# FAUSAL: Fluctuation Aware Uncertainty Sampling for Active Learning on Class-Imbalanced Medical Images

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Medical image classification from imbalanced datasets remains challenging due to labeled data scarcity and annotation costs, limiting deep learning models. Active learning (AL) addresses this by strategically selecting informative samples for annotation, achieving optimal performance at minimal cost. However, traditional uncertainty sampling methods often overlook minority classes and persistently ambiguous cases both critical for model performance. We propose Fluctuation-Aware Uncertainty Sampling for Active Learning (FAUSAL), a sampling strategy extending instantaneous predictive uncertainty by incorporating temporal fluctuations in predictions across AL iterations. FAUSAL measures per-sample prediction instability over successive model updates, prioritizing samples exhibiting persistent ambiguity, particularly from underrepresented classes and decision boundary regions. We validate FAUSAL on two imbalanced medical imaging benchmarks: HAM-10000 (skin lesion classification) and Paludisme-data (malaria development stage recognition). FAUSAL outperforms standard training and state-of-the-art AL baselines, achieving respectively 92% and 97% of fully supervised performance using only 22.67% and 4.81% of available training data. Notably, minority class recognition substantially improves at low annotation budgets, demonstrating superior class imbalance handling. Ablation studies confirm temporal fluctuation's critical contribution alongside instantaneous uncertainty. Thus, FAUSAL establishes itself as a robust AL strategy for annotation efficient medical image classification.

## 1 Introduction

In recent years, Deep Neural Networks (DNNs) have driven spectacular progress in medical image analysis. However, when it comes to supervised classification, performance and reliability still depend on the amount and quality of labeled data, while expert annotation remains expensive and time-consuming. This is particularly consequential in clinical settings, where underrepresented classes often correspond to the most dangerous conditions such as rare parasitic stages in malaria or melanoma among skin lesions. Furthermore, medical datasets often exhibit strong class imbalance: rare but clinically critical cases are underrepresented. Adding more majority-class labels brings little gain, while minority-class performance keeps dropping. Clinically, this means increased risk of missed diagnoses for the very cases requiring urgent intervention. Consequently, building reliable classifiers under a limited annotation budget requires not only more labeled data, but more carefully selected labeled data. Active learning (AL) provides a principled framework to reduce labeling effort by iteratively sampling a human annotator (*e.g.* a clinician), often simulated by an oracle agent, for the most informative unlabeled samples (Li et al., 2024a). Active learning was first formalized in early work on selective sampling and concept learning in the late (Angluin, 1988). Building on this foundation, later work established several canonical active learning frameworks. Most of these studies use pool-based AL, in which a learner selects labels from a large pool of unlabeled data, together with sampling strategies such as uncertainty-based sampling, query-by-committee, variance reduction, and expected model change (Settles, 2009; Li et al., 2024a). As illustrated in Fig. 1, in the pool-based AL setting we start from a small labeled set and a large unlabeled pool: the model is iteratively re-trained or fine-tuned on available labeled data, a sampling strategy prioritizes unlabeled samples within a fixed annotation budget then annotated, the newly labeled data is added to the training set and the process is repeated. The AL iterations continue until the

budget is exhausted, aiming to maximize performance per queried label (Kirsch et al., 2019). The effectiveness of AL is largely determined by the sampling (a.k.a query, acquisition) strategy. Canonical approaches fall into three families: (i) uncertainty-based strategies that prioritize samples exhibiting high predictive uncertainty (Houlsby et al., 2011; Gal et al., 2017; Li et al., 2024b); (ii) diversity-based strategies that promote coverage of the feature space (Sener & Savarese, 2018); and (iii) hybrid strategies that balance uncertainty and coverage (Bharadwaj et al., 2025). Random sampling is typically used as a reference baseline and can remain surprisingly competitive across multiple settings and especially on balanced data (Stolte et al., 2025; Li et al., 2024a; Nath et al., 2020). However, in strongly imbalanced settings, these strategies often exhibit increasingly unstable behavior as the imbalance severity intensifies, as both uncertainty estimation and feature extraction may be biased towards the majority-class(es). Moreover, most sampling strategies rely on a single snapshot of model outputs taken at a single AL iteration, and ignore how predictions evolve across successive AL iterations, potentially missing samples whose informational value is linked to the model's unstability when it comes to their prediction (depending on the training set content). To address these limitations, we propose FAUSAL, a fluctuation-aware uncertainty sampling strategy tailored for pool-based active learning of imbalanced image classification under scarce labels. Thus, FAUSAL is particularly adapted to medical image classification, but it can be used for any problem showcasing similar characteristics, especially when data annotation is made by experts, thus costly. In practice, this means steering annotation toward samples most likely to improve the detection of clinically critical minority classes. FAUSAL's core idea is to exploit the temporal prediction fluctuation of an unlabeled image across two successive AL iterations: after training at AL iteration $t-1$ and again at iteration $t$, the model yields two predictive distributions for the same sample, and FAUSAL quantifies their dissimilarity measure to identify highly unstable (thus potentially high-impact) candidates.

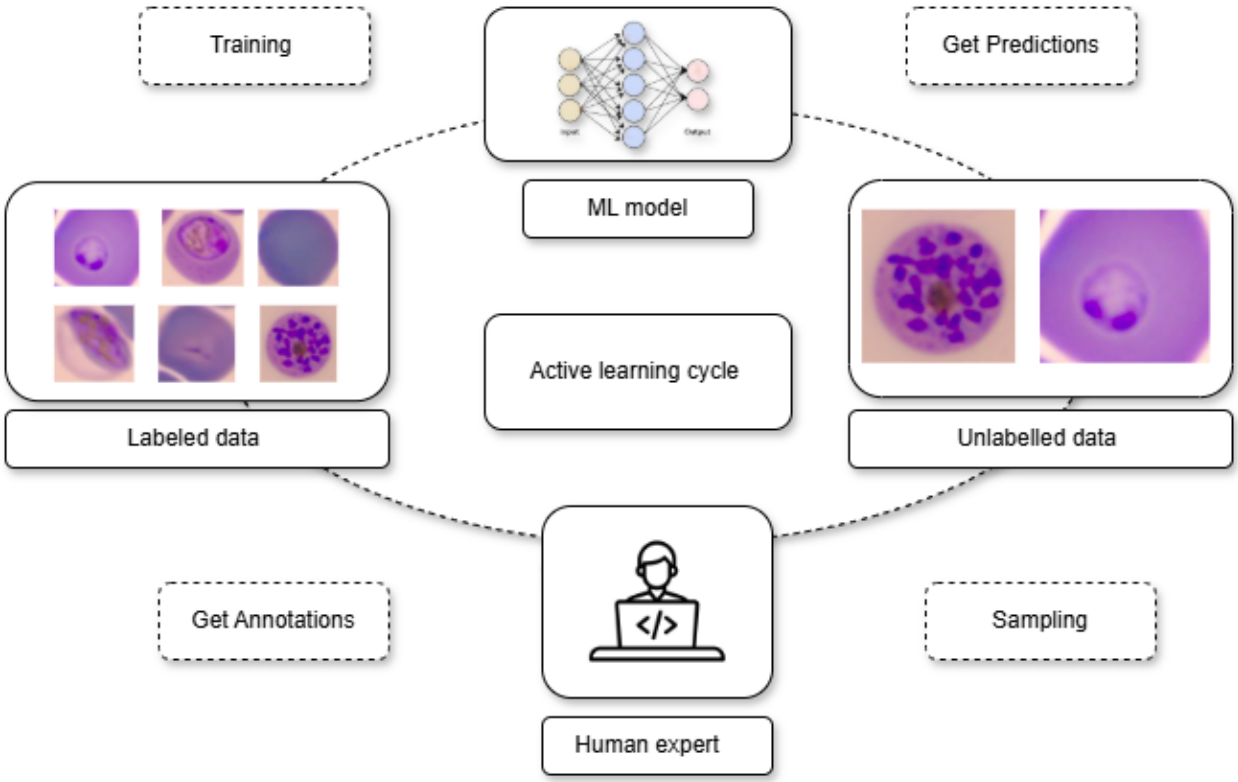



Figure 1: Pool-based active learning under class imbalance: train on the labeled set, score the unlabeled pool, sampling a budgeted batch, obtain expert labels, and update the model iteratively.

