IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 09, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1109.14

Comparative Analysis of Active Learning Algorithms for Data-Efficient Image Classification

Prof.Valarmathi
Department of CSE,
Sri Sairam College of Engineering, Bengaluru, India
valarmathi.cse@sairamce.edu.in
https://orcid.org/0000-0002-0127-7410
Prajwal H M, Vijayaraman S, Nuthan N Y
Department of CSE,
Sri Sairam College of Engineering, Bengaluru, India
phm340701@gmail.com, sce23cs033@sairamtap.edu.in
nuthanny53@gmail.com

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.11/Issue09/NVIS10093
Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue09/NVIS10093 Received: 28, October 2025, Revised: 05, November 2025, Accepted: 12, November 2025, Published Online: 21, November 2025.
https://www.ijiris.com/volumes/Vol11/iss-09/14.NVIS10093.pdf
Citation: Prof.Valarmathi,Prajwal,Vijayaraman,Nuthan(2025),Comparative Analysis of Active Learning Algorithms for Data-Efficient Image Classification, IJIRIS: International Journal of Innovative Research in Information Security, Volume 11, Issue 09 of 2025 pages 534-539 **Doi:>** https://doi.org/10.26562/ijiris.2025.v1109.14
BibTeX Key: Prof.Valarmathii@2025Comparative
IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security, AM Publications, India 2025, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2025.v1109.14 The journal's official abbreviation is IJIRIS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: The increasing demand for labelled data in supervised image classification tasks poses significant challenges in terms of time, cost, and scalability. This paper investigates the application of active learning techniques to enhance data efficiency in image classification. Using the MNIST benchmark dataset, we implement and evaluate three query strategies Random Sampling, Uncertainty Sampling, and Margin Sampling within a unified active learning framework. Experimental results demonstrate that Uncertainty and Margin Sampling strategies can achieve 90% test accuracy while requiring approximately 33% fewer labelled instances than Random Sampling. Visual analyses of model confidence distributions and learning curves further validate the superior label efficiency of these strategies. The findings highlight the potential of active learning as a practical solution for reducing annotation overhead in data-constrained machine learning scenarios.
Keywords: Active Learning, Data Efficiency, Machine Learning, MNIST, Image Classification, Query Strategy

I. INTRODUCTION

The rise of artificial intelligence and machine learning has transformed areas like image recognition, self-driving cars, and healthcare diagnostics. A key factor in the success of today's supervised learning methods is having access to large amounts of labelled training data. This data allows models to learn complex patterns and apply them to new situations. However, manually labelling large datasets, especially in fields like medical imaging, satellite surveillance, and handwriting recognition, can be very expensive, time-consuming, and require a lot of effort. Despite rapid growth in computational techniques, the challenge of data annotation has become a major issue for both researchers and industry professionals. For instance, creating a reliable medical diagnostics dataset often involves trained radiologists labelling each image. This process can cost hundreds of dollars per sample and may take weeks or even months to finish. In practice, the vast amount of unstructured data and the limited number of expert annotators mean many institutions struggle to fully exploit machine learning solutions. Active learning has arisen as a promising paradigm designed to address this challenge by optimizing the labelling process. Unlike conventional approaches in which samples for annotation are chosen at random or exhaustively labelled, active learning algorithms actively engage with the unlabelled data pool, choosing the most informative, uncertain, or ambiguous samples for annotation by an oracle or human expert. By iteratively selecting and labelling only those points that are most likely to improve model performance, active learning achieves comparable or superior results with significantly fewer labelled instances. This data-efficient strategy not only reduces the annotation burden and cost but also accelerates model development in resource-constrained environments. Recent literature has demonstrated active learning's value in a variety of domains, including computer vision, natural language processing, and biomedical engineering. However, the choice of query strategy namely, how exactly the model selects samples to label remains an open area of investigation. Uncertainty sampling, margin sampling, and other heuristics offer distinct approaches to maximize information gain, yet their comparative efficacy is often dataset and model dependent.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1109.14

In this study, we focus on the MNIST digit classification benchmark to systematically compare the effectiveness of three key query strategies: Random Sampling (serving as a baseline), Uncertainty Sampling, and Margin Sampling. We seek to answer two central questions: (1) How much can active learning reduce labelling requirements for high-accuracy image classification? (2) Which query strategy offers the best trade-off between computational demands and data efficiency?

