 116 (2019) 56–73.
[27] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: International conference on machine learning, PMLR, 2017, pp. 3987–3995.
[28] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, R. Hadsell, Progress & compress: A scalable framework for continual learning, in: International conference on machine learning, PMLR, 2018, pp. 4528–4537.
[29] H. Ritter, A. Botev, D. Barber, Online structured laplace approximations for overcoming

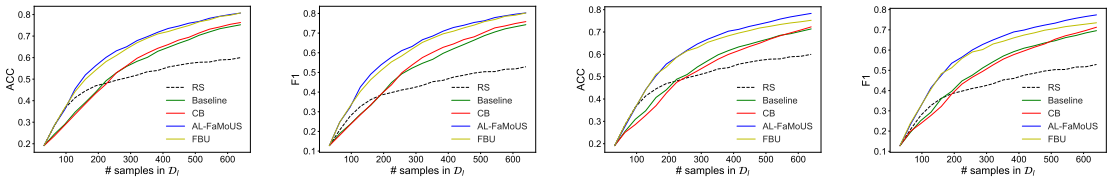
catastrophic forgetting, Advances in Neural Information Processing Systems 31 (2018).
[30] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, B. Lakshminarayanan, Simple and principled uncertainty estimation with deterministic deep learning via distance awareness, in: Advances in Neural Information Processing Systems, 2020.
[31] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations, 2018.
[32] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: Advances in Neural Information Processing Systems, 2007.
[33] Z. Lu, E. Ie, F. Sha, Mean-field approximation to Gaussian-softmax integral with application to uncertainty estimation, arXiv preprint arXiv:2006.07584 (2020).
[34] Y. LeCun, C. Cortes, The MNIST database of handwritten digits, 1998.
[35] P. W. Frey, D. J. Slate, Letter recognition using holland-style adaptive classifiers, Machine Learning 6 (1991) 161–182.
[36] H. Xiao, K. Rasul, R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, arXiv preprint arXiv:1708.07747 (2017).
[37] A. Krizhevsky, Learning multiple layers of features from tiny images, Master's thesis, University of Toronto, 2009.
[38] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, circulation 101 (2000) e215–e220.
[39] C. W. Tan, G. I. Webb, F. Petitjean, Indexing and classifying gigabytes of time series under time warping, in: Proceedings of the 2017 SIAM international conference on data mining, SIAM, 2017, pp. 282–290.
[40] J. Lines, A. Bagnall, P. Caiger-Smith, S. Anderson, Classification of household devices by electricity usage profiles, in: Intelligent Data Engineering and Automated Learning-IDEAL 2011: 12th International Conference, Norwich, UK, September 7-9, 2011. Proceedings 12, Springer, 2011, pp. 403–412.
[41] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9268–9277.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[43] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, Temporal convolutional networks for action segmentation and detection, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 156–165.
[44] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[45] J. Opitz, S. Burst, Macro f1 and macro f1, arXiv preprint arXiv:1911.03347 (2019).
[46] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds, in: International Conference on Learning Representations, 2020.

A. Detailed experiment parameter setting

Table 3
Experimental parameter setting for 4, imb.ratio indicates the ratio of the minor-class samples to the major-class samples. imb $\mathscr{D}_u$ size is the total number of samples of the imbalanced unlabeled dataset.

		LETTER	MNIST	FMINIST	CIFAR10	CROP	E-DEVICE	ECG5000
Dataset	type	image	image	image	image	time	time	time
	class num	26	10	10	10	24	7	5
	imb. ratio	10%	10%	10%	10%	10%	23%	1%
	imb $\mathscr{D}_u$ size	8274	32430	32981	27481	6668	8318	2500
	test dataset size	5000	10000	10000	10000	12000	8318	2500
SNGP	backbone	FCN	ResNet-6	ResNet-6	ResNet-18	TCN	TCN	TCN
	kernel size	8	1	5	5	2	2	1
Train	train batch	16	20	20	20	10	20	10
	optimizer	SGD	SGD	SGD	SGD	Adam	Adam	Adam
	learning rate	0.05	0.1	0.1	0.1	0.001	0.001	0.001
	momentum	0.9	0.9	0.9	0.9	-	-	-
	n_epochs	100	100	200	200	100	150	100
AL Setting	b_0	32	100	100	300	100	100	50
	b	32	100	100	300	100	100	50
	cycle	20	7	10	20	6	30	6
Class Balance	λ	0.1	0.5	0.1	0.5	0.5	0.05	0.05

