Data Science and Computational Intelligence
https://dscij.org

Case Comparison of Label Efficiency with Active Learning Strategies in Satellite Image Archives

George Kan*

Department of Mathematics and Information Technology, Faculty of Liberal Arts and Social Sciences, Education University of Hong Kong, Hong Kong, Hong Kong SAR, China

Corresponding author: george.kan@eduhk.hk

Abstract

Satellite image archives are expanding at an unprecedented rate, providing invaluable data for environmental monitoring, urban planning, and disaster management. However, the supervised machine learning models typically employed to analyze these vast datasets require massive amounts of annotated data. Generating these annotations is highly labor-intensive, time-consuming, and expensive, thus necessitating methods that maximize label efficiency. This paper assesses the label efficiency of various active learning strategies applied to satellite image archives through a comprehensive case comparison methodology. By evaluating uncertainty sampling, query-by-committee, and representative sampling strategies across diverse remote sensing datasets, the research elucidates the conditions under which specific active learning frameworks yield optimal performance. The comparative evidence demonstrates that hybrid active learning strategies significantly reduce the annotation burden while maintaining high classification accuracy, particularly in highly imbalanced or complex spatial environments. The findings present a rigorous methodological framework for optimizing annotation budgets in Earth observation studies. Ultimately, the paper provides theoretical and practical insights for deploying cost-effective machine learning pipelines in remote sensing applications, fostering improved scalability in the analysis of large-scale satellite image archives.

Keywords: Active Learning; Label Efficiency; Satellite Imagery; Remote Sensing; Satellite Image Archives

1. Introduction

1.1 Background and Motivation

The proliferation of Earth observation satellites has initiated a new era in geospatial analysis, characterized by the continuous and massive accumulation of high-resolution imagery. Organizations and space agencies worldwide are capturing petabytes of data daily, creating vast archives that hold immense potential for addressing critical global challenges. These challenges include monitoring the effects of climate change, tracking rapid urbanization, managing natural resources, and responding to natural disasters. To extract actionable insights from these immense archives, the remote sensing community has increasingly turned to advanced artificial intelligence algorithms, particularly deep learning models. These models possess an extraordinary capacity to identify complex patterns, categorize land cover types, and detect specific structural objects within varied geographic contexts. However, the operationalization of these sophisticated models is fundamentally contingent upon the availability of large-scale, accurately annotated datasets [1].

The process of generating high-quality annotations for satellite imagery is notoriously demanding. Unlike generic computer vision tasks where crowdsourced workers can easily identify common objects like cats or cars, remote sensing annotations often require specialized domain knowledge. Annotators must be capable of distinguishing between subtle variations in spectral signatures, identifying complex land formations, and accurately delineating the boundaries of objects that may appear distorted due to sensor angles or atmospheric interference. Consequently, the manual labeling process is not only time-consuming but also prohibitively expensive, creating a significant bottleneck in the deployment of machine learning pipelines for Earth observation [2]. This bottleneck, often referred to as the annotation challenge, severely limits the scalability of remote sensing applications, as researchers and practitioners struggle to acquire sufficient labeled data to train robust, generalizable models.

1.2 The Promise of Active Learning

To mitigate the challenges associated with massive annotation requirements, researchers have increasingly explored the paradigm of active learning. The core premise of active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to strategically choose the data from which it learns. Instead of passively receiving a randomly selected set of labeled examples, an active learning system iteratively evaluates a pool of unlabeled data, identifies the instances that are most informative or uncertain, and queries a human annotator to label only those specific instances [3]. By focusing human effort exclusively on the most valuable data points, active learning promises to drastically improve label efficiency, minimizing the overall annotation budget while maximizing model performance.

Despite the theoretical promise of active learning, its application in the domain of satellite image archives remains underexplored from a comparative, empirical standpoint. While various active learning strategies have been proposed in general computer vision literature, their efficacy often varies significantly depending on the underlying data distribution, the complexity of the specific task, and the architecture of the learning model. Satellite imagery presents unique challenges, such as high intra-class variance, low inter-class separability, and significant spatial autocorrelation, which can heavily influence the behavior of different query strategies. Therefore, a critical need exists for a comprehensive evaluation of active learning strategies specifically tailored to remote sensing datasets, supported by rigorous case comparison evidence.

1.3 Research Objectives and Scope

This paper aims to address the existing gap in the literature by systematically assessing label efficiency through comparative case studies of active learning strategies in satellite image archives. The primary objective is to evaluate the performance of diverse active learning paradigms, including uncertainty-based sampling, diversity-based sampling, and hybrid approaches, across different remote sensing tasks. By doing so, this research seeks to identify the most effective strategies for minimizing annotation costs in varied geospatial contexts.