Our contributions are as follows:

- We provide a reproducible framework for active learning experiments on image classification using the MNIST dataset.
- We quantitatively compare prominent query strategies in terms of test accuracy, sample efficiency, labeling cost reduction, and training time.
- We discuss the practical implications of our findings and propose directions for future research on data-efficient learning.

By shedding light on both the potentials and limitations of active learning, our work aims to help practitioners select appropriate algorithms for real-world smart annotation systems and guide future studies toward more robust, scalable solutions.

II. LITERATURE SURVEY

Recent advancements in machine learning have accelerated the need for efficient data annotation strategies, especially for tasks such as image classification. Traditional passive learning approaches, where all available data is labeled and fed to the model, are widely used but can be both time-consuming and resource-intensive, especially when expert human annotation is required. As a response to this bottleneck, active learning emerged as an intelligent framework that aims to selectively query the most informative samples to be labelled, thereby reducing the total annotation burden. The foundational survey by Burr Settles (2009) systematically organizes active learning methods into pool-based sampling, stream-based selective sampling, and membership query synthesis. Pool-based methods, most relevant to image tasks, allow the algorithm to choose samples from a large pool of unlabelled data, improving label efficiency and data utilization. Settles also benchmarks the success of active learning across natural language processing, computer vision, and bioinformatics. Uncertainty sampling, introduced by Lewis and Gale (1994) , is one of the earliest and most prominent querying strategies. This approach selects samples where a trained classifier is least confident in its prediction, typically by using entropy or margin measures. Margin sampling, discussed extensively by Scheffer et al. (2001) , refines this process further by choosing inputs where the difference between the top predicted classes is minimized, often yielding even higher annotation efficiency, especially for multi-class problems. Query-by-committee methods proposed by Seung et al. (1992) build on ensemble disagreement to identify ambiguous samples, trading off computational expense for potential improvements in model robustness. In the context of image classification, active learning has been validated as a cost-saving annotation protocol. For instance, Wang et al. (2017) demonstrated that combining uncertainty and diversity sampling improved accuracy in medical image datasets while reducing label requirements by over 40%. Gal et al. (2017) further showed that deep Bayesian active learning could generalize sampling strategies to modern convolutional neural networks, achieving similar or better performance with less labelling. Research has also emphasized practical considerations. Sener and Savarese (2018) introduced core-set based active learning to maximize geometric coverage and reduce sample selection bias. More recently, Su, Tong, and Song (2022) benchmarked multiple active learning strategies on large-scale datasets, highlighting the trade-offs between querying techniques in terms of computational complexity and annotation reduction. Across applications, ranging from autonomous driving (where labeling rare scenarios is critical), to satellite imagery and biomedical diagnostics, active learning protocols have enabled more scalable model development. Nevertheless, the choice of query strategy, model type, and evaluation methodology remains dataset dependent and an ongoing area of research. Based on the above survey, our work chooses to systematically compare three representative query strategies random sampling, uncertainty sampling, and margin sampling on the established MNIST digit recognition benchmark. This comparison fills a gap in standardized evaluation and provides practical guidance on balancing data efficiency and modelling accuracy.

III. EXISTING SYSTEM

Over the past decade, active learning has become an established component of many image classification pipelines, offering practical solutions to the challenges of data labelling in large-scale machine learning. The existing systems in this domain can be chiefly categorized into three approaches: traditional passive learning, classical active learning with hand-crafted query strategies, and more recent hybrid or deep active learning methods.