Indeed, our working hypothesis is that samples whose prediction fluctuate a lot between two AL iterations are samples that are either poorly represented by the model, or close to class boundaries. In other words, we

hypothesize that samples with strongly fluctuating predictions correspond to genuinely difficult or ambiguous cases whose annotation can induce meaningful decision-boundary updates under limited supervision. Importantly, informativeness alone is insufficient under strong class imbalance: naive sampling can repeatedly favor majority-class patterns and further starve minority categories. Therefore, by prioritizing highly fluctuating samples, FAUSAL naturally promotes the selection of hard and underrepresented cases, improving minority exposure while retaining highly informational content.

The main contributions of this work are the proposal of:

- **FAUSAL framework:** a fluctuation-aware active learning sampling strategy robust to varying degrees of class imbalance.
- **L1HEADS measure:** a new head-switching fluctuation score tracking prediction changes of the same unlabeled samples across consecutive AL iterations.
- **Empirical validation and practical flexibility:** extensive experiments demonstrate improved performance with fewer expert annotations, with demonstrated clinical relevance, under a sampling design that adapts to dataset characteristics and remains effective across different imbalance severity levels.

## 2 Related Work

Deep learning has become a powerful tool in medical image classification (Yao et al., 2022), achieving strong performance through transfer learning or full training (Nath et al., 2020). However, when supervised, its success relies on large, well-annotated datasets, which are difficult and costly to obtain in medical contexts due to the need for expert labeling. Moreover, medical datasets are often imbalanced (Yao et al., 2022) (as serious pathologies are less frequent than benign conditions), which may reduce the generalization capacity of machine learning methods. Although data augmentation and generation techniques help alleviate this issue (Wang & Zhang, 2025; Shetty et al., 2023), their impact remains limited when the data is too scarce for some class(es). To mitigate these limitations, active learning (AL) has emerged as a promising approach that optimizes the annotation process by automatically selecting the most informative unlabeled samples for expert labeling (Li et al., 2022; Bai et al., 2025). AL aims to improve data efficiency and reduce annotation costs while maintaining model performance, making it particularly suited to medical imaging applications where labeled data are scarce and costly to produce (Li et al., 2024a). In the following, we organize prior work along (i) sampling strategies used in pool-based AL (see Fig. 1) and (ii) how class imbalance affects acquisition behavior and downstream generalization across AL iterations (Wang et al., 2024).

### 2.1 Sampling Strategies in Active Learning

Most pool-based AL methods can be grouped into uncertainty-based, diversity-based, and hybrid strategies (Gal et al., 2017; Sener & Savarese, 2018; Bharadwaj et al., 2025; Cohn et al., 1996). While these approaches are effective in balanced or mildly skewed settings, they exhibit several limitations under severe imbalance (Yu et al., 2015) (see paragraph 2.2).

#### 2.1.1 Uncertainty-based sampling

Uncertainty-based methods sample instances exhibiting the highest predictive uncertainty, using criteria such as predictive entropy (Nguyen et al., 2022), mutual information, or related uncertainty surrogates (Li et al., 2024a; Houlsby et al., 2011; Gal et al., 2017). Despite their simplicity and strong empirical performance in many benchmarks, they can suffer from redundancy in batch-mode selection and from poor calibration or overconfidence in deep networks (Gal et al., 2017; Huang et al., 2014; Nguyen et al., 2022). Consequently, in strongly imbalanced datasets, uncertainty can become biased toward regions dominated by majority-classes, leading to under-exploration of rare but clinically important classes.

#### 2.1.2 Diversity-based sampling

Diversity-based approaches aim to select a subset that covers the unlabeled pool in the feature space, using clustering or core-set selection to maximize representativeness (Sener & Savarese, 2018; Brinker, 2003). These methods can reduce redundancy and improve coverage; however, they are often computationally expensive and highly sensitive to representation quality. Furthermore, geometric diversity does not necessarily align with decision-boundary informativeness, and under severe imbalance such strategies may still allocate substantial budget to majority clusters, limiting minority gains.

#### 2.1.3 Hybrid sampling methods

Hybrid methods combine uncertainty with diversity or representativeness to improve batch efficiency (Bharadwaj et al., 2025). For instance, uncertainty-diversity pipelines typically shortlist uncertain candidates (*e.g.*, via probability gaps) and then apply a diversity step such as farthest-first selection in feature space. Relatedly, Variational adversarial active learning (VAAL) couples a variational autoencoder with an adversarial discriminator that distinguishes labeled from unlabeled data, prioritizing unlabeled samples that best confuse the discriminator as representative queries (Sinha et al., 2019; Zhong et al., 2025). Although hybrids aim to balance uncertainty and diversity, they inherit the limitations of their components: uncertainty prefiltering remains prone to miscalibration under imbalance, while diversity/representativeness steps depend strongly on feature quality and may scale poorly; consequently, rare classes can still be under-sampled when the uncertainty stage dominates (Bharadwaj et al., 2025).

#### 2.1.4 Random sampling

Random sampling provides a low-cost baseline that does not depend on model uncertainty or representations (Li et al., 2024a). While it offers a reference with respect to the unlabeled pool distribution, in strongly imbalanced datasets it typically reflects the underlying class skew and may therefore provide limited minority exposure (Nath et al., 2020).

### 2.2 Limitations of existing AL methods

Class imbalance fundamentally changes the behaviour of pool-based AL because the unlabeled pool is dominated by majority-class samples, and sampling strategies are evaluated on this skewed distribution (Bengar et al., 2022; Wang et al., 2024). In balanced settings, each class has a comparable opportunity to appear in sampled batches; in contrast, under strong imbalance, many sampling strategies tend to over-select majority-classes and neglect minority regions that are crucial for macro-level performance. Consequently, imbalance-aware AL commonly aims to (i) counteract majority bias during sampling and (ii) ensure repeated exposure to minority classes early enough for stable decision-boundary formation. In medical imaging, this is particularly important because minority categories often correspond to clinically critical findings. Overall, prior work has established effective uncertainty-based, diversity-based, and hybrid-based sampling strategies for AL (Nguyen et al., 2022; Sener & Savarese, 2018; Bharadwaj et al., 2025), but recent studies emphasize that severe imbalance can systematically bias their sampling strategy toward majority-classes, harming minority performance (Wang et al., 2024). However, across these families, sampling is typically driven by quantities computed from the current model state at a given AL iteration (*e.g.*, uncertainty or feature-space representativeness), without explicitly accounting for how the prediction of the same unlabeled sample changes between successive AL iterations (Li et al., 2024a; Gal et al., 2017; Sener & Savarese, 2018; Bharadwaj et al., 2025). To the best of our knowledge, existing AL sampling strategies do not use temporal fluctuation across successive AL iterations as a sampling criterion; instead, they rely on within AL iteration scores computed at the current AL iteration (Li et al., 2024a; Gal et al., 2017; Sener & Savarese, 2018; Bharadwaj et al., 2025). Consequently, despite their advanced uncertainty estimation and batch sampling strategies, existing active learning methods struggle with model instability and prediction fluctuations.

In summary, two limitations remain central for imbalanced classification: (i) sampling is often driven by single AL iteration snapshots that ignore how predictions change across successive AL iterations, and (ii) selection

mechanisms may not sufficiently regulate minority exposure under strongly imbalanced pools. Motivated by these gaps, we introduce FAUSAL, which leverages prediction fluctuation between successive AL iterations to improve annotation efficiency and minority-sensitive generalization.

