Various Active Learning Strategies Analysis in Image Labeling: Maximizing Performance with Minimum Labeled Data

Arnav Tyagi[1], Harshvardhan Aditya[2], Nitin Arvind Shelke[3], Jagendra Singh[4*], Yagna Jadeja[5], Anil V Turukmane[6], Rishabh Khandelwal[7]

[1,2,3,4,7]School of Computer Science Engineering & Technology, Bennett University, Greater Noida, India
[5]College of Science and Engineering, University of Derby, Derby, United Kingdom
[6]School of Computer science and Engineering, VIT - AP University, Amaravati, India
*jagendrasngh@gmail.com

Abstract. The use of active learning in supervised machine learning is proposed in this study to reduce the expenses associated with labeling data. Active learning is a technique that includes iteratively selecting the most informative unlabeled data points and asking a human expert to label them. Active learning can achieve high accuracy while utilizing fewer labeled examples than typical supervised learning algorithms by selecting the most informative data points. This study conducts and provides an in-depth examination and analysis of numerous active learning algorithms and their applications to various machine learning labeling problems, especially focusing on image classification. The experiments are carried out using Fashion MNIST as a benchmark dataset. This study compares the performance of five popular active learning methods BALD, DBAL, coreset, least confidence and ensemble varR for the given problem. The best performing algorithm was BALD with a mean classification accuracy of 91.31%, when 50% of the data is considered labeled, closely followed by all other techniques, making each suitable for specific use cases. The trials conducted by the study illustrates how active learning may lower the time and cost of data labeling while also maintaining high accuracy.

Keywords: Active Learning, CNN, Data Labeling, Diversity Sampling, Ens-varR, Image Classification, Uncertainty Sampling.

1 Introduction

1.1 Background and Motivation

To obtain high accuracy, supervised machine learning and deep learning algorithms require a significant amount of labeled data. Labeling data, on the other hand, can be time-consuming and costly, especially when working with huge datasets. This

limitation has sparked considerable interest in the development of active learning approaches, which aim to reduce the cost and time involved with data labeling by labeling just the most informative data points. Active learning takes a few labeled samples and uses those samples to make a labeled dataset for multiple unlabeled samples of the same problem. Active learning has been used successfully in a variety of machine learning applications such as text categorization, image recognition, and object detection [1]. Despite its potential benefits, active learning is still underutilized in many practical applications, and more research is needed to investigate its effectiveness and limitations. As a result, the purpose of this study is to investigate the application of active learning in supervised machine learning for minimizing data labeling costs, as well as its potential to revolutionize the way we train machine learning models. Through this research, we hope, will help to design more efficient and cost-effective deep learning systems, particularly in domains where huge labeled datasets are difficult or expensive to collect [2].

1.2 Hypothesis

The primary research question addressed in this work is whether active learning may reduce data labeling costs in supervised machine learning while preserving or improving model accuracy. We specifically intend to investigate the following issues:

- In terms of the sub number of labeled instances necessary for training and the consequent accuracy, how does active learning compare to classic supervised learning methods?
- How does the effectiveness of active learning differ depending on task complexity and dataset characteristics?
- What are the advantages and disadvantages of utilizing active learning to reduce data labelling costs, and how may these be addressed in actual applications?

1.3 Objective and Contribution

The following are the key goals of this paper:

- To give an in-depth examination of active learning algorithms and their applications to various machine learning problems such as text classification, image recognition, and object detection.
- To examine the efficacy of active learning in decreasing data labeling costs in supervised machine learning and to compare it to traditional supervised learning approaches in terms of the number of labeled instances required for training and the resulting accuracy.
- To examine the trade-offs and constraints of employing active learning to reduce data labeling costs, and to make recommendations on how to resolve these issues in real implementations.
- Empirical assessments using benchmark datasets will be employed to illustrate the practical application of active learning.

This paper provides the following contributions:

- A comprehensive examination of active learning, including its potential advantages and drawbacks in the context of minimizing data labeling expenses in supervised machine learning.
- Empirical findings from standardized datasets demonstrating how active learning can diminish data labeling costs while preserving or even improving accuracy.
- Exploration of the trade-offs and constraints associated with the adoption of active learning to reduce data labeling expenses, along with practical suggestions for addressing these issues during real-world implementations.
- An enhancement of our comprehension regarding the potential benefits and limitations of active learning for cost-effective data labeling in supervised machine learning, coupled with recommendations for practitioners seeking to incorporate this methodology.

2 Problem Statement

In the realm of supervised machine learning, the substantial expenses and time-intensive process of data labeling pose significant challenges. This study looks at the efficacy of active learning as a method for lowering data labeling costs while maintaining or enhancing model performance across a variety of machine learning tasks.

3 Literature Review

In recent years, the field of supervised machine learning has grown rapidly, with applications in fields as diverse as healthcare, finance, and natural language processing. The availability of big, high-quality labeled datasets is one of the important variables contributing to the success of supervised learning algorithms. The paper [3]offers a thorough examination of active learning strategies for on-road vehicle detection using computer vision. The authors examine and assess three common active learning algorithms in terms of data costs, recall, annotation costs, and precision. The detectors used in this work are based on histograms and SVM classification (HOG-SVM) [4], and Adaboost classification (Haar-Adaboost), and Haar-like features and [5].

To address the limitations of learning from such data streams, the authors suggest an online-knn classifier that joins self-labeling with demand-based active learning. The study starts by outlining the considered setup and reiterating the idea of concept drift. From a theoretical standpoint, the authors then justify the use of supervised learning for non-stationary data streams. They provide a classification of drift behaviours as well as generated self-labeling problems [6]. They provide a detailed description of their proposed online learning system, which combines self-labeling and demand-based active learning to enhance classification accuracy while lowering labelling costs. The authors test their technique on a variety of real-world datasets, including social media, cellphones, and industrial process monitoring. According to the results, their suggested strategy surpasses existing state-of-the-art approaches in terms of classification

accuracy while needing fewer labeled samples. Overall, this study contributes significantly to the area of machine learning by presenting a unique method to the problem of learning from non-stationary data streams with limited labelling. The suggested online-knn classifier, which combines self-labeling with demand-based active learning, has demonstrated encouraging results in real-world settings and has the potential to be applied in a variety of fields where data is accessible as streams [7].

The study [8] looks at how to label soundscape ecology data using visual active learning techniques. According to the scientists, appropriately labelling such data is critical for effective soundscape ecology research throughout the world [8]. However, retrieving information from this sort of data may be difficult and costly. As a result, the authors suggest a multidisciplinary strategy combining ecoacoustics, machine learning, and visualization. The concept, implementation, and testing of a Visual Active Learning technique for labelling soundscape ecology data is the major contribution of this study. The authors employ multidimensional projections to underpin the process of user-centered labelling. To summaries data detailed in visualizations, they suggest "Time Line Spectrogram" (TLS) visualizations. The authors examine the efficacy of their approach in labelling soundscape ecology data using actual data on birds, frogs, and insects. They compare their method to others and demonstrate that it exceeds them in terms of accuracy and efficiency [9, 10].

In another publication [11], the authors offer a unique hybrid framework for mining data streams that combines active and semi-supervised learning. The authors offer two techniques families based on blind and informed approaches, which result in seven algorithms for enabling active learning with self-labeling. They undertake a rigorous experimental analysis on real data streams with varying labelling budgets, demonstrating the benefits of adopting hybrid solutions when accessible class labels are few, particularly in extremely low budget scenarios [12]. The article emphasizes the difficulties of mining data streams in real-time and on a limited budget, where labelling vast volumes of data may be costly and time-consuming. The proposed hybrid technique combines active learning, in which the algorithm picks the most informative examples to label, with self-labeling, in which the system labels unlabeled data using its own predictions. This method decreases labelling costs while retaining excellent accuracy. The authors present thorough experimental data demonstrating the efficacy of their suggested algorithms in a variety of settings. They also make recommendations on where these algorithms should be used. Overall, this preprint gives useful insights on how to mine data streams quickly and cheaply utilizing active learning and self-labeling approaches [13, 14].

The research article "Automatically Labelling Video Data Using Multi-class Active Learning" provides a novel way to labelling video data. According to the scientists, manually labelling video footage is time-consuming and prone to human mistake, and it finally becomes impractical for enormous volumes of data. To solve this issue, the authors present a unified multi-class active learning strategy that use active learning techniques to choose the most informative instances for labelling while requiring the least amount of human work. The study analyses the efficacy of this technique and its prospective applications in visual information retrieval, object identification, and human activity modelling [15]. The authors broaden the active learning technique from

binary to many classes, allowing the learning algorithm to choose the most useful unlabeled input for all classes rather than just binary classes. They also offer and assess a variety of practical sample selection procedures [16].

This study has important implications for industries such as video surveillance and content analysis, where enormous volumes of video data must be reliably and effectively labeled. Overall [15] offers a novel approach to a prevalent challenge in computer vision applications. The suggested method has demonstrated promising results in terms of minimizing human labor while retaining good labelling accuracy in video data.

4 Proposed Methodology

We will utilize a combination of literature research and empirical evaluations to explore the efficiency of active learning for decreasing data labeling costs in supervised machine learning. We will begin by conducting a thorough study of the literature on active learning and its applications to diverse machine learning problems such as text classification, image recognition, and object identification. This review aims to shed light on the potential advantages and limitations of active learning while also aiding in the selection of the most effective active learning algorithms for various task categories.

Subsequently, we will undertake empirical evaluations to gauge the effectiveness of active learning in mitigating data labeling expenses within the realm of supervised machine learning. We will leverage benchmark datasets spanning diverse domains, including text classification, image recognition, and object detection. For each dataset, we will compare the performance of active learning against conventional supervised learning methods like random sampling and full labeling in terms of both the quantity of labeled instances required for training and the resulting accuracy. Furthermore, we will delve into factors affecting active learning success, such as task complexity and dataset attributes.

To conclude, we will explore the trade-offs and constraints inherent in the utilization of active learning to curtail data labeling costs, accompanied by practical recommendations for addressing these challenges in real-world implementations. Our goal is to provide a thorough understanding of the possible benefits and limitations of active learning, as well as to assist practitioners in determining whether active learning is a viable alternative for their unique applications.

4.1 Dataset Description

The Fashion MNIST dataset has been used for this active learning task. The Fashion MNIST dataset, which consists of 70,000 grayscale images of 28x28 pixels displaying ten different apparel item categories, is a commonly used benchmark dataset in computer vision research. The dataset is divided into two parts: a training set of 60,000 photographs and a test set of 10,000 images, with each image labeled with the clothing item category to which it belongs [8].

T-shirt/top, Trouser, Sneaker, Bag, Pullover, Sandal, Shirt, Dress, Coat and Ankle boot are the dataset's ten classes. The photos in the dataset are grayscale, with the pixel values preprocessed to center and normalize them.

4.2 Processing

The Fashion MNIST dataset active learning procedure consists of four steps:

- Initialization: 30000 samples from the training dataset have been chosen at random, and only their labels are considered to be existing. All the other 30000 images are considered unlabeled.
- Querying: The active learning algorithm selects a subset of unlabeled data points from the dataset that are most relevant to the model after training the initial model. This is accomplished through the use of a query method that finds samples about which the model is unclear or samples that are on the decision border. The chosen samples are labeled by an expert or annotator, and their labels are added to the labeled dataset. Because this procedure can be time-consuming and costly, the goal of active learning is to reduce the number of samples that must be labeled in order to obtain high model performance. The revised labeled dataset is then used to train a new model, and the procedure is repeated until the required level is reached.
- Labeling: When the desired accuracy has been reached, this model is used to label the unlabeled data points.
- Evaluation: After the labeling, the model's accuracy is tested and compared to that of random sampling, which is considered the baseline approach for labeling. We measure the performance of the model's using accuracy, precision, recall, and F1-score.

4.3 Approaches Used

Uncertainty Sampling: One of the most popular approaches in active learning is uncertainty sampling. On our labeled data of 30000 images, we trained a deep convolutional neural network (CNN) and its performance was evaluated on a test set[13].

Then, using the unlabeled data, we employ the uncertainty sampling approach to select the most informative samples. We specifically choose the samples for which the model is most uncertain, i.e., the samples for which the model produces the probability distribution over all possible classes with the maximum entropy. The reasoning behind this technique is that the model will benefit the most from the most uncertain samples.

- Least Confidence: For each item, the difference between 1 (100% confidence) and the most confidently predicted label is used to calculate the least confidence. Although confidence alone can be used to rank order, it can be advantageous to transform the uncertainty scores to a 0-1 range, with 1 being the most uncertain score. In that situation, the score must be normalized. The value is subtracted from 1, then multiplied by n/(1-n), where n is the number of labels. This is because the minimal confidence can never be less than one divided by the number of labels, indicating

that all labels have the same expected confidence. The least confidence approach is the most basic and widely used; it provides a ranked list of predictions in which you sample objects with the lowest confidence for their anticipated label.

- Deep Bayesian Active Learning: The DBAL(Deep Bayesian Active Learning) method includes picking the most informative samples from a huge pool of unlabeled Fashion MNIST photos repeatedly and using the real labels of those data points. In a real world setting, instead of using already labeled data, it would be better to ask for labels from an expert in that specific field. The model is then trained on the newly labeled samples, and the process is continued until the required level of performance is attained. The most useful samples for labeling are picked using acquisition functions depending on model uncertainty, such as entropy and variation ratios [10].
- Diversity Sampling: In active learning, diversity sampling is an approach for selecting samples that are diverse and representative of the underlying distribution. The aim behind diversity sampling is to choose samples that differ from those that have previously been labeled in order to cover a greater range of the input space and eliminate redundancy in the labeled data [9]. There are various approaches to measuring diversity, but one popular strategy is to employ a distance metric, such as Euclidean distance or cosine similarity, to quantify the dissimilarity between the new and labeled samples. The underlying assumption is that varied samples are ones that are far off from the labeled samples in the input space.

In practice, this method is suitable for the Fashion MNIST dataset since it is an image dataset. This strategy decreases labeling costs while retaining excellent model performance by picking a limited selection of useful photos for labeling.

- Query by Committee Sampling: QBC sampling is a prominent active learning approach for selecting samples for annotation based on disagreement among a committee of multiple classifiers. The primary principle underlying QBC is to choose samples that are challenging to classify in order to increase the classifier's performance. The QBC method entails training a committee of multiple classifiers on the labeled data available. To capture distinct characteristics of the underlying distribution, each classifier in the committee is trained using a different model architecture or a separate set of hyperparameters. After the committee has been taught, the unlabeled samples are queried depending on the committee members' disagreements. The samples with the highest levels of disagreement are then chosen for annotation.
- Ensemble Variation Ratio (Ens-varR): The ENS-Var approach was created primarily for image classification applications involving convolutional neural networks (CNNs). The strategy entails training an ensemble of CNNs on labeled data and then using them to predict on unlabeled data. The variation of these forecasts is then used to calculate uncertainty, with more variance suggesting greater uncertainty. The samples with the greatest uncertainty are then chosen for labeling in order to increase the model's accuracy. This method is continued iteratively until the desired degree of precision is attained or the labeling budget is depleted. On increasingly complicated datasets like MNIST and CIFAR-10, our experiments suggest that ENS-Var outperforms alternative active learning methods.

5 Results

The results showed that all active learning algorithms outperformed the baseline model that used all available labeled data. BALD achieved the highest mean classification accuracy of 91.31%, followed closely by ensemble varR with a mean accuracy of 90.56%. The accuracies of DBAL and coreset were 89.88% and 89.65%, respectively, which are not far behind those of DBAL and ensemble varR. Least confidence had the lowest mean accuracy of 89.12%. Table 1 showcases the accuracies achieved by all of the above stated algorithms on the Fashion MNIST dataset. Table 2 further showcases the classification report per label or class for the best performing model, BALD.

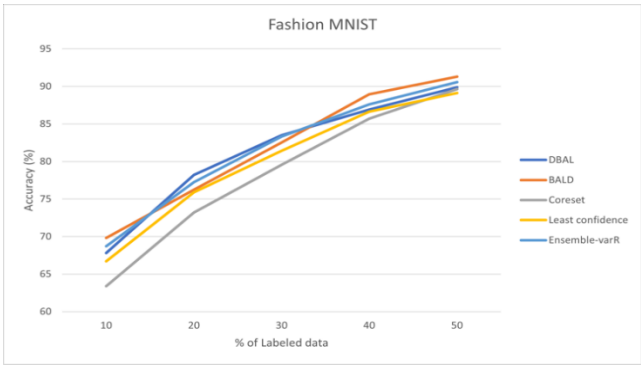

Fig. 1. Shows the graph for Accuracy comparison for different AL methods.

Table 1. Accuracy of Different Algorithms when 50% of data is considered labeled.

Algorithm	Accuracy
DBAL	89.88%
BALD	91.31%
Corset	89.65%
Least Confidence	89.12%
Ensemble varR	90.56%

Table 2. Classification report of best performing algorithm (BALD)

Class	Precision	Recall	F1-Score	Support
0 T-shirt/Top	0.89	0.88	0.88	100
1 Trouser	0.95	0.92	0.93	100
2 Pullover	0.92	0.89	0.90	100
3 Dress	0.89	0.91	0.90	100
4 Coat	0.80	0.92	0.86	100
5 Sandal	0.93	0.91	0.92	100
6 Shirt	0.81	0.72	0.75	100
7 Sneakers	0.95	0.95	0.95	100
8 Bag	0.97	0.98	0.98	100
9 Ankle Boot	0.96	0.97	0.97	100
Avg/Total	0.91	0.91	0.91	1000

6 Conclusion

The findings of the studies showed that active learning algorithms may greatly lower the quantity of labeled instances required to achieve high classification accuracy, hence lowering the overall costs associated with data labeling. In in- stance, when 50% of the data is taken into account to be labeled, the BALD algorithm attained a high accuracy of 91.31%.

These results have significant repercussions for real-world supervised ma- chine learning applications, particularly when big datasets are involved. Active learning methods may drastically reduce the expense and time needed for data labeling, which can make the process of constructing and deploying machine learning models more effective and efficient. This is done by minimizing the number of labeled instances needed for training. Therefore, this study emphasizes the potential advantages of active learning for cutting costs associated with data labeling in supervised machine learning.

The outcomes demonstrate that the BALD algorithm, alongside DBAL, coreset, least confidence, and ensembled varR are successful technique for attaining high classification accuracy while lowering data labeling expenses, and it is anticipated that future research will further investigate the possibilities of active learning in other applications and with additional datasets.

References

1. W. Jiang, “A Machine Vision Anomaly Detection System to Industry 4.0 Based on Variational Fuzzy Autoencoder,” Computational Intelligence and Neuroscience, vol. 2022, 2022, doi: 10.1155/2022/1945507.
2. A Goswami, D Sharma, H Mathuku, SMP Gangadharan, CS Yadav, “Change Detection in Remote Sensing Image Data Comparing Algebraic and Machine Learning Methods”, Electronics, Article id: 1505208, 2022.
3. Chin-Teng Lin, Mukesh Prasad, Chia-Hsin Chung, Deepak Puthal, Hesham El-Sayed, Sharmi Sankar, Yu-Kai Wang, Arun Kumar Sangaiah, “IoT-based Wireless Polysomnography Intelligent System for Sleep Monitoring", IEEE Access, Vol 6, Oct 2017.
4. Saurabh Kumar, S.K. Pathak, "A Comprehensive Study of XSS Attack and the Digital Forensic Models to Gather the Evidence". ECS Transactions, Volume 107, Number 1, 2022.
5. N. Sharma et al., “A smart ontology-based IoT framework for remote patient monitoring,” Biomedical Signal Processing and Control, vol. 68, no. March, p. 102717, 2021, doi: 10.1016/j.bspc.2021.102717.
6. Shachi Mall, “Heart Diagnosis Using Deep Neural Network", accepted in 3rd International Conference on Computational Intelligence and Knowledge Economy ICCIKE 2023, Amity University, Dubai, 2023.
7. Aditi Sharan, “Term Co-occurrence and Context Window based Combined Approach for Query Expansion with the Semantic Notion of Terms”, International Journal of Web Science(IJWS), Inderscience, Vol. 3, No. 1, 2017.
8. Yadav, C.S.; Yadav, A.; Pattanayak, H.S.; Kumar, R.; Khan, A.A.; Haq, M.A.; Alhussen, A.; Alharby, S. “Malware Analysis in IoT & Android Systems with Defensive Mechanism”. Electronics 2022, 11, 2354. https://doi.org/10.3390/electronics11152354.

9. T. Berghout, M. Benbouzid, and S. M. Muyeen, “Machine learning for cybersecurity in smart grids: A comprehensive review-based study on methods, solutions, and prospects,” International Journal of Critical Infrastructure Protection, vol. 38, no. May, p. 100547, 2022, doi: 10.1016/j.ijcip.2022.100547.
10. K. Upreti, A. K. Gupta, N. Dave, A. Surana and D. Mishra, "Deep Learning Approach for Hand Drawn Emoji Identification," 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), Bhopal, India, 2022, pp. 1-6, doi: 10.1109/CCET56606.2022.10080218.
11. Mohammad Sajid, Ranjit Rajak,” Capacitated Vehicle Routing Problem Using Algebraic Particle Swarm Optimization with Simulated Annealing Algorithm”, In Artificial Intelligence in Cyber-Physical Systems, CRC Press, 2023.
12. Aruna Yadav, A., Kumar, “A Review of Physical Unclonable Functions (PUFs) and Its Applications in IoT Environment”. In: Hu, YC., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds) Ambient Communications and Computer Systems. Lecture Notes in Networks and Systems, vol 356. Springer, Singapore, 2022.
13. Mukesh Prasad, Yousef Daraghmi, Prayag Tiwari, Pranay Yadav, Neha Bharill, “Fuzzy Logic Hybrid Model with Semantic Filtering Approach for Pseudo Relevance Feedback-based Query Expansion”, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 2017.
14. Rakesh Kumar, “Lexical Co-Occurrence and Contextual Window-Based Approach with Semantic Similarity for Query Expansion”, International Journal of Intelligent Information Technologies (IJIIT), IGI, Vol. 13, No. 3, pp. 57-78, 2017.
15. Aditi Sharan, “Term Co-occurrence and Context Window based Combined Approach for Query Expansion with the Semantic Notion of Terms”, International Journal of Web Science(IJWS), Inderscience, Vol. 3, No. 1, 2017.
16. Vijay Kumar Bohat, “Neural Network Model for Recommending Music Based on Music Genres”, In 10th IEEE International Conference on Computer Communication and Informatics (ICCCI -2021), Jan. 27-29, 2021, Coimbatore, INDIA.