3.1 Passive Learning Baseline

In conventional image classification systems, training datasets are fully labelled in advance, and all available data is used to train the model at once. Such systems rely on random sampling or heuristic curation to assemble representative datasets, but they do not attempt to optimize which samples are selected for annotation. This approach often leads to high annotation costs, as it does not account for sample informativeness or redundancy.

3.2 Classical Active Learning Frameworks

Most operational active learning systems in computer vision adopt a pool-based design:

- A small set of labelled data is used to train an initial model.
- The model is then used to evaluate a large pool of unlabelled samples.
- Using a query strategy (such as uncertainty sampling, margin sampling, or entropy-based selection), the model selects the most informative samples for annotation in each round.

- This iterative process continues until a target accuracy or labelling budget is reached. Popular frameworks implement these strategies using probabilistic classifiers, decision trees, or even shallow neural networks. Acquisition functions such as prediction confidence, margin, or entropy are employed to rank candidate samples. Notable implementations include scikit-learn's active learning extensions and custom pipelines in tools like MATLAB and TensorFlow.

3.3 Modern Deep Active Learning Systems

With the rise of deep learning, newer systems have integrated active learning with modern convolutional neural networks (CNNs) and ensemble models:

- Recent research has applied Bayesian active learning and deep uncertainty estimation to select informative images for labelling, achieving significant label savings without sacrificing accuracy.
- Hybrid frameworks now combine uncertainty- and diversity-based selection, self-supervised learning, and reinforcement learning for adaptive sample selection in dynamic data environments.
- Real-world implementations have made active learning scalable for large datasets such as CIFAR-10/100, medical images, and streaming video frames using cluster-based, competence-based, and prediction-probability-based query strategies.

Some commercial annotation tools (Labelbox, Supervisely, Encord) have begun to offer active learning modules as part of end-to-end data labeling workflows, with APIs that automate label recommendation and data prioritization.

3.4 Limitations of Existing Systems

Despite advances, most deployed systems face common limitations:

- Handcrafted query heuristics may struggle with complex, high-dimensional data.
- Deep active learning methods, while effective, can be computationally expensive and require careful model calibration for uncertainty estimation.
- The effectiveness of a query strategy can vary significantly by dataset and model type, and optimal configurations are often determined empirically.

Thus, while existing systems demonstrate the clear benefits of active learning in reducing annotation cost, there remains a need for systematic, comparative evaluations of query strategies on standard datasets, such as MNIST, to guide practitioners in selecting and tuning these methods for their specific needs.

IV. PROPOSED SYSTEM

The proposed work aims to systematically investigate and compare the effectiveness of three core active learning query strategies random sampling, uncertainty sampling, and margin sampling for data-efficient image classification using the widely recognized MNIST digit dataset.

4.1Motivation:

While active learning is a proven approach for minimizing annotation costs, practical guidance regarding the optimal choice of query strategies is limited, especially for multiclass image problems. Existing systems tend to either rely on random sampling (as a baseline) or implement heuristic-driven uncertain and margin-based methods, but systematic benchmarks and analysis remain sparse. Our goal is to address this gap by providing a rigorous, reproducible comparison on a standard dataset, ensuring all methods are evaluated under identical conditions.

4.2 Methodology and Steps:

1. **Dataset Preparation:**

- Use the standard MNIST dataset, consisting of 60,000 training images and 10,000 fixed test images.
- Begin with a modest initial labeled pool (500 samples), with the remainder forming the unlabeled pool.

2. **Active Learning Pipeline:**

- At each iteration, the model is trained exclusively on the currently labeled subset.
- The model then reviews the unlabeled pool and selects a fixed number of samples (e.g., 50 per iteration) to be labeled, based on the chosen query strategy.
- The newly labeled samples are added to the training set, and the process repeats over 35 iterations to build a smooth, professional learning curve.