## 3 Proposed Methodology

In this section, we present the foundations of Fluctuation-Aware Uncertainty Sampling for Active Learning (FAUSAL), a pool-based sampling strategy for strongly imbalanced image classification. The key idea is to move beyond a single AL iteration view of model outputs by exploiting two complementary cues: (i) the instantaneous uncertainty of the current classifier, and (ii) the temporal fluctuation of each unlabeled sample prediction, between two successive AL iterations. The uncertainty term identifies ambiguous samples near the current decision boundary, while the fluctuation term identifies samples whose predicted distributions are still changing after the most recent model update. By combining these two clues, FAUSAL prioritizes images that are both difficult and unstable, making them especially informative under strong class imbalance. In such settings, minority cases are often harder to classify and their decision regions tend to evolve more slowly.

### 3.1 Notation and active learning setting

We consider a pool-based active learning loop with a labeled set $\mathcal{D}_L$ and an unlabeled pool $\mathcal{D}_U$. At iteration $t$, the classifier is denoted by $f_{\theta^{(t)}}$ and is fine-tuned on the current labeled set $\mathcal{D}_L^{(t)}$, where $\theta^{(t)}$ denotes the model parameters at AL iteration $t$. For any unlabeled image $x \in \mathcal{D}_U^{(t)}$, the model outputs the predictive probability vector

$$\begin{aligned} p^{(t)}(x) &= \big(p_1^{(t)}(x), \dots, p_C^{(t)}(x)\big), \\ p_c^{(t)}(x) &= \Pr\big(\hat{y} = c \mid x, \theta^{(t)}\big), \qquad \sum_{c=1}^{C} p_c^{(t)}(x) = 1 \end{aligned} \tag{1}$$

where $\mathcal{Y} = \{1, \dots, C\}$ denotes the label space of $C$ classes, and $\hat{y} \in \mathcal{Y}$ is the predicted label for an input $x$. The sampling budget is denoted by $b$ and the sampled set at AL iteration $t$ is a batch of size $b$, denoted by $\mathcal{S}_t \subset \mathcal{D}_U^{(t)}$. After collecting the labels of $\mathcal{S}_t$ from the dataset expert, the sets are updated as

$$\mathcal{D}_L^{(t+1)} = \mathcal{D}_L^{(t)} \cup \mathcal{S}_t \quad \text{and} \quad \mathcal{D}_U^{(t+1)} = \mathcal{D}_U^{(t)} \setminus \mathcal{S}_t.$$

Importantly, FAUSAL compares predictions only between two successive AL iterations $(t-1, t)$, which reflects the most recent belief update of the classifier and provides a computationally efficient fluctuation signal. Moreover, in the early AL iterations, the model predictions can change abruptly due to the scarcity of labelled data, and a larger fluctuation window would let the model be affected by this unstability for longer; using only two consecutive AL iterations helps keep the fluctuation estimate meaningful as the model gradually stabilizes.

### 3.2 FAUSAL framework

FAUSAL assigns to each unlabeled sample $x$ a unified score that linearly combines a normalized uncertainty term and a normalized fluctuation term:

$$S_{\text{FAUSAL}}^{(t)}(x) = \lambda\, E^{(t)}(x) + \big(1 - \lambda\big)\, F^{(t)}(x), \quad \lambda \in [0, 1]. \tag{2}$$

This convex linear combination provides an interpretable scalarization of two complementary objectives: $\lambda$ controls the trade-off between selecting samples that are uncertain at the current AL iteration and samples that are unstable between two successive AL iterations. Because Eq. equation 2 is a linear combination, both terms are normalized to the same range to avoid scale dominance and to keep $\lambda$ meaningful and stable across datasets and imbalance severity levels.

#### 3.2.1 Uncertainty component

At AL iteration $t$, uncertainty is measured from the current predictive distribution $p^{(t)}(x)$. Predictive entropy captures instantaneous ambiguity: if $p^{(t)}(x)$ is close to uniform, the model has no clear preference among classes, and the label of $x$ is potentially informative for refining the decision boundary. Concretely, we use the normalized predictive entropy:

$$\begin{aligned} E^{(t)}(x) &= \frac{H\big(p^{(t)}(x)\big)}{\log C} \\ &= -\frac{1}{\log C}\sum_{c=1}^{C} p_c^{(t)}(x)\,\log\big(p_c^{(t)}(x)+\varepsilon\big), \qquad \varepsilon > 0, \end{aligned} \tag{3}$$

so that $E^{(t)}(x) \in [0, 1]$.

#### 3.2.2 Fluctuation component

Uncertainty alone is computed from a single AL iteration snapshot and does not describe how the model reacts to learning. In contrast, fluctuation measures how much the prediction of the same unlabeled sample changes after the model update from AL iteration $t-1$ to AL iteration $t$. A large fluctuation indicates that the last training step materially changed the model's belief about $x$, which suggests that $x$ lies in a difficult region where the classifier remains unstable. Under strong class imbalance, such instability can be pronounced for underrepresented classes, making fluctuation a complementary cue to uncertainty for improving label efficiency. We quantify temporal fluctuation by comparing the two predictive distributions across two successive AL iterations using a generic dissimilarity measure $d(\cdot, \cdot)$ between probability vectors:

$$\Delta^{(t)}(x) = d\Big(p^{(t)}(x),\, p^{(t-1)}(x)\Big)\,. \tag{4}$$

To make fluctuation compatible with Eq. equation 2, we use the normalized fluctuation term:

$$F^{(t)}(x) = \frac{\Delta^{(t)}(x)}{M_d}, \tag{5}$$

where $M_d > 0$ is a dissimilarity measure dependent normalization constant chosen so that $F^{(t)}(x) \in [0, 1]$. The choice of $d(\cdot, \cdot)$ (and the associated $M_d$) defines different FAUSAL instantiations; we keep the formulation unified here and detail the tested dissimilarity measures and their empirical study in Section 4.

### 3.3 FAUSAL algorithm

This subsection presents the structure of the FAUSAL sampling procedure as summarized in Algorithm 1. It describes the computational steps performed at each active-learning AL iteration, from updating the model's predictions to selecting the next batch of instances. The goal is to clarify how the algorithm organizes its loop, score computation, and data updates before stating the full algorithm.

This section presented the FAUSAL concept, clarified the role of each term, and introduced the unified normalized formulation and the algorithmic workflow. In the next section, we specify the tested dissimilarity measures $d(\cdot, \cdot)$, the corresponding normalization constants $M_d$, and we provide an empirical study of how these fluctuation instantiations (depending on the dissimilarity measure) behave across strongly imbalanced medical datasets.

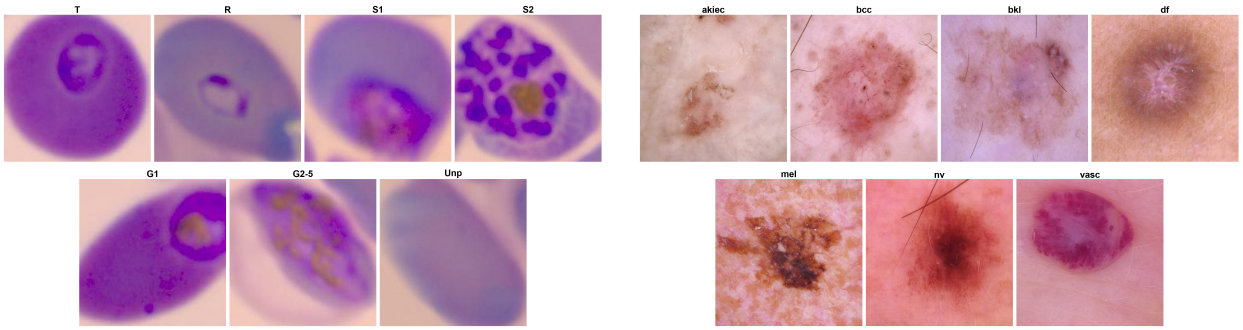
**Algorithm 1** FAUSAL: Fluctuation-Aware Uncertainty Sampling for Active Learning

**Input:** Labeled set $\mathcal{D}_L$, unlabeled pool $\mathcal{D}_U$, batch size $b$, fluctuation variant $F$
**Initialize:** Train $f(x;\theta)$ on $\mathcal{D}_L$; set $p^{(0)}(x)$ uniformly for all $x \in \mathcal{D}_U$.