B. Performance comparison on different datasets



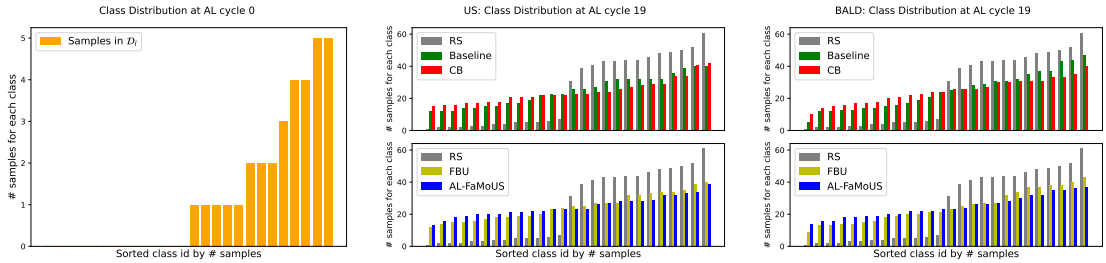
(a) ACC of US on LETTER (b) F1 of US on LETTER (c) ACC of BALD on LETTER (d) F1 of BALD on LETTER

Figure 7: The performance on LETTER dataset

LETTER This dataset consists of 26 classes. As illustrated in Fig. 7c, both FBU and AL-FaMoUS demonstrate significantly better performance in comparison to CB and their respective baselines. For example, we noted that an additional annotation of approximately 50% is required to achieve 70% accuracy when using the baseline method compared to AL-FaMoUS. Moreover, in terms of Macro F1 score, AL-FaMoUS outperforms FBU, exhibiting a mean improvement of 1.2% and 1.5% for the US and BALD, respectively.

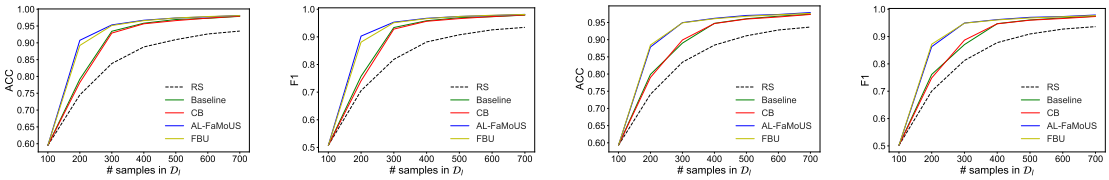
Table 4
The accuracy and Macro F1 of all strategies averaged over ten repetitions. The best F1 score and accuracy achieved on each dataset are marked in bold.

		LETTER		MNIST		FMNIST		CIFAR10		CROP		E-DEVICE		ECG5000	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
RS		.494	.414	.834	.811	.722	.716	.580	.531	.440	.345	.688	.616	.926	.479
US	Baseline	.547	.512	.886	.868	.747	.740	.630	.598	.423	.336	.696	.632	.936	.538
	CB	.552	.525	.883	.864	.750	.742	**.639**	.606	.446	.365	.695	.632	.935	.530
	FBU	.618	.597	.905	.891	.748	.742	.625	.591	.440	.361	.702	.639	.936	.528
	AL-FaMoUS	**.627**	**.609**	**.908**	**.895**	**.755**	**.748**	**.639**	**.607**	**.455**	**.376**	**.704**	**.643**	**.938**	**.541**
BALD	Baseline	.534	.500	.876	.855	.741	.735	.631	.597	.425	.337	.684	.608	.934	.528
	CB	.523	.492	.875	.855	.741	.734	.641	.610	.435	.351	.686	.618	.935	.532
	FBU	.600	.572	**.901**	**.886**	.747	.741	.630	.593	**.450**	.367	.693	.620	.935	.528
	AL-FaMoUS	**.612**	**.591**	**.901**	**.886**	**.751**	**.746**	**.647**	**.617**	.449	**.370**	**.695**	**.626**	**.937**	**.537**