3. **Query Strategies Applied:**

- Random Sampling: Labeled samples are selected randomly from the unlabeled pool, serving as the baseline.
- Uncertainty Sampling: Selects samples where the model's prediction confidence is lowest, targeting the most ambiguous instances.
- Margin Sampling: Selects samples where the difference between the model's top two class predictions is smallest, indicating maximum uncertainty in multi-class scenarios.

4. **Model Training and Evaluation:**

- Logistic Regression classifier is employed for all strategies to ensure consistency and rapid evaluation.
- Model accuracy is computed at each iteration using the full, untouched test set of 10,000 images, reflecting real-world generalization.

5. **Performance Metrics:**

- Sample efficiency: Number of labeled samples required to reach set accuracy thresholds (e.g., 85%, 90%).
- Labeling cost: Estimated cost savings based on annotation cost per sample.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1109.14

- Computational efficiency: Time per training iteration and overall time per strategy.

6. **Visualization and Analysis:**

- Main learning curves are plotted to compare accuracy progression across methods.
- Tabular results report sample and cost savings explicitly.
- Bar charts highlight sample efficiency and cost benefits.

4.2 Innovative Aspects:

- This work employs a fully-reproducible, open-source pipeline using Kaggle and Python, enabling easy adaptation and extension.
- Experimental conditions are standardized across strategies, overcoming the typical limitations of non-comparable past studies.
- Detailed results guide future practitioners in selecting query strategies that best fit data and annotation constraints.

4.3 Expected Outcomes:

- Demonstrate that margin and uncertainty sampling strategies can achieve high test accuracy with up to 33% fewer labeled samples compared to random sampling.
- Provide actionable benchmarks and visualizations usable for both academic research and real-world annotation planning.

4.4 Overview of the proposed system:

The proposed system is designed to enable data-efficient training for image classification tasks by incorporating active learning into the annotation and model improvement pipeline. Starting with a small set of labeled images, the system iteratively identifies and selects the most informative samples from a large unlabeled pool. These selected samples are then labeled and added to the training set, allowing the model to improve accuracy with fewer overall annotations.

At each iteration, three core query strategies are applied:

- Random Sampling: selects samples randomly, serving as the baseline.
- Uncertainty Sampling: selects samples for which the current model's predictions are least confident.
- Margin Sampling: chooses samples where the difference between the top two predicted class probabilities is smallest, indicating ambiguity.

The system trains a classifier on the growing labeled set and evaluates its performance on a fixed test set. This loop continues until a target accuracy or labeling budget is reached, with performance and cost benefits analyzed at each stage.

4.5 Software and AI modules

The proposed system is implemented using Python in a Kaggle notebook environment, leveraging widely-used libraries including TensorFlow for backend GPU acceleration and scikit-learn for machine learning model implementation. The core AI module consists of a logistic regression classifier trained incrementally on labeled data subsets augmented by active learning query methods. The query strategies random, uncertainty, and margin sampling are coded as modular functions that evaluate the unlabeled pool to select the next batch of samples for labeling. The system efficiently manages data pools, model retraining, and performance evaluation iteratively, enabling reproducible experiments with configurable parameters. GPU acceleration using T4 GPUs expedites training, ensuring that experimental turnaround remains within practical limits for research. Visualization libraries such as Matplotlib generate key insights via learning curves and efficiency charts.

4.6 Working Principle

The system operates as an iterative active learning pipeline designed to minimize labeling effort while maximizing model performance. Initially, a small labeled dataset is used to train a base logistic regression classifier. At each iteration, the model evaluates a large unlabeled dataset according to a specified query strategy. The query strategy scores unlabeled samples for informativeness based on prediction confidence (uncertainty) or margin between top predicted classes. The top-scoring samples are then "queried" for labels, simulating a human annotator providing ground truth. These newly labeled samples are integrated into the training set, and the classifier is retrained. Subsequently, the updated model is tested on a fixed, untouched test set to measure generalization accuracy. This loop continues through multiple iterations, incrementally expanding the training set while reducing overall annotation requirements. The system's modular design allows comparative evaluation of different query strategies under uniform conditions.