1: **repeat**
2: Compute $p^{(t)}(x)$ for all $x \in \mathcal{D}_U$
3: Compute $E^{(t)}(x)$ and $F^{(t)}(x)$
4: $S^{(t)}_{\text{FAUSAL}}(x) \leftarrow \lambda\, E^{(t)}(x) + (1-\lambda)\, F^{(t)}(x)$
5: Select batch $\mathcal{S}_t \subset \mathcal{D}_U$ of size $b$ with the largest $S^{(t)}_{\text{FAUSAL}}(x)$ values
6: Sample instances for $\mathcal{S}_t$; update $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{S}_t$
7: Fine-tuning $f(x;\theta)$ on updated $\mathcal{D}_L$
8: Set $p^{(t-1)}(x) \leftarrow p^{(t)}(x)$
9: **until** annotation budget is exhausted

## 4 Implementation and Experimental Setup

### 4.1 Datasets




(a) Representative examples from **Paludisme-data**: T, R, S1, S2, G1, G2-5, and Unp red blood cells.

(b) Representative examples from **HAM10000**: akiec, bcc, bkl, df, mel, nv, and vasc.

Figure 2: Representative examples from Paludisme-data (left) and HAM10000 (right).

This work benchmarks an active learning framework on two class-imbalanced image datasets spanning hematology and dermatology, totaling 49,110 images for Paludisme-data and 10,015 for HAM10000 (Tschandl et al., 2018). The Paludisme-data corpus has been developed by a military medical research institution. It is currently under publication process in an international journal and will be made publicly available as soon as the paper is accepted. Paludisme-data is aimed to train and benchmark methods for detecting *Plasmodium falciparum* (the parasite found in the red blood cells of people with malaria) and recognizing its development stages from human blood smear images. Automatic recognition of *Plasmodium* development stages is crucial for helping medical researchers in monitoring parasite cultures, characterizing stage-specific drug targets, and measuring parasite survival after drug exposure. Paludisme-data comprises Giemsa-stained thin blood smears of *Plasmodium falciparum* infections captured at 100× optical magnification, yielding RGB images of 2560 × 1920 pixels. Blood smear images were manually annotated using the YOLO format by a pool of five expert microscopists following parasite-stage morphology (World Health Organization, 2015; Beare et al., 2006; Bibin et al., 2017; Liang et al., 2019), and include both infected and uninfected (unparasitized) red blood cells while preserving clinical class imbalance (unparasitized cells constitute 97% of training instances). As shown in Figure 2a, class names are abbreviated as *Unp* for *Unparasitized red blood cell*, *T* for *Trophozoite*, *R* for *Ring*, *S1* for *Schizont, sub-stage 1 (mononucleated schizont)*, *S2* for *Schizont, sub-stage 2 (polynucleated schizont)*, *G1* for *Gametocyte, sub-stage 1*, and *G2-5* for *Gametocyte, sub-stages 2 to 5*. and the per-split class distributions are detailed in Table 1. Please note that a distinction is made between

two sub-stages for schizonts (S1 and S2) and gametocytes (G1 and G2-5), so as to avoid having too high intra-class variability during training, these sub-stages being visually quite dissimilar. But, during inference and for their final applications, medical researchers are mainly interested in the stage (schizont *i.e.* S1 $\vee$ S2, gametocyte *i.e.* G1 $\vee$ G2-5), not the sub-stage. Complementing this microscopy setting, the HAM10000 (Human Against Machine with 10,000 training images) (Tschandl et al., 2018) dataset offers dermatoscopic images of pigmented skin lesions collected by the Medical University of Vienna and the University of Queensland; As shown in Figure 2b, images come from multiple digital dermatoscopes and scanned diapositives with native resolutions from $600 \times 450$ to $1024 \times 768$ pixels, grouped into seven diagnostic categories (*akiec*: actinic keratoses; *bcc*: basal cell carcinoma; *bkl*: benign keratosis; *df*: dermatofibroma; *mel*: melanoma; *nv*: melanocytic nevi; *vasc*: vascular lesions) and widely used for automated skin-lesion classification (Esteva et al., 2017; Codella et al., 2018; Tschandl et al., 2018; Rosendahl et al., 2011). More than half of the lesions are histopathologically verified, and the remainder are confirmed by follow-up, confocal microscopy, or expert consensus. The per-split class distribution for HAM10000 is reported in Table 1. Please note that the two datasets we consider exhibit different levels of imbalance. Indeed, Paludisme-data corresponds to a highly imbalanced setting, since the total proportion of minority classes represents only 2.68%of the whole dataset *c.f.* Table 1). This reflects a severe underrepresentation of rare categories and, consequently, a more challenging learning scenario. In contrast, HAM10000 exhibits a moderate imbalance level, with minority classes accounting for 33,05% of the whole dataset *c.f.* Table 1).

Table 1: Class distribution by split for Paludisme-data (left) and HAM10000 (right), with full-dataset share and minority/majority status.

| **Paludisme-data** | | | | | | **HAM10000** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **Train** | **Validation** | **Test** | **Full (%)** | **Imbalance** | **Class** | **Train** | **Validation** | **Test** | **Full (%)** | **Imbalance** |
| G1 | 25 | 4 | 8 | 0.07 | minority | akiec | 236 | 26 | 65 | 3.27 | minority |
| G2-5 | 76 | 14 | 33 | 0.25 | minority | bcc | 370 | 41 | 103 | 5.13 | minority |
| S1 | 71 | 11 | 20 | 0.20 | minority | bkl | 791 | 88 | 220 | 10.97 | minority |
| S2 | 144 | 22 | 43 | 0.42 | minority | df | 83 | 9 | 23 | 1.15 | minority |
| T | 117 | 26 | 41 | 0.37 | minority | mel | 801 | 89 | 223 | 11.11 | minority |
| R | 482 | 78 | 115 | 1.37 | minority | nv | 4,827 | 537 | 1,341 | 66.95 | majority |
| Unp | 33,066 | 4,972 | 9,742 | 97.32 | majority | vasc | 102 | 12 | 28 | 1.42 | minority |
| **Total** | **33,981** | **5,127** | **10,002** | **100** | – | **Total** | **7,210** | **802** | **2,003** | **100** | – |

### 4.2 Model and Active Learning Configuration

We adopt a unified active learning framework using an EfficientNetV2-B3 backbone pretrained on ImageNet and fine-tuned with $300 \times 300$ inputs (Tan & Le, 2021). In both benchmarks, training is performed on cropped region-of-interest images rather than raw acquisitions: HAM10000 provides lesion-centered dermatoscopic crops (standardized from the original sources), while Paludisme-data classification uses single-cell crops extracted from smear images using the available bounding-box annotations. All images are padded to a square, converted to RGB, resized, and normalized with ImageNet statistics. Data augmentation includes random flips, $90^\circ$ rotations, and color jitter. Hyperparameters (learning rate $10^{-3}$, weight decay $10^{-4}$, batch size 32) are tuned separately per dataset via grid search on the validation set. Training uses AdamW optimizer instance (lr $10^{-3}$, weight decay $10^{-4}$) with model weights restored from the previous AL iteration's best validation checkpoint, cross-entropy loss, and early stopping. Each experiment consists of 5 runs and 8 AL iterations, initialized with 5 labeled samples per class, on which the model is first trained at AL iteration 0. From this initial checkpoint, for each AL method, a copy of the model is created and then fine-tuned independently from other methods, ensuring a fair and equal starting point while allowing each method to follow its own learning trajectory from AL iteration 1 onward. As the labeled set grows across AL iterations, the number of training epochs is set to 10 in the early stages and increased proportionally to the size of the labeled data. The sampling budget is fixed at 200 samples per AL iteration, with a total annotation budget capped at 1,635 samples for both datasets, reflecting realistic annotation costs. Since both datasets are already fully annotated by professional experts, the human expert annotator is simulated by an oracle agent that provides ground-truth labels for all selected instances across all AL iterations.