(a) Initial imbalanced $\mathcal{D}_l$ (b) US based $\mathcal{D}_l$ on LETTER (c) BALD based $\mathcal{D}_l$ on LETTER

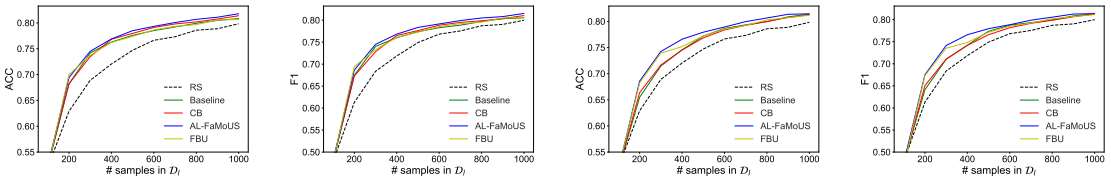
Figure 8: The class distribution on LETTER dataset



(a) ACC of US on MNIST (b) F1 score of US on MNIST (c) ACC of BALD on MNIST (d) F1 of BALD on MNIST

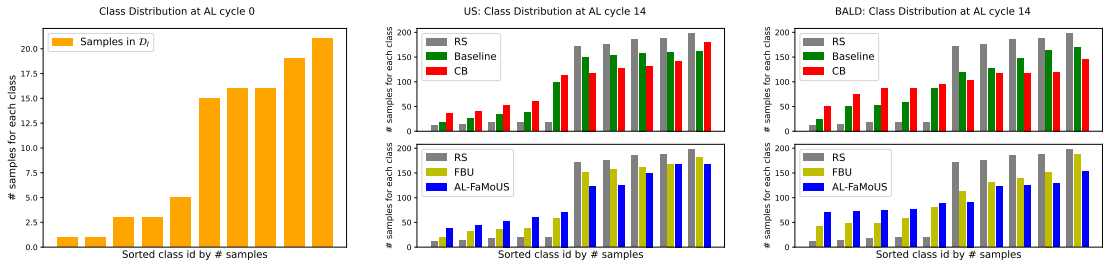
Figure 9: The performance on MNIST dataset

MNIST A similar finding was also observed in the experiments on MNIST, as shown in Fig 9d. For total budgets below 400, both FBU and AL-FaMoUS maintained a considerable advantage, underscoring the significance of sample diversity in the selection process. Besides, BC and the BALD baseline achieved comparable results. The main reason could be that MNIST is a well-structured dataset, where the existing learning patterns could be relatively tedious. Therefore, using a class balance selection strategy in AL cycles can not create a significant benefit. One possible reason is that the dataset is relatively simple, as we observed in the Fig 9d that even with fewer samples in the minority class, the accuracy on test dataset can reach 95% rapidly.



(a) ACC of US on FMNIST (b) F1 of US on FMNIST (c) ACC of BALD on FMNIST (d) F1 of BALD on FMNIST

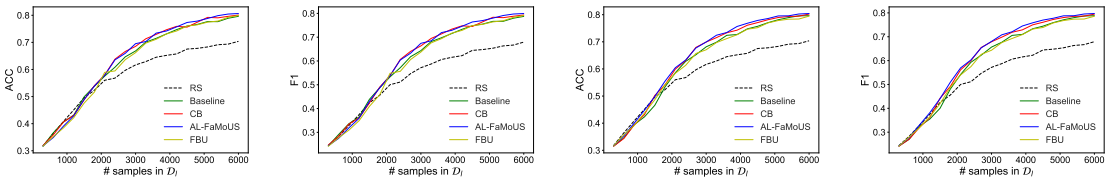
Figure 10: The performance on FMNIST dataset



(a) Initial imbalanced $\mathcal{D}_l$ (b) US based $\mathcal{D}_l$ on FMNIST (c) BALD based $\mathcal{D}_l$ on FMNIST

Figure 11: The class distribution on FMNIST dataset

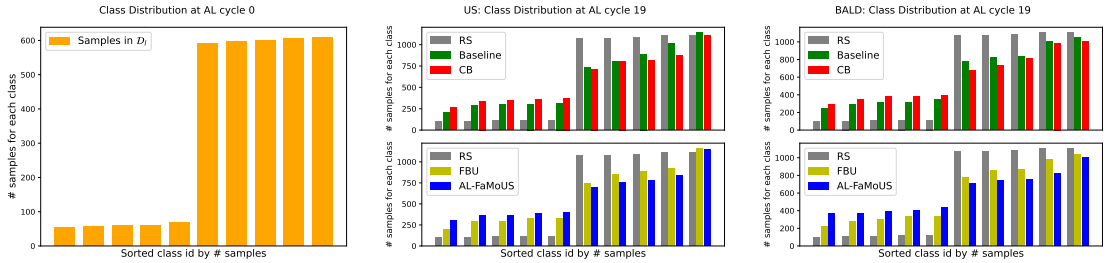
FMNIST According to the results in Table 1, we observed the US-based FBU and CB performed equally well and better than the US baseline by about 0.2% of the mean F1 score. By comparison, the mean accuracy of the BALD baseline and the corresponding CB dropped by up to 0.9%. It suggests the selection of AL strategies had a considerable influence on this dataset. Furthermore, AL-FaMoUS improved the performance of both US and BALD AL strategies in terms of accuracy and F1 score. It is noteworthy that AL-FaMoUS with BALD achieved an improvement by 1.5% and 2.7% in terms of accuracy compared to FBU and CB, respectively, as shown in Fig. 10c.



(a) ACC of US on CIFAR10 (b) F1 of US on CIFAR10 (c) ACC of BALD on CIFAR10 (d) F1 of BALD on CIFAR10

Figure 12: The performance on CIFAR10 dataset

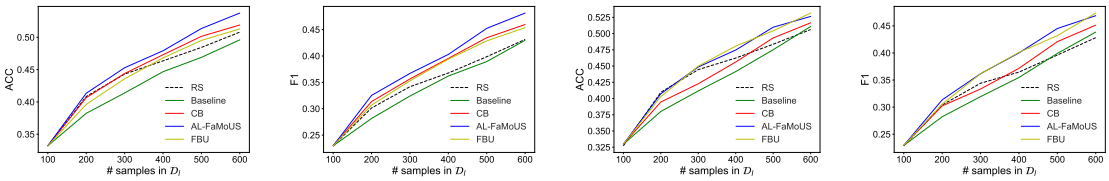
CIFRA10 In Fig. 12c, RS outperformed the others when the total budget was below 1500, which indicates the learning patterns in this dataset may be more complex and challenging for



(a) Initial imbalanced $\mathscr{D}_l$ (b) US based $\mathscr{D}_l$ on CIFAR10 (c) BALD based $\mathscr{D}_l$ on CIFAR10

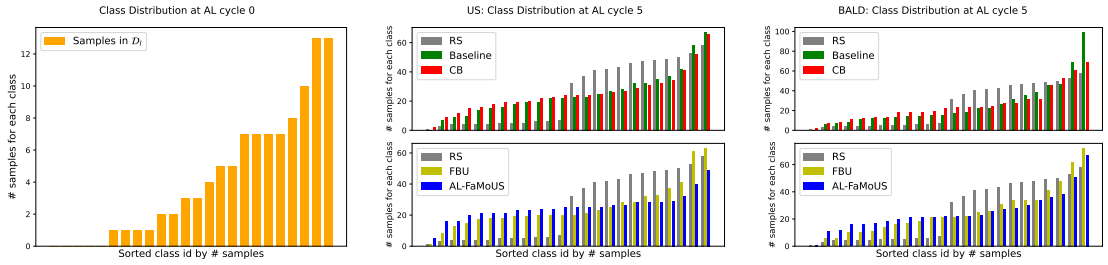
Figure 13: The class distribution on CIFAR10 dataset

training. We especially noted that the accuracy and F1 score of the US-based FBU dropped by 0.5% and 0.7% compared to the US baseline. This result is consistent with the findings in [17], which suggested that one promising solution could be an SNGP network consisting of multiple layers with Laplace approximation. Furthermore, Table 1 indicates the AL-FaMoUS strategy based on both US and BALD achieved a higher mean accuracy than the corresponding FBU. It may be explained by the FBU falling into the local optima due to the monotonous major-class samples selected according to the single selection criteria, i.e., the top-ranked utility scores. However, the regularization term for the class balance weighted the minor-class samples, leading to the best performance of AL-FaMoUS.



(a) ACC of US on CROP (b) F1 of US on CROP (c) ACC of BALD (d) F1 of BALD on CROP

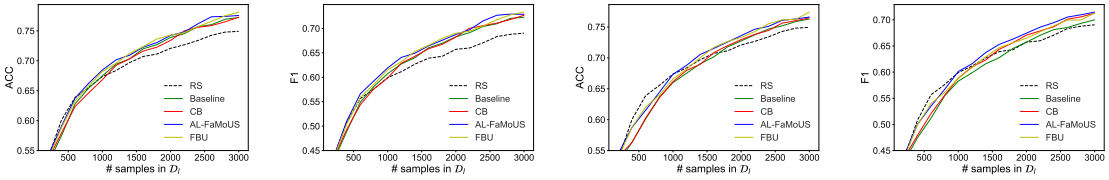
Figure 14: The performance on CROP dataset



(a) Initial imbalanced $\mathscr{D}_l$ (b) US based $\mathscr{D}_l$ on CROP (c) BALD based $\mathscr{D}_l$ on CROP

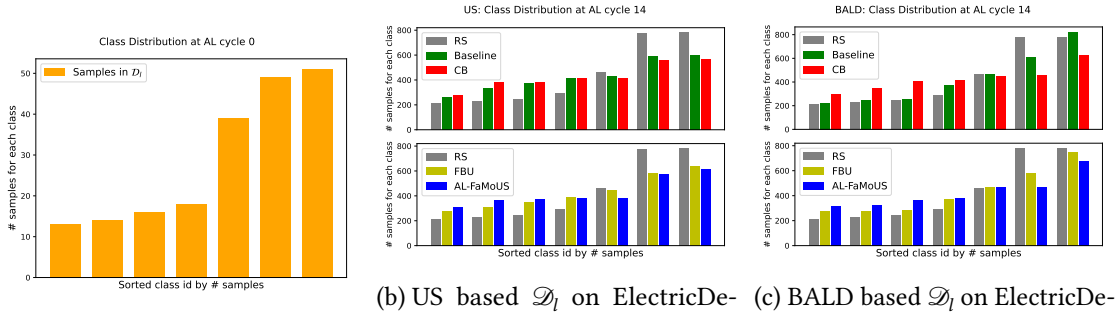
Figure 15: The class distribution on CROP dataset

CROP This dataset consists of 24 classes. Despite a lack of consideration for class balance, the US and the BALD baseline were inferior to RS, as listed in Table 1. This observation highlights the crucial role played by sample diversity. The performances of FBU and CB vary depending on the AL strategy chosen for the baseline, but AL-FaMoUS maintains the optimal results.



(a) ACC of US on E-Dev (b) F1 of US on E-Dev (c) ACC of BALD (d) F1 of BALD

Figure 16: The performance on ElectricDevices dataset

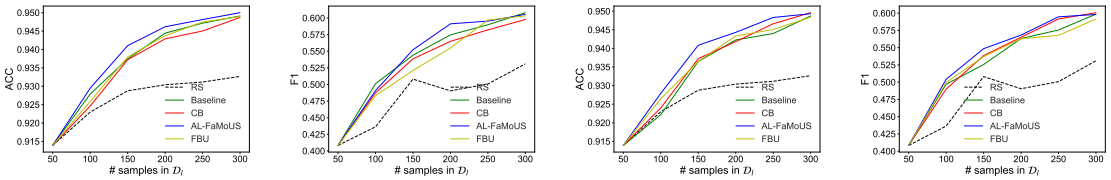


(a) Initial imbalanced $\mathcal{D}_l$ (b) US based $\mathcal{D}_l$ on ElectricDevices (c) BALD based $\mathcal{D}_l$ on ElectricDevices

Figure 17: The class distribution on ElectricDevices dataset

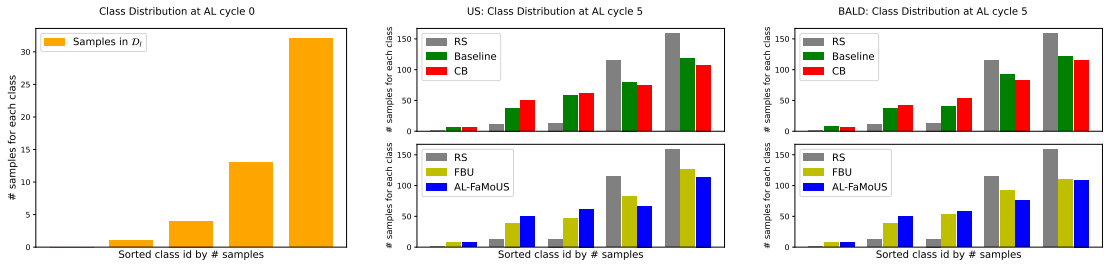
Electric Devices This dataset comprises 7 classes, with the minority class samples accounting for approximately 23% of the majority class samples. As shown in Table 1, BC and the US baseline perform similarly. Meanwhile, AL-FaMoUS achieves an average F1 improvement of 1.1% and 0.4% compared to the US baseline and FBU, respectively. The mean accuracy of all methods based on BALD decreases by 1.2% relative to the US baseline. Nevertheless, AL-FaMoUS consistently demonstrates the best performance. As Figure 16d is shown, when F1 reaches 65%, compared with the BALD baseline, the AL-FaMoUS method only needs to label 70% of the samples.

ECG5000 This dataset exhibits extreme class imbalance, with the minority class samples representing only 1% of the majority class samples. Since Macro F1 is sensitive to minority class performance, the curve of RS in Figure 18b, rises first and then falls. As the AL cycle increases, more and more samples of the majority class are selected, resulting in the model being biased toward the majority class, leading to poor performance of the minority class. Notably, FBU's F1 score is significantly lower than the US baselines, with a maximum difference of 3%.



(a) ACC of US on ECG5000 (b) F1 of US on ECG5000 (c) ACC of BALD on ECG5000 (d) F1 of BALD on ECG5000

Figure 18: The performance on ECG5000 dataset



(a) Initial imbalanced $\mathscr{D}_l$ (b) US based $\mathscr{D}_l$ on ECG5000 (c) BALD based $\mathscr{D}_l$ on ECG5000

Figure 19: The class distribution on CROP dataset

The experimental results indicate that this method overlooks class balance, leading to poor classification performance in the minority class. In contrast, AL-FaMoUS achieves the best results, with the highest F1 improvement of 6% compared to FBU. Additionally, Figures 18a and 18c show that AL-FaMoUS attains optimal accuracy in most AL cycles. The experimental results demonstrate that our proposed method not only enhances the recognition accuracy of the minority class through class balance but also maintains the performance of the majority class recognition.