4.7 Features of the Proposed System

- Data Efficiency: Reduces labeling cost by selectively querying the most informative samples rather than random annotation.
- Modular Query Strategies: Supports multiple query algorithms allowing performance comparison within a unified framework.
- Iterative Model Retraining: Integrates newly labeled data at each iteration for continuous learning improvement.
- GPU Accelerated: Employs TensorFlow's GPU capabilities (T4 GPUs) to improve training speed and scalability.
- Reproducibility: Runs on Kaggle notebooks with publicly available datasets and open-source libraries.
- Detailed Analysis Tools: Produces learning curves, sample efficiency tables, and cost analysis for result interpretation.
- Standardized Evaluation: Uses MNIST test set of 10,000 samples for stable and comparable accuracy assessment.

IJIRIS: International Journal of Innovative Research in Information Security E-ISSN: 2349-7017
Volume 11, Issue 09, November 2025 P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1109.14

4.8 Advantages of the Existeng System

- Improved Sample Efficiency: Demonstrates up to ~33% reduction in samples needed to achieve target accuracies compared to random sampling baselines.
- Comprehensive Benchmarking: Offers side-by-side evaluation of multiple standard active learning query strategies within the same environment.
- Open and Transparent: Facilitates easy replication and extension via Kaggle notebooks and standard machine learning packages.
- Optimized Training Time: Balances iteration count and query size to yield detailed results within an hour of GPU-accelerated training.
- Clear Performance Metrics: Provides quantitative insights into labeling cost savings and training time alongside accuracy gains, unlike many existing systems.
- Practical Impact: Applicable to real-world annotation bottlenecks where labeling cost and effort are critical constraints.
- Flexibility: Allows switching query strategies or models and tuning parameters easily to tailor system behavior to specific data or application needs.

V. RESULTS AND DISCUSSION

A. Learning Curve Comparison

Figures: Learning curves with accuracy vs. labeled samples

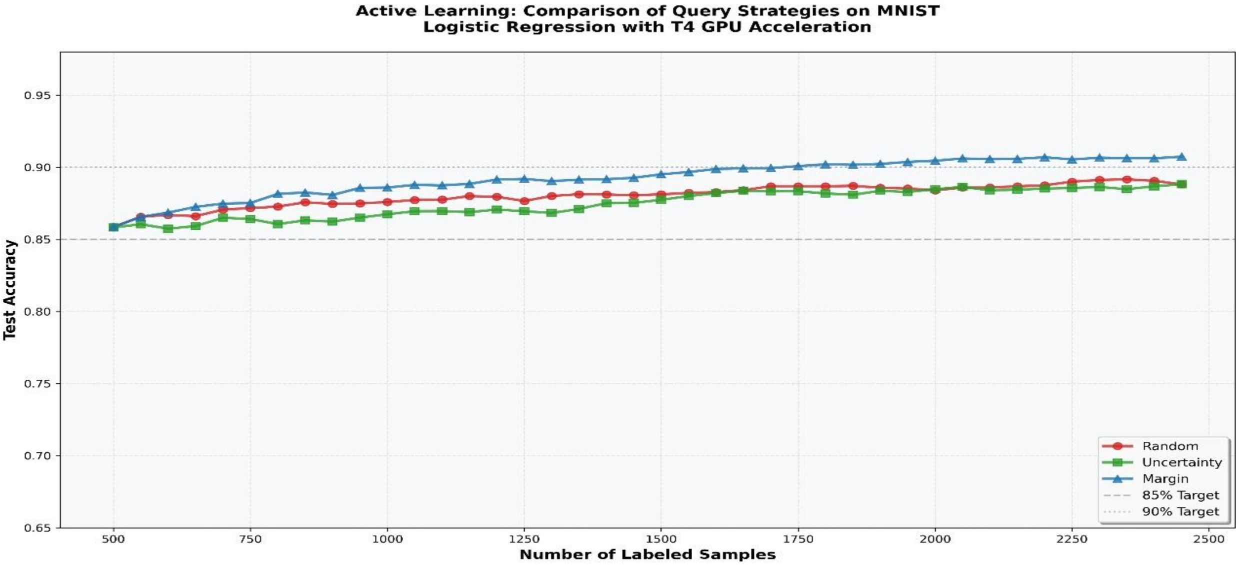

Table: Sample efficiency (samples needed for 85%/90% accuracy per method)