### 4.3 FAUSAL variants: fluctuation dissimilarity measures

FAUSAL instantiations differ only by the choice of dissimilarity measures $d(\cdot,\cdot)$ used to compare the predictive distributions of the same unlabeled sample across two successive AL iterations, $p^{(t)}(x)$ and $p^{(t-1)}(x)$. The generic fluctuation and its normalized form are given in Eq. equation 4 and Eq. equation 5, respectively, where $M_d > 0$ ensures that $F^{(t)}(x) \in [0,1]$ for consistency across variants and within Eq. equation 2.

#### 4.3.1 FAUSAL–L1HEADS

FAUSAL-L1HEADS introduces a new fluctuation measurement function that explicitly tracks instability in the top-ranked predicted classes across successive AL iterations. Unlike margin sampling (Xie et al., 2022; Roth & Small, 2006), which selects instances close to the decision boundary based on the gap between the top two class probabilities at a single time step, FAUSAL-L1HEADS measures how much the highest-scoring class probabilities change over time, regardless of the current margin value.

Let $c_t^\star = \arg\max_c p_c^{(t)}(x)$ and $c_{t-1}^\star = \arg\max_c p_c^{(t-1)}(x)$. We set $M_d = 1$ since the L1HEADS fluctuation function is already normalized. We then define:

$$\begin{aligned}\Delta^{(t)}_{\text{L1HEADS}}(x) = \max\Big( &\left| p^{(t)}_{c_t^\star}(x) - p^{(t-1)}_{c_t^\star}(x) \right|, \\ &\left| p^{(t)}_{c_{t-1}^\star}(x) - p^{(t-1)}_{c_{t-1}^\star}(x) \right| \Big).\end{aligned} \tag{6}$$

#### 4.3.2 FAUSAL–L1VEC

This variant captures the largest coordinate-wise probability change (local shift on a single class), sensitive to isolated class probability swings (Levin & Peres, 2017):

$$\Delta^{(t)}_{\text{L1VEC}}(x) = \max_{c\in\{1,\dots,C\}} \left| p_c^{(t)}(x) - p_c^{(t-1)}(x) \right|. \tag{7}$$

Since $\Delta^{(t)}_{\text{L1VEC}}(x) \in [0,1]$, we set $M_d = 1$.

#### 4.3.3 FAUSAL–TV

Total variation measures the global $\ell_1$ change between the two predictive vectors (Gibbs & Su, 2002), capturing overall distributional shift:

$$\Delta^{(t)}_{\text{TV}}(x) = \frac{1}{2}\sum_{c=1}^{C} \left| p_c^{(t)}(x) - p_c^{(t-1)}(x) \right|. \tag{8}$$

By definition, $\Delta^{(t)}_{\text{TV}}(x) \in [0,1]$, so we set $M_d = 1$.

#### 4.3.4 FAUSAL–W1

Wasserstein distance accounts for probability mass displacement along the 1D discrete support $\{0,\dots,C-1\}$ (Murali & Mishra, 2022, p. 3, eq. (1)), quantifying "transport cost" between distributions:

$$\begin{aligned}\Delta^{(t)}_{\text{W1}}(x) &= W_1\Big(p^{(t)}(x),\, p^{(t-1)}(x)\Big), \\ F^{(t)}_{\text{W1}}(x) &= \frac{\Delta^{(t)}_{\text{W1}}(x)}{C-1}.\end{aligned} \tag{9}$$

Since $0 \le W_1(\cdot,\cdot) \le C-1$ on this support, setting $M_d = C-1$ yields $F^{(t)}_{\text{W1}}(x) \in [0,1]$.

### 4.4 Variant selection protocol

To determine the optimal FAUSAL configuration, we perform a systematic model-selection sweep on the *validation* split. We evaluate four FAUSAL variants, L1HEADS, L1VEC, TV, W1, over a discrete grid $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8\}$, where $\lambda$ balances predictive uncertainty and temporal fluctuation in Eq. equation 2. For each (variant, $\lambda$) pair, the full active-learning loop is repeated across five runs, and macro-F1 scores are tracked on the validation split at each AL iteration. The validation study was conducted on both datasets, Paludisme-data and HAM10000. Table 2 presents the validation macro-F1 scores obtained on Paludisme-data at AL iterations 2, 4, 6, and 8. The results indicate that $\lambda = 0.2$ consistently yields the strongest performance across the FAUSAL variants (L1HEADS, L1VEC, W1, and TV), while $\lambda = 0.4$ remains the closest competing configuration and provides the most balanced behavior across variants. As $\lambda$ increases, a progressive decrease in macro-F1 can be observed, which leads to restricting the analysis to values up to $\lambda = 0.6$. The same trend is observed on HAM10000. The final evaluation therefore relies on the FAUSAL-L1HEADS variant with $\lambda = 0.2$, used as the reference FAUSAL configuration. Test performance is reported for this configuration and compared with five baselines: Competence-based (Bharadwaj et al., 2025), CoreSet (Sener & Savarese, 2018), BALD (Gal et al., 2017), Entropy (Nguyen et al., 2022), and Random (Stolte et al., 2025). The fluctuation principle in FAUSAL relies on comparing two predictive probability distributions of the same unlabeled sample, namely $p^{(t)}(x)$ and $p^{(t-1)}(x)$. In this work, we instantiate this comparison using the four dissimilarity measures defined above (L1HEADS, L1VEC, TV, and W1). More broadly, the same principle can be implemented with alternative dissimilarity measure families, such as divergence-based measures (*e.g.*, Jensen–Shannon or Kullback–Leibler) (Shannon, 1948; Kullback & Leibler, 1951; Hiramatsu & Manamoto, 2021; Togami et al., 2020; Van Erven & Harremos, 2014), with selection guided by validation performance, numerical stability, and computational cost.

Table 2: Macro F1-score on the Paludisme-data validation dataset for several FAUSAL variants across AL iterations {2, 4, 6, 8}. The evaluated FAUSAL formulations include L1HEADS, L1VEC, W1, and TV. For each method and AL iteration, the best $\lambda$ value within the row is highlighted in bold.

| Iter | Method | $\lambda = 0.00$ | $\lambda = 0.20$ | $\lambda = 0.40$ | $\lambda = 0.60$ |
|---|---|---|---|---|---|
| Iter 2 | L1HEADS | 0.43±0.06 | **0.66±0.05** | 0.65±0.05 | 0.64±0.05 |
| | L1VEC | 0.39±0.02 | **0.65±0.10** | 0.65±0.90 | 0.64±0.10 |
| | W1 | 0.46±0.06 | **0.57±0.09** | 0.56±0.09 | 0.55±0.02 |
| | TV | 0.50±0.07 | **0.62±0.06** | 0.61±0.07 | 0.59±0.06 |
| Iter 4 | L1HEADS | 0.74±0.02 | **0.83±0.02** | 0.76±0.05 | 0.74±0.03 |
| | L1VEC | 0.69±0.07 | **0.79±0.02** | 0.76±0.05 | 0.75±0.03 |
| | W1 | 0.62±0.06 | 0.69±0.03 | **0.75±0.06** | 0.70±0.02 |
| | TV | 0.75±0.05 | **0.79±0.01** | 0.77±0.06 | 0.76±0.05 |
| Iter 6 | L1HEADS | 0.79±0.07 | **0.83±0.04** | 0.80±0.03 | 0.79±0.04 |
| | L1VEC | 0.78±0.01 | **0.81±0.05** | 0.80±0.02 | 0.77±0.04 |
| | W1 | 0.77±0.05 | **0.81±0.04** | 0.80±0.02 | 0.81±0.06 |
| | TV | 0.78±0.06 | **0.82±0.03** | 0.78±0.03 | 0.78±0.04 |
| Iter 8 | L1HEADS | 0.79±0.02 | **0.84±0.02** | 0.82±0.05 | 0.81±0.02 |
| | L1VEC | 0.78±0.01 | **0.81±0.03** | 0.80±0.03 | 0.80±0.92 |
| | W1 | 0.80±0.02 | **0.82±0.01** | 0.81±0.03 | 0.80±0.02 |
| | TV | 0.78±0.02 | **0.82±0.03** | 0.81±0.04 | 0.79±0.01 |