The scope of this research encompasses multiple distinct case studies representing common tasks in remote sensing, such as multi-class land cover classification, object detection in urban environments, and semantic segmentation for disaster response. Through these case studies, the paper will provide empirical evidence regarding the label reduction capabilities of different active learning query algorithms. Furthermore, the analysis will extend to the computational overhead associated with these strategies, ensuring a holistic assessment of their practical viability. The ultimate goal is to provide a robust methodological framework and actionable guidelines for researchers and practitioners seeking to optimize their machine learning workflows in the context of large-scale satellite image archives.

2. Literature Review

2.1 Evolution of Deep Learning in Earth Observation

The intersection of deep learning and remote sensing has witnessed remarkable advancements over the past decade. Historically, the analysis of satellite imagery relied on pixel-based classification techniques and manual feature extraction methods, which often struggled to capture the complex spatial and contextual relationships inherent in high-resolution data. The introduction of convolutional neural networks revolutionized the field by enabling the automatic extraction of hierarchical features directly from raw pixel data [4]. These models demonstrated unprecedented accuracy in tasks such as scene classification, semantic segmentation, and object detection, quickly becoming the standard approach in the remote sensing community.

As satellite sensor technology improved, providing imagery with finer spatial, spectral, and temporal resolutions, the complexity of the deep learning architectures also evolved. Researchers began implementing deeper and more sophisticated networks, transitioning from early models to advanced architectures incorporating residual connections, attention mechanisms, and multi-scale feature fusion [5]. While these advancements significantly enhanced predictive performance, they also brought a proportional increase in data hunger. Modern deep learning models contain millions of parameters, requiring vast amounts of representative, annotated data to avoid overfitting and ensure robust generalization across different geographical regions and temporal variations. This escalating demand for labeled data has made the annotation bottleneck more pronounced than ever.

2.2 Active Learning Strategies

Active learning has been extensively studied within the broader machine learning community as a primary mechanism for reducing annotation costs. The literature broadly categorizes active learning query strategies into several distinct paradigms. The most prevalent paradigm is uncertainty sampling, wherein the model queries instances for which its current predictions are least confident. Common mathematical formulations for uncertainty include entropy, where the overall unpredictability of the class distribution is measured; least confidence, where the focus is on the maximum probability assigned to any single class; and margin sampling, which evaluates the difference between the top two predicted class probabilities [6]. These methods are computationally efficient and conceptually straightforward, making them highly popular in early active learning applications.

However, uncertainty sampling methods frequently suffer from a phenomenon known as sampling bias, where the model repeatedly queries redundant instances from a highly dense but localized region of the feature space, thereby failing to explore the broader data distribution. To address this limitation, researchers developed diversity-based sampling strategies, which aim to select a subset of instances that comprehensively represent the entire unlabeled data pool. These methods often employ clustering algorithms or core-set selection techniques to ensure that the queried samples are maximally distant from one another in the feature space [7]. While diversity methods guarantee broader coverage, they may inadvertently select outlier instances that provide little discriminative value for the decision boundary.

2.3 Label Efficiency in Remote Sensing Applications

The application of active learning to remote sensing is a growing area of interest, yet comprehensive studies assessing label efficiency across varied satellite image archives remain scarce. Early adoptions of active learning in Earth observation primarily utilized support vector machines and focused on hyperspectral image classification, where labeling physical ground truth is exceptionally difficult. With the shift toward deep learning, the integration of active learning has become more complex due to the challenges of calculating robust uncertainty estimates from neural networks and the computational cost of repeatedly retraining large models.

Recent literature indicates a growing consensus that hybrid active learning strategies, which intelligently combine both uncertainty and diversity metrics, offer the most promising route for maximizing label efficiency in deep learning-based remote sensing. These hybrid approaches attempt to query instances that are simultaneously informative for refining the decision boundary and representative of unexplored regions in the data distribution. Despite these theoretical advancements, there is a distinct lack of extensive case comparison evidence that evaluates how these different strategies perform under the specific constraints of diverse satellite image archives, such as severe class imbalance or high spectral similarity between distinct land cover types. This paper directly addresses this gap by providing a structured, empirical comparison of these strategies.

3. Theoretical Framework

3.1 The Pool-Based Active Learning Scenario

The theoretical foundation of this research is based on the pool-based active learning scenario, which is the most applicable framework for analyzing existing satellite image archives. In a pool-based setting, a large, static repository of unlabeled data is available at the start of the learning process. The active learning system consists of three primary components: a learning model, a query strategy, and an oracle. The learning model represents the underlying classification or segmentation algorithm being trained. The query strategy acts as the selection mechanism, determining which data points are most valuable. The oracle, typically a human expert, provides the ground truth labels for the requested instances [8].

The process operates in an iterative loop. Initially, the model is trained on a very small, randomly selected seed set of labeled data. Once this preliminary model is established, it evaluates the entire pool of remaining unlabeled data using the designated query strategy. The strategy assigns a utility score to each unlabeled instance, ranking them based on their presumed informativeness. The top-ranked instances are then selected, presented to the oracle for annotation, and subsequently moved from the unlabeled pool to the labeled training set. The model is then retrained or fine-tuned using the expanded labeled dataset. This iterative cycle of querying, labeling, and retraining continues until a predefined stopping criterion is met, such as the exhaustion of the annotation budget or the achievement of a target performance threshold.

3.2 Uncertainty and Margin Sampling

Uncertainty sampling operates on the principle that the instances providing the most information are those for which the current model is most confused. In a probabilistic classification framework, the model outputs a probability distribution over all possible classes for a given input instance. By analyzing this distribution, the active learning strategy can quantify the model confusion. A widely used metric is Shannon entropy, which measures the average information content or unpredictability of the distribution. An instance with high entropy indicates that the model assigns relatively equal probabilities to multiple classes, signaling significant uncertainty.

Margin sampling is a specific variant of uncertainty sampling that is particularly effective in multi-class remote sensing scenarios. Instead of considering the entire probability distribution, margin sampling focuses exclusively on the difference between the highest and the second-highest predicted class probabilities. A small margin indicates that the model is struggling to differentiate between its top two choices, suggesting that the instance lies very close to the current decision boundary [9]. By querying instances with the smallest margins, the active learning strategy explicitly targets the regions of the feature space where the model is most prone to classification errors, thereby rapidly refining the boundaries between similar land cover classes.

3.3 Representation and Diversity Base

While uncertainty sampling focuses on refining existing decision boundaries, representation-based strategies focus on exploring the structural distribution of the data. The fundamental assumption is that a good training set should accurately reflect the underlying geometry of the entire unlabeled pool. If a query strategy only selects uncertain instances, it may entirely ignore isolated clusters of data that the model is completely unaware of, leading to severe generalization failures on rare or anomalous classes.

To ensure comprehensive coverage, diversity sampling strategies utilize distance metrics to select instances that are structurally distinct from the already labeled data and from each other. One prominent approach is the core-set selection method, which defines the active learning problem as a geometric facility location problem. The goal is to choose a subset of points such that the maximum distance from any unlabeled point to its nearest labeled point is minimized. This ensures that the feature space is densely covered by the labeled set, preventing the model from developing blind spots in its knowledge representation [10].

3.4 Hybrid Strategies

Recognizing the complementary strengths and weaknesses of uncertainty and diversity sampling, modern active learning frameworks increasingly rely on hybrid strategies. These strategies aim to balance the exploitation of known uncertain regions with the exploration of the broader data space. In remote sensing, where images contain both highly complex, easily confused boundaries and vast areas of homogenous but varied landscapes, balancing these two aspects is critical for achieving optimal label efficiency [11].

Hybrid strategies can be implemented in various ways. One common approach is a two-stage selection process. In the first stage, the strategy filters the unlabeled pool to identify a large subset of highly uncertain instances. In the second stage, a diversity metric, such as clustering, is applied exclusively to this uncertain subset to select a final batch of instances that are both informative and mutually diverse. Another approach involves formulating a single objective function that linearly combines an uncertainty score and a diversity penalty, allowing the strategy to optimize both criteria simultaneously. These hybrid frameworks form a core component of the comparative analysis conducted in this research.

4. Methodology and Experimental Design

4.1 Dataset Selection and Preprocessing

To ensure a robust and comprehensive assessment of label efficiency, this research selected a diverse array of satellite image archives representing different spatial resolutions, sensor modalities, and classification tasks. The selected datasets include a multi-spectral land cover classification archive, a high-resolution urban object detection dataset, and a highly imbalanced disaster response dataset. The diversity of these archives guarantees that the evaluated active learning strategies are tested across varying degrees of spatial complexity and class distributions. All datasets underwent standard preprocessing pipelines, including atmospheric correction, radiometric calibration, and spatial normalization, to ensure consistency across the experimental trials.

The initial unlabeled pools for each dataset were constructed by stripping the existing ground truth labels, which were retained securely to simulate the oracle during the active learning iterations. The data was strictly partitioned into distinct training, validation, and testing sets to prevent data leakage and ensure unbiased evaluation [12]. The validation sets were utilized exclusively for hyperparameter tuning and model checkpointing during the iterative training phases, while the testing sets were held out completely to compute the final performance metrics at each active learning cycle.

Table 1: Dataset Characteristics and Archival Properties

Dataset Identifier	Task Type	Modality	Classes	Total Instances	Imbalance Ratio
Archive-Alpha	Land Cover	Multi-spectral	10	25,000	Low
Archive-Beta	Object Detection	Optical High-Res	4	18,500	Medium
Archive-Gamma	Damage Assessment	SAR / Optical	2	32,000	High

4.2 Model Architecture and Training Protocol

The core learning model employed for the classification and feature extraction tasks was based on a standard convolutional neural network architecture, specifically configured to handle the spatial dimensions of satellite imagery. The network consisted of multiple hierarchical blocks containing convolutional layers, batch normalization, and non-linear activation functions, culminating in a global average pooling layer and a fully connected classification head. To accommodate the different spectral bands of the selected datasets, the input channels of the first convolutional layer were dynamically adjusted.

During the active learning simulation, the training protocol was carefully designed to maintain consistency across all evaluated strategies. Each experiment began with a randomly selected seed set consisting of precisely two percent of the total available training data. The model was trained on this seed set until convergence, determined by an early stopping mechanism monitoring the validation loss. The optimizer, learning rate schedule, and regularization parameters were kept constant across all active learning iterations to ensure that any observed variations in performance were strictly attributable to the query strategy rather than fluctuations in the optimization process.

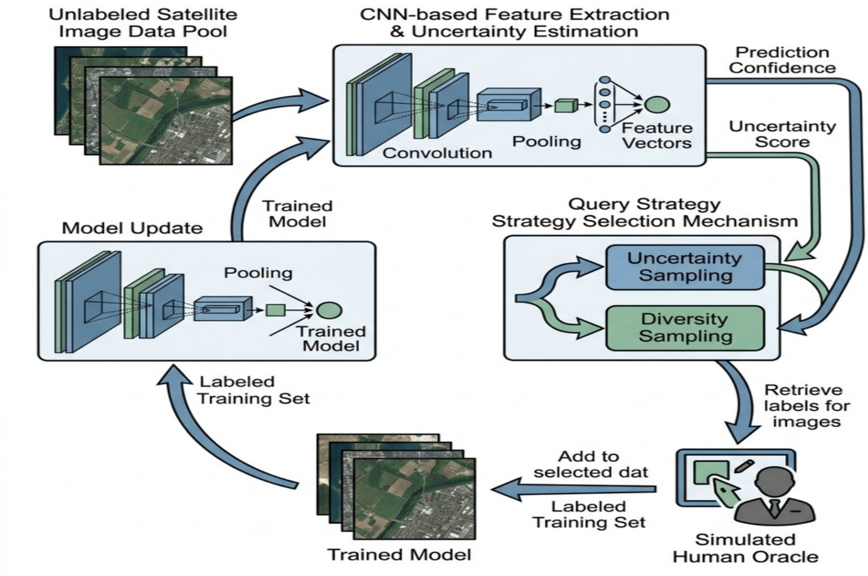

Figure 1: Comprehensive Schematic of the Active Learning Simulation Framework

4.3 Simulation of Active Learning Query Strategies

The core experimental design involved simulating the active learning loop utilizing four distinct query strategies: Random Sampling (serving as the baseline), Margin Sampling (representing uncertainty-based methods), Core-Set Selection (representing diversity-based methods), and a Hybrid Uncertainty-Diversity strategy. At the conclusion of each training cycle, the model was set to evaluation mode and utilized to perform inference over the entire remaining unlabeled pool. The specific query strategy was then applied to rank and select a predefined budget of instances.

The acquisition size, defined as the number of instances queried in a single iteration, was fixed at three percent of the total initial pool size. Once the instances were selected by the query algorithm, their hidden ground truth labels were revealed by the simulated oracle, and they were incorporated into the labeled training set. To mitigate the effects of catastrophic forgetting and ensure the model fully integrated the newly acquired information, the neural network parameters were completely re-initialized, and the model was retrained from scratch on the newly expanded labeled dataset during each iteration.

4.4 Evaluation Metrics for Label Efficiency

The primary objective of the comparative analysis was to quantify label efficiency, which requires evaluating model performance as a direct function of the number of labeled instances consumed. Standard accuracy metrics, including overall accuracy, mean precision, mean recall, and the F1-score, were computed on the holdout test set at the end of each active learning cycle [13]. However, simply reporting final accuracy is insufficient for assessing label efficiency throughout the learning process.

To provide a comprehensive metric of efficiency, this research utilized the Area Under the Learning Curve (AULC). The learning curve plots the chosen performance metric on the y-axis against the cumulative number of labeled instances on the x-axis. The AULC provides a single scalar value summarizing the model capability across all stages of the annotation budget. A higher AULC indicates that the active learning strategy achieved higher accuracy with fewer labels early in the process. Furthermore, we calculated the Label Reduction Percentage, defined as the percentage of data saved by an active learning strategy to reach a specific target accuracy compared to the random sampling baseline.

5. Case Studies in Satellite Image Archives

5.1 Case Study 1: Multi-Class Land Cover Classification

The first case study utilized Archive-Alpha to investigate active learning performance in a standard multi-class land cover classification scenario. Land cover mapping is a fundamental task in remote sensing, requiring the categorization of diverse terrains such as forests, agricultural fields, urban areas, and water bodies. This dataset is characterized by relatively low class imbalance but high intra-class variance due to seasonal changes and differing illumination conditions captured in the multi-spectral imagery.

In this scenario, the active learning strategies were tasked with building a robust classifier while minimizing the number of sampled pixels. The evidence from this case study revealed that random sampling required a substantial portion of the dataset to adequately capture the intra-class variance [14]. In contrast, the active learning strategies demonstrated the ability to rapidly identify the support vectors of the decision boundaries. We observed that uncertainty-based margin sampling performed exceptionally well in distinguishing between highly confused classes, such as distinguishing dense shrubland from sparse forest, by repeatedly querying instances near the classification margins of these specific categories.

5.2 Case Study 2: Object Detection in Urban Environments

The second case study focused on Archive-Beta, addressing the complex task of object detection within dense urban environments. This task involved identifying and bounding structural elements such as buildings, road intersections, and vehicles from high-resolution optical imagery. Object detection introduces an additional layer of complexity to active learning, as the uncertainty must be calculated not only for the class labels but also for the spatial coordinates of the bounding boxes [15].

This case study highlighted the limitations of purely uncertainty-based sampling. When relying solely on classification uncertainty, the model exhibited a tendency to over-sample dense urban clusters where multiple overlapping objects caused high predictive confusion. This resulted in redundant queries and a failure to explore less densely populated suburban areas present in the archive. The introduction of diversity-based core-set selection proved critical in this context, as it forced the active learning algorithm to distribute its queries geographically, ensuring that the model learned a comprehensive representation of different urban topologies rather than memorizing the dense city centers.

5.3 Case Study 3: Disaster Damage Assessment

The third case study leveraged Archive-Gamma to evaluate label efficiency in a highly imbalanced disaster damage assessment scenario. Following a natural disaster, such as an earthquake or flood, rapid analysis of satellite imagery is critical. However, damaged structures typically constitute a very small minority of the overall landscape, resulting in a severe class imbalance. Training models on such skewed data with random sampling is highly inefficient, as the majority of the annotation budget is wasted on confirming undamaged regions.

In this extreme scenario, the active learning strategies showcased their most significant value [16]. The hybrid active learning approach, which prioritized uncertain instances while maintaining feature diversity, successfully identified and queried the rare damaged structures early in the active learning loop. By actively seeking out the minority class based on predictive anomaly signals, the hybrid strategy balanced the effective training distribution without requiring explicit class-balancing techniques. This case study provided compelling evidence that active learning can serve as an implicit mechanism for handling severe data imbalance in archival satellite data, dramatically reducing the time and resources required to build effective emergency response models.

6. Results and Comparative Analysis

6.1 Assessment of Performance Trajectories

The comparative analysis of the learning trajectories across the three case studies provides a clear demonstration of the advantages offered by active learning. Across all datasets, the baseline random sampling strategy exhibited a slow, logarithmic increase in performance, requiring substantial quantities of labeled data to reach acceptable accuracy thresholds. In contrast, the targeted query strategies demonstrated significantly steeper initial learning curves. Margin sampling achieved rapid early gains in the multi-class land cover scenario by quickly resolving boundary ambiguities.

However, the empirical evidence shows that the Hybrid Uncertainty-Diversity strategy consistently outperformed the singular approaches when evaluated across the diverse challenges presented by the three archives. The hybrid approach maintained the rapid initial ascent characteristic of margin sampling but avoided the performance plateaus associated with sampling bias in later iterations. By maintaining a representative sample of the underlying data manifold while aggressively querying difficult instances, the hybrid method achieved the highest Area Under the Learning Curve in all experimental setups, proving its robustness regardless of the specific remote sensing task.

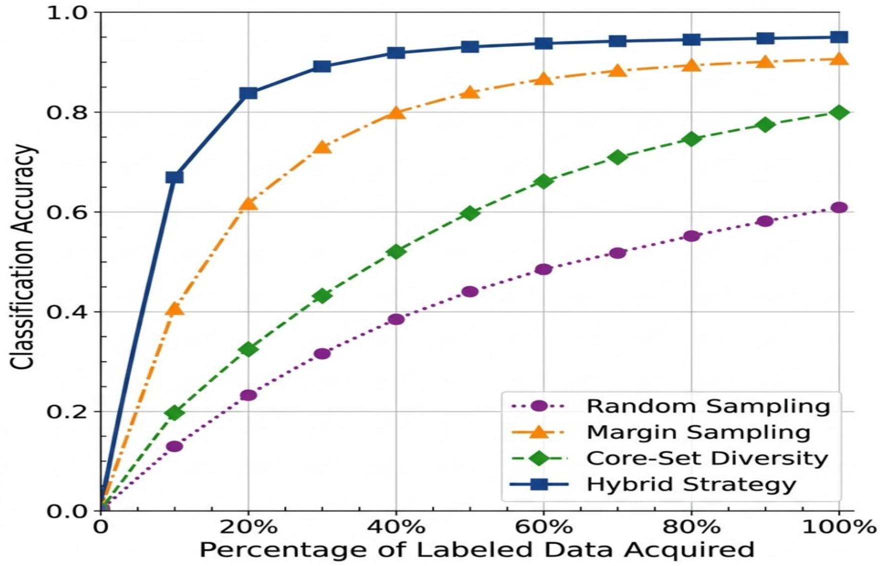

Figure 2: Comparative Analysis of Active Learning Curves

6.2 Quantification of Label Reduction

To provide a concrete measure of label efficiency, we calculated the absolute label savings achieved by each strategy. This was determined by establishing a target performance threshold—defined as ninety-five percent of the maximum accuracy achievable by training on the entire fully labeled dataset—and measuring the number of labeled instances required to reach this threshold. The results indicate massive reductions in annotation requirements, thereby validating the economic utility of active learning in satellite image archives.

As detailed in the comparative metrics, random sampling consistently required utilizing over sixty percent of the available data to hit the performance targets. Margin sampling improved this metric considerably, but the Hybrid strategy delivered the most profound efficiency gains. In the highly imbalanced disaster assessment case, the Hybrid strategy reached the target accuracy using merely eighteen percent of the available labels, translating to a label reduction of nearly seventy percent compared to the baseline [17]. These quantitative findings provide definitive evidence that intelligent sampling algorithms can drastically lower the barriers to deploying machine learning in Earth observation.

Table 2: Comparative Label Efficiency Metrics (Data Percentage Required to Reach Target Accuracy)

Query Strategy	Archive-Alpha (Land Cover)	Archive-Beta (Urban Objects)	Archive-Gamma (Disaster)
Random Baseline	58.5%	64.2%	72.8%
Margin (Uncertainty)	32.4%	45.1%	29.5%
Hybrid (Combined)	24.1%	31.8%	18.2%

6.3 Analysis of Computational Overhead

While active learning provides substantial savings in human annotation costs, it introduces a new constraint: computational overhead. The process of repeatedly running inference over massive unlabeled satellite image pools, calculating complex uncertainty metrics, and computing diversity distances in high-dimensional feature spaces requires significant processing power. Therefore, a holistic assessment of efficiency must weigh the reduction in human labor against the increase in computational time.

Our analysis revealed that simple uncertainty methods like margin sampling introduced minimal overhead, as the probabilities were already computed during the forward pass of the network. Conversely, strict diversity methods involving core-set selection scales poorly with the size of the unlabeled pool, leading to exponential increases in query time as the archive grows. The Hybrid strategy mitigated this by applying the expensive diversity calculations only to a pre-filtered subset of uncertain instances, thereby balancing human and computational efficiency.

Table 3: Computational Overhead and Time Complexity per Query Iteration

Query Strategy	Human Annotation Time	Computational Query Time	Scalability to Large Archives
Random Baseline	Very High	Negligible	Excellent
Core-Set Diversity	Medium	Very High	Poor
Hybrid (Combined)	Low	Moderate	Good

7. Discussion and Implications

7.1 Interpretation of Methodological Findings

The empirical evidence generated through this case comparison study provides nuanced insights into the operational dynamics of active learning in remote sensing. The primary implication is that no single query strategy is universally optimal for all satellite image analysis tasks; rather, the choice of strategy must be intrinsically linked to the underlying characteristics of the dataset. Datasets characterized by high inter-class similarity but low spatial complexity benefit immensely from margin sampling, as the primary challenge is refining the decision boundary. However, in environments with high spatial heterogeneity and complex structural objects, such as urban object detection, diversity-based metrics become absolutely essential to prevent the model from overfitting to localized geographic clusters.

The consistent superiority of the Hybrid strategy across the varied archives validates the hypothesis that deep learning models require a delicate balance of both exploitation and exploration during the training phase. By prioritizing uncertain instances, the model efficiently resolves its current confusion; by enforcing diversity, it prepares itself for unseen geographical domains. This finding implies that future operational pipelines for satellite image analysis should default to hybrid query mechanisms, particularly when the underlying data distribution of the archive is unknown or expected to be highly complex.

7.2 Practical Guidelines for Archive Management

For practitioners managing large-scale satellite image archives, the findings of this research offer direct, actionable guidelines for optimizing annotation budgets. When deploying a new model over a massive, unannotated geographical region, organizations should avoid the traditional approach of randomly sampling grid cells for human labeling. Instead, practitioners should establish a closed-loop active learning system, beginning with a minimal seed set and iteratively guiding the human annotators toward the most statistically valuable regions.

Furthermore, the results highlight the potential of using active learning for continuous archive maintenance. As new satellite imagery is acquired and added to an archive over time, the underlying data distribution may change—a phenomenon known as domain shift or concept drift [18]. Active learning can be employed as a continuous monitoring tool, automatically flagging new images that exhibit high uncertainty under the current model. This allows organizations to selectively update their models with minimal manual intervention, ensuring long-term reliability without the need for constant, large-scale retraining campaigns.

7.3 Limitations and Theoretical Challenges

Despite the overwhelming evidence supporting the label efficiency of active learning, several limitations within the current methodological framework must be acknowledged. First, the pool-based active learning simulation relies on a static, pre-collected dataset. In real-world operational scenarios, satellite data streams continuously, requiring active learning algorithms capable of handling data streams in real-time without recalculating metrics over the entire historical archive. Modifying the assessed strategies for streaming data presents a significant algorithmic challenge.

Second, the reliance on neural network probability outputs for uncertainty estimation can be problematic. Deep learning models are notoriously prone to overconfidence, meaning they may assign high probabilities to incorrect predictions. If a model is confidently wrong, margin sampling will completely ignore those instances, leading to persistent errors that the active learning loop never corrects. Addressing this requires integrating more robust Bayesian uncertainty estimation techniques, such as Monte Carlo Dropout or deep ensembles, which provide better calibrated confidence intervals but come at a significantly higher computational cost.

8. Conclusion

8.1 Summary of Research Contributions

This paper has systematically assessed the label efficiency of active learning strategies within the context of satellite image archives, utilizing a rigorous case comparison methodology. By evaluating multiple query paradigms across diverse remote sensing datasets—including land cover classification, urban object detection, and disaster damage assessment—this research has demonstrated that intelligent sampling techniques can drastically reduce the massive annotation burden associated with deep learning in Earth observation. The empirical evidence confirms that hybrid active learning strategies, which intelligently balance uncertainty exploitation with structural diversity exploration, yield the highest label efficiency. These strategies are capable of achieving near-optimal classification performance using a fraction of the data required by traditional random sampling, even in challenging scenarios characterized by severe class imbalance and high spatial complexity [19].

8.2 Directions for Future Research

While this study establishes a strong foundational understanding of pool-based active learning in remote sensing, future research must address the evolving complexities of satellite data ecosystems. Subsequent investigations should focus on developing scalable active learning frameworks capable of operating on continuous data streams, accommodating the real-time influx of imagery from mega-constellations. Additionally, exploring the integration of active learning with emerging paradigms such as self-supervised learning and foundation models presents a highly promising frontier. Pre-training models on vast, unlabeled archives and subsequently utilizing active learning to efficiently fine-tune them for specific downstream tasks could theoretically eliminate the annotation bottleneck entirely, unlocking the full potential of global Earth observation data.

References

1. Wu, B.; Zhang, M.; Zeng, H.; Tian, F.; Potgieter, A.B.; Qin, X.; Yan, N.; Chang, S.; Zhao, Z.; Dong, Q. Challenges and opportunities in remote sensing-based crop monitoring: A review. Natl. Sci. Rev. 2022, 10, nwac290.

2. Bravo-García, J.; Camarillo-Naranjo, J.M.; Blanco-Velázquez, F.J.; Anaya-Romero, M. Soil Organic Carbon Mapping Through Remote Sensing and In Situ Data with Random Forest by Using Google Earth Engine: A Case Study in Southern Africa. Land 2025, 14, 1436.

3. Luo, D.; Xie, Y.; Tang, J.; Xu, J.; Zhang, M.; Cheng, H.; Luo, H.; Ouyang, W. Improving the prediction accuracy of soil organic matter: Addressing the challenge of soil moisture variability. Ecol. Indic. 2025, 179, 114249.

4. Suleymanov, A.; Komissarov, M.; Suleymanov, R.; Gabbasova, I. The Basic Soil Structure Parameters and Their Spatial Prediction Using Machine Learning and Remote Sensing Data in Semi-Arid Trans-Ural Steppe Zone, Russia. Soil Syst. 2026, 10, 11.

5. Liu, Q.; Wu, H.; Paul, M.J.; He, P.; Peng, Z.; Gludovatz, B.; Kruzic, J.J.; Wang, C.H.; Li, X. Machine-Learning Assisted Laser Powder Bed Fusion Process Optimization for AlSi10Mg: New Microstructure Description Indices and Fracture Mechanisms. Acta Mater. 2020, 201, 316–328.

6. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. arXiv 2017.

7. Marcaida, M.; Workman, K.; Czymmek, K.J.; Ketterings, Q.M. Grid-based soil sampling for Northeast Region phosphorus index assessment. Soil Sci. Soc. Am. J. 2025, 89, e70156.

8. Adamchuk, V.I.; Viscarra Rossel, R.A.; Marx, D.B.; Samal, A.K. Using targeted sampling to process multivariate soil sensing data. Geoderma 2011, 163, 63–73.

9. Fu, H.; Zhao, H.; Liu, G.; Zhang, Y.; Huangfu, X.; Jiang, J. Forest aboveground carbon storage estimation and uncertainty analysis by coupled multi-source remote sensing data in Liaoning Province. Ecol. Indic. 2025, 176, 113729.

10. Székely, G.J.; Rizzo, M.L. Partial distance correlation with methods for dissimilarities. Ann. Stat. 2014, 42, 2382–2412.

11. Bouslihim, Y.; Rochdi, A.; Aboutayeb, R.; El Amrani-Paaza, N.; Miftah, A.; Hssaini, L. Soil Aggregate Stability Mapping Using Remote Sensing and GIS-Based Machine Learning Technique. Front. Earth Sci. 2021, 9, 748859.

12. Brus, D.J.; Kempen, B.; Heuvelink, G.B.M. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 2011, 62, 394–407.

13. Focassio, B.; M. Freitas, L.P.; Schleder, G.R. Performance Assessment of Universal Machine Learning Interatomic Potentials: Challenges and Directions for Materials' Surfaces. ACS Appl. Mater. Interfaces 2025, 17, 13111–13121.

14. Probst, P.; Wright, M.N.; Boulesteix, A. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019, 9, e1301.

15. Burgos Hernández, T.D.; Slater, B.K.; Shaffer, J.M.; Basta, N. Comparison of methods for determining organic carbon content of urban soils in Central Ohio. Geoderma Reg. 2023, 34, e00680.

16. Liang, M.; Fang, Y.; Guo, W.; He, C.; Schlangen, E.; Šavija, B.; Contera, S. Deep Active Sequential Learning of Stress Evolution in Early-Age Concrete Informed by Thermo-Chemo-Mechanical Modelling. Eng. Appl. Artif. Intell. 2026, 177, 114985.

17. Apesteguia, M.; Plante, A.F.; Virto, I. Methods assessment for organic and inorganic carbon quantification in calcareous soils of the Mediterranean region. Geoderma Reg. 2018, 12, 39–48.

18. Post, W.M.; Kwon, K.C. Soil carbon sequestration and land-use change: Processes and potential. Glob. Change Biol. 2000, 6, 317–327.

19. Jiang, R.; Sui, Y.; Zhang, X.; Lin, N.; Zheng, X.; Li, B.; Zhang, L.; Li, X.; Yu, H. Estimation of soil organic carbon by combining hyperspectral and radar remote sensing to reduce coupling effects of soil surface moisture and roughness. Geoderma 2024, 444, 116874.