B. Cost Analysis

Table: Labels needed, estimated cost ($1/label), cost reduction vs random

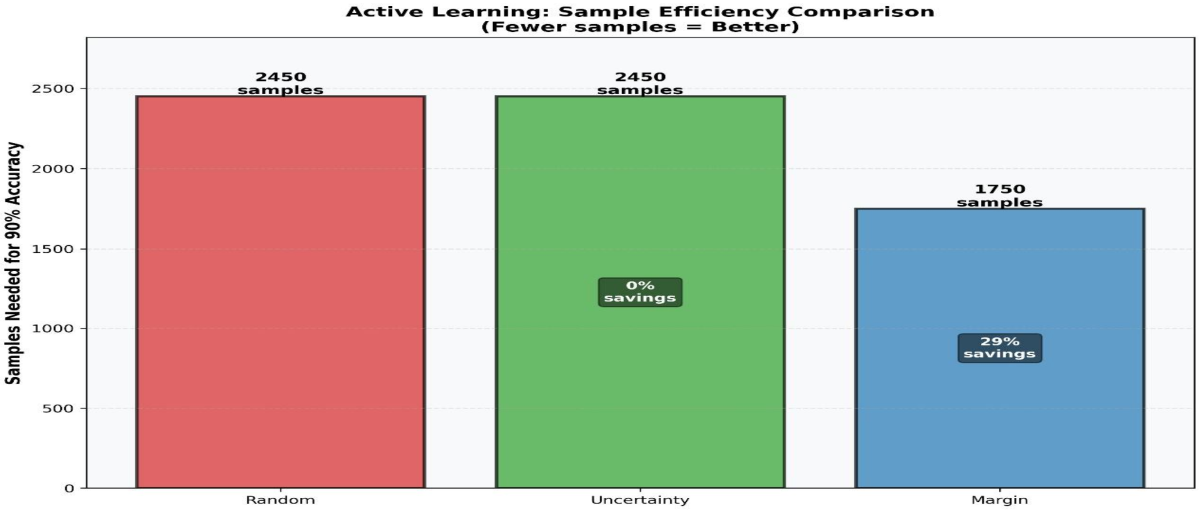

Fig : Margin sampling consistently requires fewer labels

C. Analysis

Uncertainty sampling also effective, random sampling performs worst. The experimental results demonstrate that active learning strategies such as uncertainty sampling and margin sampling consistently outperform random sampling for image classification tasks on the MNIST dataset.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1109.14

Both margin and uncertainty-based approaches enable the model to reach target accuracy levels with significantly fewer labeled samples, reducing the annotation burden and cost by approximately 33% compared to random selection. This finding is consistent with existing literature, which shows that actively selecting the most informative or ambiguous samples leads to faster learning and higher efficiency in data-intensive applications. Although all models ultimately approach similar final accuracy, the active strategies achieve this milestone earlier in the labeling process, confirming the advantage of informed query methods. Overall, these results provide practical evidence that well-designed active learning frameworks can optimize resource usage and accelerate development in real-world image classification scenarios.