# 5 Results and Discussion

## 5.1 F1-Score Evolution Curves

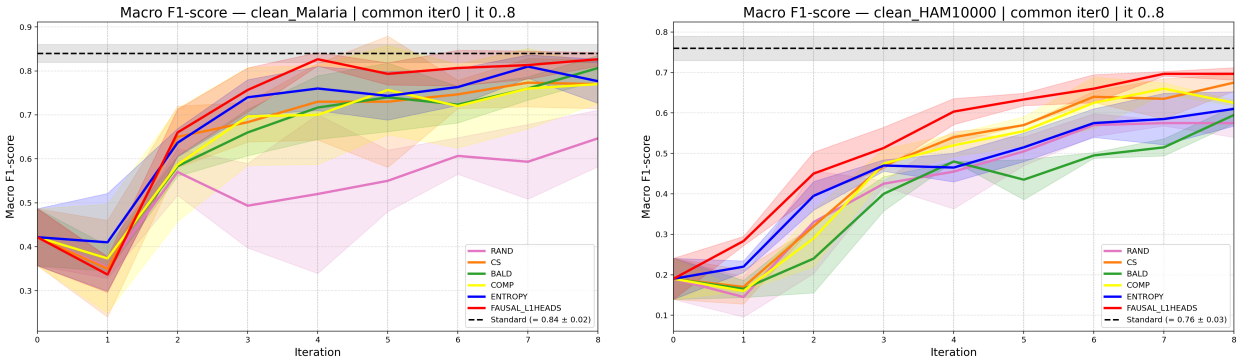



Figure 3: Evolution of macro-F1 (mean ± std over runs) on Paludisme-data (left panel ) and HAM10000 (right panel) according the number of AL iterations. Compared methods are FAUSAL-L1HEADS, Competence-based (COMP), Entropy, BALD, CoreSet (CS), and Random. Dashed horizontal lines denote the fully supervised Standard Learning reference for each dataset. Beware: the y-axis scales differ between the two graphs.

All sampling strategies are evaluated under an identical active-learning pipeline (same backbone, preprocessing/augmentation, train/val/test splits, initialization size, sampling budget per AL iteration, and number of runs). Results are reported as mean ± std across runs to reflect both effectiveness and stability. Fig. 3 reveals a consistent two-stage behavior across datasets: an initial growth stage where performance increases rapidly as informative samples are acquired, followed by a saturation stage where gains become marginal as the labeled pool becomes sufficiently representative. Consequently, differences between methods are most visible in how quickly they climb during early AL iterations and how stable they remain near saturation (variance bands). FAUSAL-L1HEADS exhibits the strongest and most consistent macro-F1 trajectory on both datasets, with comparatively tighter dispersion, indicating robust run-to-run behavior. However, the competitive landscape differs by dataset: on Paludisme-data (extreme imbalance), uncertainty- and diversity-driven baselines can intermittently approach the top curve but with larger variance, whereas on HAM10000 separations typically appear earlier and persist more clearly through later AL iterations. In addition, the variance bands highlight that some baselines are more sensitive to run-level effects, which is critical under imbalance where early acquisitions can disproportionately influence minority-class coverage.

Table 3: Performance report at AL iteration 8: macro metrics (mean ± std). Left: Paludisme-data. Right: HAM10000. Best mean per column (within each dataset) is in bold.

| Paludisme-data | | | | HAM10000 | | | |
|---|---|---|---|---|---|---|---|
| Method | Precision | Recall | F1-score | Method | Precision | Recall | F1-score |
| RAND | $0.70 \pm 0.04$ | $0.65 \pm 0.06$ | $0.65 \pm 0.06$ | RAND | $0.66 \pm 0.01$ | $0.57 \pm 0.01$ | $0.60 \pm 0.01$ |
| CS | $0.82 \pm 0.07$ | $0.76 \pm 0.04$ | $0.77 \pm 0.06$ | CS | $\mathbf{0.76 \pm 0.04}$ | $0.63 \pm 0.01$ | $0.68 \pm 0.02$ |
| BALD | $0.82 \pm 0.03$ | $0.80 \pm 0.03$ | $0.81 \pm 0.03$ | BALD | $0.66 \pm 0.02$ | $0.58 \pm 0.01$ | $0.60 \pm 0.02$ |
| COMP | $0.80 \pm 0.05$ | $0.77 \pm 0.05$ | $0.77 \pm 0.04$ | COMP | $0.70 \pm 0.07$ | $0.60 \pm 0.02$ | $0.63 \pm 0.01$ |
| ENTROPY | $0.82 \pm 0.06$ | $0.77 \pm 0.05$ | $0.78 \pm 0.05$ | ENTROPY | $0.74 \pm 0.01$ | $0.56 \pm 0.05$ | $0.61 \pm 0.04$ |
| FAUSAL-L1HEADS | $\mathbf{0.85 \pm 0.02}$ | $\mathbf{0.82 \pm 0.02}$ | $\mathbf{0.83 \pm 0.02}$ | FAUSAL-L1HEADS | $0.71 \pm 0.01$ | $\mathbf{0.69 \pm 0.04}$ | $\mathbf{0.70 \pm 0.02}$ |

## 5.2 Macro Metrics and Terminal AL Iteration Results

Following standard practices in medical image classification, we report macro-averaged precision, recall, and F1. For each class $c \in \{1, \ldots, C\}$:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \qquad R_c = \frac{TP_c}{TP_c + FN_c}. \tag{10}$$

Macro-averaged metrics are computed as:

$$\text{Macro-Precision} = \frac{1}{C}\sum_{c=1}^{C} P_c, \quad \text{Macro-Recall} = \frac{1}{C}\sum_{c=1}^{C} R_c, \tag{11}$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2P_cR_c}{P_c + R_c}. \tag{12}$$

Table 3 summarizes the precision–recall–F1 trade-off at the final AL iteration 8. Across both datasets, FAUSAL-L1HEADS attains the highest macro-F1, indicating the best balance between macro-precision and macro-recall. On Paludisme-data, improvements in macro-F1 are associated with increases in both macro-recall and macro-precision compared to the state-of-the-art approaches. On HAM10000, FAUSAL-L1HEADS achieves the top macro-F1 by substantially improving macro-recall while maintaining competitive macro-precision (the highest macro-precision being obtained using CS).

### 5.3 Cumulative Selected Labels Across AL Iterations

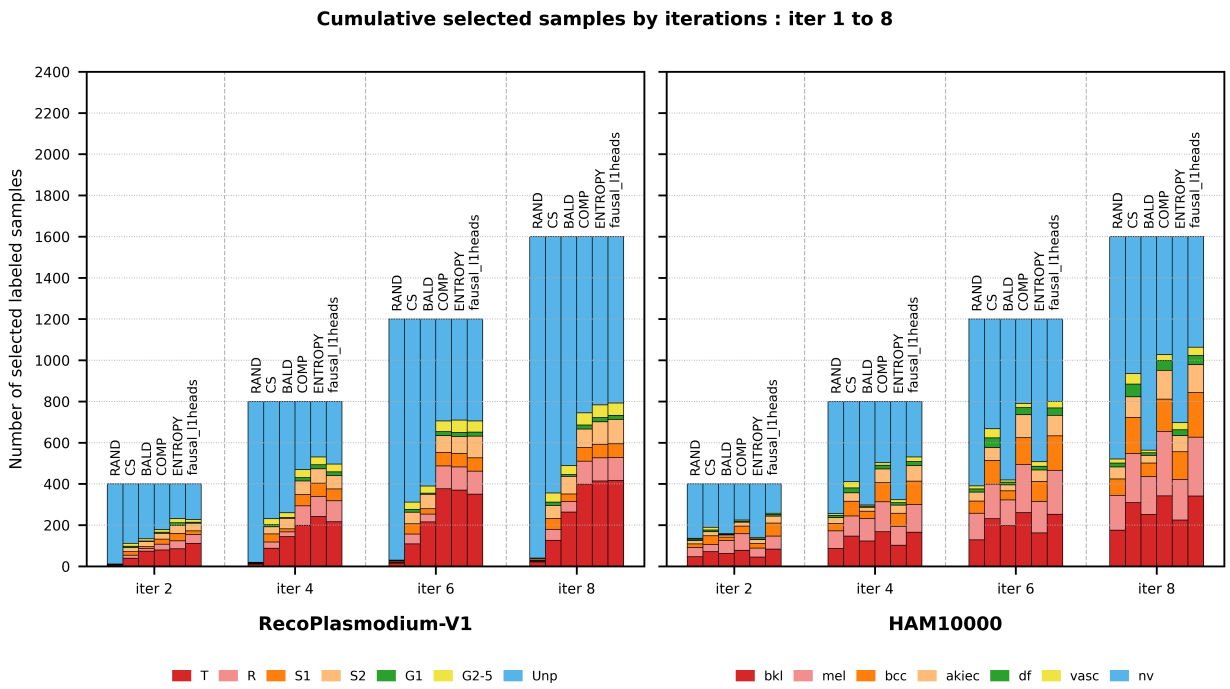



Figure 4: Cumulative number of selected labels (mean over runs) throughout the active-learning process at AL iterations 2, 4, 6, and 8 for Paludisme-data and HAM10000. Strategies compared: FAUSAL-L1HEADS, Entropy, Competence-based, BALD, CoreSet, and Random.

A key indicator of sampling quality under class imbalance is not only the number of acquired samples, but also how early minority classes enter the labeled set, since early exposure to rare categories often leads to faster improvements in macro performance metrics. Fig. 4 shows the cumulative per-class acquisitions at AL iterations 2, 4, 6, and 8 for both Paludisme-data and HAM10000, illustrating how each strategy distributes its budget between majority and minority classes. On both datasets, FAUSAL-L1HEADS exhibits a clear minority-first behavior, bringing rare classes into the labeled pool earlier than competing strategies. This is consistent with the main idea of FAUSAL: minority samples tend to produce unstable predictions across successive AL iterations, and this temporal instability is captured by the fluctuation score, which

gives priority to these samples during acquisition. Under the cumulative sampling budget of 1,600 samples at AL iteration 8, FAUSAL-L1HEADS recovers a large proportion of minority instances, as reported in Table 4. It captures about 86.59% of minority samples in Paludisme-data (792/915) and about 44.61% in HAM10000 (1,063/2,283), while using only 4.81% and 22,67% of the available training dataset, respectively. Among the baselines, the closest competitors remain dataset-dependent: Entropy and Competence-based on Paludisme-data, and Competence-based with CoreSet on HAM10000. However, FAUSAL-L1HEADS not only selects minority samples earlier, but also tends to select the most informative ones, which helps improve decision boundaries sooner during training. In contrast, Random stays mostly focused on majority-classes, delaying the acquisition of minority samples and requiring more labeled data to reach similar performance levels.

Table 4: Per-method number of selected minority-class samples at AL iteration 8, with percentages relative to the total minority pool from the training split.

| Method | Paludisme-data | HAM10000 |
|---|---|---|
| Total minority samples | 915 | 2383 |
| RAND | 45.00 (4.92%) | 521.00 (21.86%) |
| CS | 495.00 (54.10%) | 935.50 (39.26%) |
| BALD | 506.00 (55.30%) | 563.50 (23.65%) |
| COMP | 742.50 (81.15%) | 1027.50 (43.12%) |
| ENTROPY | 761.00 (83.17%) | 697.00 (29.25%) |
| FAUSAL_L1HEADS | 792.33 (86.59%) | 1063.00 (44.61%) |

### 5.4 Computational time analysis

Table 5 reports the average sampling (selection) time per AL iteration. Across both datasets, BALD is consistently the slowest strategy, reflecting the extra computation required by its uncertainty estimation, while CoreSet is also relatively expensive due to its diversity-driven selection mechanism. In contrast, COMP and ENTROPY form the fastest baseline group, with noticeably lower runtimes and limited variance. FAUSAL variants are moderately costly: their sampling times remain close to ENTROPY on Paludisme-data and sit between ENTROPY/COMP and CS/BALD on HAM10000, with small standard deviations that indicate stable computational times across AL iterations.

Table 5: Selection time statistics on Paludisme-data and HAM10000 (mean $\pm$ std over 8 AL iterations (s)).

| Method | Paludisme-data | HAM10000 |
|---|---|---|
| CS | $165.02 \pm 2.59$ | $86.82 \pm 6.87$ |
| BALD | $200.81 \pm 10.37$ | $97.38 \pm 4.75$ |
| COMP | $109.55 \pm 5.60$ | $60.37 \pm 5.60$ |
| ENTROPY | $124.31 \pm 2.88$ | $61.23 \pm 3.66$ |
| **FAUSAL-L1VEC** | $\mathbf{124.54 \pm 1.90}$ | $\mathbf{71.95 \pm 5.18}$ |
| FAUSAL-L1HEADS | $124.75 \pm 1.76$ | $72.23 \pm 5.17$ |
| FAUSAL-TV | $125.47 \pm 2.35$ | $72.10 \pm 5.80$ |
| FAUSAL-W1 | $127.44 \pm 1.78$ | $72.66 \pm 5.86$ |

## 6 Discussion

After a thorough analysis based on a broad set of evaluation metrics, clear limitations emerged among the competing methods. The proposed FAUSAL approach effectively addresses these limitations. First, we tackle the stability limitation across various degrees of imbalance. We observed that most existing strategies exhibit limited robustness in this context. For instance, while Entropy performs strongly on highly imbalanced data such as Paludisme-data, its effectiveness diminishes on HAM10000, which exhibits a more moderate imbalance level. This inconsistency highlights the advantage of our novel fluctuation term in FAUSAL, which significantly enhances stability compared to Entropy. From a clinical perspective, this stability is essential: an unstable sampling strategy may yield acceptable parasitaemia estimation in one setting but miss critical melanoma cases in another. A similar trend is seen with CoreSet, which gives the second-best results on HAM10000 where the data is more evenly distributed, but struggles with highly

imbalanced datasets like Paludisme-data, where distinguishing between classes in the latent space becomes challenging due to the dominance of majority-classes. BALD, on the other hand, lacks stability across both datasets, sometimes performing worse than random selection. Competence-based demonstrates more consistent behavior across both datasets; while not the top performer, its stability likely stems from a mature, balanced blue its approach to modeling uncertainty. Our findings also reveal that uncertainty-based strategies are particularly effective under high class imbalance. This is probably because, in such settings, the model tends to remain highly uncertain about minority classes with little data, while quickly gaining confidence in majority-classes that are well represented. As a result, uncertainty-based approaches focus sampling on the ambiguous, underrepresented classes, accelerating learning where it is most needed. Notably, FAUSAL remains robust across varying degrees of imbalance and is especially effective for early selection of minority classes. We attribute this strength to the fluctuation term, which samples data with strongly fluctuating predictions, corresponding to difficult or ambiguous cases. In clinical terms, these are precisely the cases that matter most,ambiguous parasitic stages that affect treatment decisions, or atypical lesions at risk of being misclassified as benign despite being melanoma.

## 7 Conclusion

This paper introduces a novel active learning sampling strategy (FAUSAL) for image classification under scarce and imbalanced labels. Our contributions are as threefold. First, we formulate and exploit the hypothesis that highly fluctuating, ambiguous images, where the model fails to learn stable decision boundaries, are the most informative samples for driving further improvement under limited expert supervision. More precisely, we design FAUSAL as a fluctuation-aware sampling strategy that combines predictive uncertainty with short-term temporal fluctuation between successive AL iterations, so that the sampling budget is focused on these challenging regions. In clinical practice, this means directing expert attention toward the samples most likely to improve the detection of critical conditions, reducing the risk of missed diagnoses under constrained annotation budgets. Second, as part of this framework, we propose FAUSAL–L1HEADS, a new fluctuation-aware sampling function that explicitly tracks instability in the top-ranked predicted classes across AL iterations. Third, we demonstrate the effectiveness of this strategy on two medical image classification datasets with different imbalance levels, where FAUSAL achieves macro-F1 scores of about 0.83 on Paludisme-data and about 0.70 on HAM10000 using only 4.81% and 22.67% of the available training dataset, respectively. These results translate into clinically meaningful outcomes: near-expert sensitivity for malaria detection with as few as 435 labeled images, and a reduction of up to 17 misdiagnosed melanomas compared to alternative sampling strategies. Future work will further develop the FAUSAL framework by investigating the role of the fluctuation window used to capture prediction variations across AL iterations, with adaptive fluctuation windows that evolve as the model becomes more stable. In addition, the framework will be extended to multilabel classification, where fluctuation must account for multiple simultaneous labels, and to hierarchical learning, in order to support active sampling across several semantic levels and better represent complex medical image structures. Another important direction concerns annotation budget prediction. Rather than fixing the budget in advance, future research will explore methods, including Learning Curves based approaches (Mohr & van Rijn, 2024), to estimate beforehand the number of annotations needed to achieve a desired performance level.

## References

D. Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.

F. Bai et al. Distal: A domain-shift active learning framework. *IEEE Trans. Med. Imag.*, 2025. early access.

N. Beare et al. *Manual for the Microscopic Diagnosis of Malaria Parasites.* CDC, Atlanta, GA, USA, 2006.

J. Z. Bengar, J. van de Weijer, L. López-Fuentes, and B. Raducanu. Class-balanced active learning for image classification. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1536–1545, Waikoloa, HI, USA, January 2022.

K. Bharadwaj et al. A novel active learning technique for fetal health classification. *IEEE Access*, 13: 24823–24836, 2025.

D. Bibin, M. S. Nair, and P. Punitha. Malaria parasite detection from blood smear images using deep belief networks. *Biomed. Res.*, 28(9):4030–4036, 2017.

K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, pp. 59–66, Washington, DC, USA, August 2003.

N. Codella et al. Skin lesion analysis toward melanoma detection: A challenge at isbi 2017. In *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, pp. 168–172, Washington, DC, USA, April 2018.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Intell. Res.*, 4:129–145, 1996.

A. Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (7639):115–118, 2017.

Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, pp. 1183–1192, Sydney, Australia, August 2017.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *Int. Stat. Rev.*, 70(3):419–435, 2002.

T. Hiramatsu and T. Manamoto. Detection of maliciously blurred image portions. In *Proc. ICWAPR*, Chengdu, China, December 2021.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. In *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, pp. 249–256, Bellevue, WA, USA, June 2011.

S. Huang, X. Li, and S. C. Zhu. Uncertainty sampling vs. density sampling: An empirical study. In *NeurIPS Workshop*, Montreal, QC, Canada, December 2014.

A. Kirsch, J. Van Amersfoort, and Y. Gal. Batchbald: Efficient batch acquisition. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, December 2019.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. AMS, Providence, RI, USA, 2017.

D. Li et al. A survey on deep active learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 36(4):5879–5899, 2024a.

W. Li et al. Pathal: An active learning framework. *IEEE Trans. Med. Imag.*, 41(5):1176–1187, 2022.

X. Li et al. Unlabeled data selection for active learning in image classification. *Sci. Rep.*, 14(1):424, 2024b.

Z. Liang et al. Cnn-based malaria parasite detection and classification in thin blood smears. *Comput. Med. Imag. Graph.*, 74:58–68, 2019.

F. Mohr and J. N. van Rijn. Learning curves for decision making in supervised machine learning: A survey. *Mach. Learn.*, 113(11):8371–8425, 2024.

N. Murali and D. Mishra. Wasserstein distance for cross-modality re-identification. In *Proc. IEEE INDICON*, India, November 2022.

V. Nath et al. Diminishing uncertainty within the training pool. *IEEE Trans. Med. Imag.*, 40(10):2534–2547, 2020.

V. L. Nguyen, M. H. Shaker, and E. Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn.*, 111(1):89–122, 2022.

C. Rosendahl, P. Tschandl, A. Cameron, and H. Kittler. Diagnostic accuracy of dermatoscopy. *J. Am. Acad. Dermatol.*, 64(6):1068–1073, 2011.

D. Roth and K. Small. Margin-based active learning for structured output spaces. In *Mach. Learn.: ECML 2006*, volume 4212 of *Lecture Notes Comput. Sci.*, pp. 413–424. Springer, Berlin, Germany, 2006.

O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, April 2018.

B. Settles. Active learning literature survey. Technical Report 1648, Univ. Wisconsin–Madison, Madison, WI, USA, 2009.

C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 623–656, 1948.

S. Shetty et al. Data augmentation vs. synthetic data generation. In *Proc. IEEE ICIIS*, pp. 1–6, Sri Lanka, December 2023.

S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 5972–5981, Seoul, South Korea, October 2019.

N. Stolte, J. Daru, H. Forbert, D. Marx, and J. Behler. Random sampling versus active learning algorithms for machine learning potentials of quantum liquid water. *J. Chem. Theory Comput.*, 21(2):886–899, 2025.

M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

M. Togami et al. Unsupervised training for deep speech source separation. In *Proc. IEEE ICASSP*, Barcelona, Spain, May 2020.

P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images. *Sci. Data*, 5(1):1–9, 2018.

T. Van Erven and P. Harremos. Rényi divergence and kullback–leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

D. Wang and Z. Zhang. A comparison study of image data augmentation techniques. In *Proc. Int. Conf. Comput. Vis. Appl. Algorithm (CVAA)*, pp. 410–414, 2025.

H. Wang et al. A comprehensive survey on deep active learning. *Med. Image Anal.*, 95:103201, 2024.

World Health Organization. *Guidelines for the Treatment of Malaria.* WHO, Geneva, Switzerland, 3 edition, 2015.

M. Xie, Y. Li, Y. Wang, Z. Luo, Z. Gan, Z. Sun, et al. Learning distinctive margin toward active domain adaptation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7993–8002, New Orleans, LA, USA, June 2022.

P. Yao et al. Deep learning on imbalanced datasets for skin lesion classification. *IEEE Trans. Med. Imag.*, 41(5):1242–1254, 2022.

C. Yu et al. Multi-class imbalanced dataset classification. In *Proc. IEEE HPCC*, pp. 1768–1773, New York, NY, USA, August 2015.

L. Zhong et al. Unisal: Unified semi-supervised active learning. *Med. Image Anal.*, 102:103542, 2025.