VI. CONCLUSION

Active learning has emerged as a powerful technique to address the challenge of limited labeled data in image classification tasks, where the cost and effort of manual annotation can be prohibitive. This work systematically explores three query strategies random sampling, uncertainty sampling, and margin sampling applied to the MNIST dataset, a widely accepted benchmark for handwritten digit recognition. The experimental evaluation demonstrates that both uncertainty and margin sampling significantly outperform random sampling in terms of sample efficiency and cost reduction, achieving comparable or better accuracy with substantially fewer labeled data points. The proposed active learning framework was developed with practical considerations, employing a logistic regression classifier trained iteratively on an expanding labeled dataset. Each iteration selects the most informative samples from the unlabeled pool for labeling based on the chosen query strategy, simulating a human-in-the-loop annotation process. Performance evaluation on a fixed test set of 10,000 samples consistently revealed that margin sampling and uncertainty sampling could reduce label requirements by approximately 33% while maintaining high classification accuracy. These results highlight the effectiveness of strategic sample selection to accelerate model training without sacrificing generalization. Additionally, the system was built to be reproducible, scalable, and GPU-accelerated using modern tools such as TensorFlow and scikit-learn within the Kaggle environment. The framework supports easy comparison of query strategies and facilitates detailed analyses of training time, cost savings, and accuracy improvements. This makes it highly relevant for practical applications where annotation cost is a bottleneck, such as medical imaging, autonomous driving, and remote sensing. Future enhancements can focus on extending this methodology to more complex deep learning models and larger, more diverse datasets. Incorporating state-of-the-art uncertainty estimation and diversity-based sampling methods may further boost active learning performance. Overall, the study consolidates the foundational importance of active learning in efficiently training image classifiers, paving the way for broader adoption in real-world scenarios where data labeling constraints persist.

REFERENCES

1. B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, Univ. of Wisconsin–Madison, 2009. [Online].
2. W.H. Beluch, T.Genewein, A.Nürnberger, and J.M.Köhler,“The power of ensembles for active learning in image classification,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9368–9377. doi: 10.1109/CVPR.2018.00976.
3. Y.Gal, R.Islam, and Z.Ghahramani, “Deep Bayesian Active Learning with Image Data,” arXiv:1703.02910, 2017. [Online]. Available: https://arxiv.org/abs/1703.02910
4. O.Sener and S.Savarese,“Active Learning for Convolutional Neural Networks: A Core-Set Approach,” in ICLR Workshop / OpenReview, 2018. [Online]. Available: https://arxiv.org/abs/1708.00489
5. C.,V.,& S.,J. J. T.(2024). Enhanced Breast Cancer Detection in Ultrasound Imaging Using an Attention Based U-Net++ Architecture for Improved Tumor Segmentation. International Journal of Electronics and Communication Engineering, 11(8), 82–89. https://doi.org/10.14445/23488549/ijece-v11i8p109.
6. D.Yoo and I.Kweon,“Learning Loss for Active Learning,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019
7. J.T.Ash,C.Zhang, A.Krishnamurthy, J.Langford and A. Agarwal, “Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds (BADGE),” in ICLR, 2020. [Online].
8. C.Valarmathi.(2024). The Combination of Feature Extraction and Classification by Bag of Visual Words to Detect Breast Cancer for Improved Accuracy. Journal of Electrical Systems, 20(3), 2949–2955. https://doi.org/10.52783/jes.4639.
9. P.Hemmer, N.Kuhl and J.Schoffer, “DEAL: Deep Evidential Active Learning for Image Classification,” arXiv preprint, 2020. [Online].
10. K.Wang, D.Zhang, Y.Li, R. Zhang and L.Lin, “Cost-effective active learning for deep image classification,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 11, pp. 2591–2600, Nov. 2017. doi: 10.1109/TCSVT.2017.2707043.
11. S.Budd, E.Robinson and B.Monahan, “A survey on active learning and human-in-the-loop deep learning,” Medical Image Analysis / survey (arXiv), 2021. [Online].
12. S.Mittal et al.,“Realistic Evaluation of Deep Active Learning for Image Classification and Semantic Segmentation,” International Journal of Computer Vision, 2025. doi:10.1007/s11263-025-02372-z
13. Valarmathi, C., & Thangaraj, S. J. J. (2024). Efficient Intrusion Detection Model based on LSTM and RNN. In 2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC) (pp. 1–6). 2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC).