Active Learning for Annotation of Burn Scars on Satellite Imagery

Saugat Adhikari
Indiana University
700N N Woodlawn Ave
adhiksa@iu.edu

Jiyeong Oh
Indiana University
700N N Woodlawn Ave
ohjiye@iu.edu

Yongjun Cho
Indiana University
700N N Woodlawn Ave
yc134@iu.edu

Yudai Hirata
Indiana University
700N N Woodlawn Ave
yuhirata@iu.edu

1. Introduction

Wildfires have become more frequent and destructive due to the combined effects of climate change and human activities. They cause massive damage to homes, infrastructure, and natural environments, and cost billions of dollars each year. Knowing exactly where burn scars are located is important for organizing recovery efforts and figuring out how to prevent future fires. Satellite imagery provides valuable data for mapping burn scars; however, the problem is that we lack sufficient labeled data to train machine learning models effectively. Manually labeling burn scars is slow, expensive, and prone to mistakes. This paper proposes a solution to this problem using an active learning approach, where the model and human experts work together to improve the quality of burn scar labels quickly.

2. Background and Related Work

There is a large gap between the availability of high-resolution satellite imagery and the availability of high-quality labeled data for burn scar detection. While organizations like NASA [1], USGS [9] and NOAA [4] provide abundant satellite images, creating accurate burn scar labels remains challenging because it requires manual effort. This is time-consuming, expensive, and often inconsistent. We show two examples of burn scars in Fig 1 and Fig 2. Fig 1 shows burn scars left behind in Maui in 2023, and Fig 2 shows that in Pacific Palisades in 2025, which look totally different. Thus, we need to develop an efficient method for generating accurate burn scar labels, not only to improve the quality of fire-related research but also to support decision-making for resource management and disaster response.

U-Net [8] is the first model to successfully apply a convolutional network for biomedical image segmentation. Since medical images require precise localization—where the output must correspond to specific labels—this approach is also beneficial for segmentation in other fields beyond medicine. The U-Net architecture features a unique U-shaped structure that includes both downsampling and upsampling processes.

Active learning [7] presents an effective solution to this problem by reducing the need for large annotated datasets. In an active learning framework, the model learns more efficiently by focusing on the most informative or uncertain samples, allowing for faster improvement with fewer human-labeled examples. This approach has the potential to accelerate the annotation process while improving the overall performance of burn scar segmentation models.

Training deep learning models for image segmentation relies on accurately annotated ground-truth labels. To generate these labels, annotation tools are essential for visualizing images and enabling manual annotation. One such tool is FloodTrace [3], designed specifically for annotating flood and dry regions in satellite imagery. It supports real-time 3D terrain visualization and allows users to annotate directly within a web browser.

3. Methods

This section details the approach used to address research questions related to the application of active learning to burn scar annotation.

3.1. Research Questions

The core research questions that guide this paper are:

1. Using active learning, can we create high-quality labels suitable for burn scar detection?
2. Can active learning reduce the annotation workload required for burn scars?
3. Can we achieve high accuracy in burn scar detection

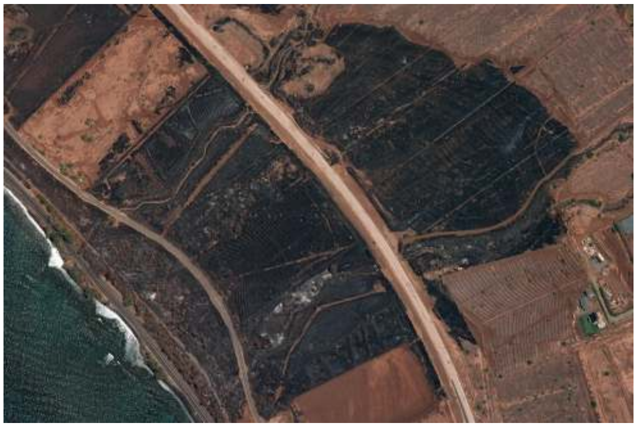


Figure 1. This image shows the burn scars from the 2023 Maui Fire. A large burn area is visible in the center, characterized by dark black and deep green tones. The burn area spans a wide region, highlighting the severity and reach of the wildfire.

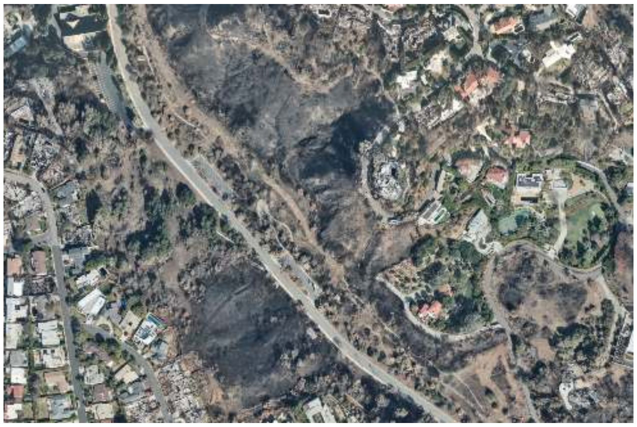


Figure 2. This image shows burn scars from the 2025 California Fire. The affected areas appear in dark black tones, indicating severe fire damage. This burn region is situated very close to a residential area, highlighting the intensity and potential danger of the wildfire.

models using the annotated labels generated through active learning?

3.2. Active Learning Method

There are four steps in our active learning.

Step 0. Train Model: This training phase will involve preprocessing the data and applying data augmentation techniques to make the model more robust to burn scars. We use the publicly available HLS burn scars dataset [6] to train an initial UNet model.

Active learning round is illustrated in Fig 3.

Step 1. Infer on Unlabeled Data: We begin by training a model that is used to generate burn scar predictions for new, unlabeled satellite images.

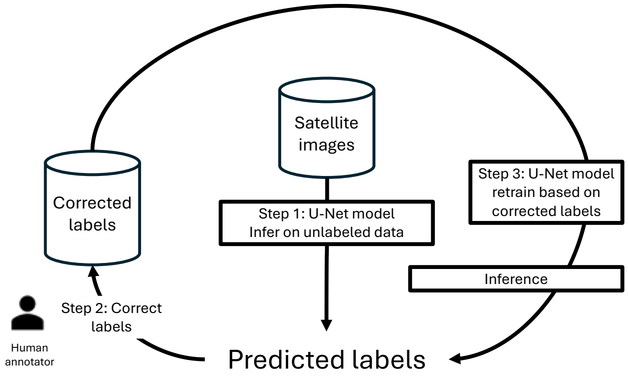

Satellite images
Corrected labels
Step 1: U-Net model Infer on unlabeled data
Step 3: U-Net model retrain based on corrected labels
Inference
Step 2: Correct labels
Human annotator
Predicted labels

Figure 3. This figure illustrates the Active Learning process. In Step 1, a satellite image is used as input to a U-Net model, which generates initial predicted labels. In Step 2, a human annotator corrects these predicted labels. In Step 3, the corrected labels are used to retrain the U-Net model for three epochs, after which the process loops back to Step 1 to re-infer the model and continues iteratively.

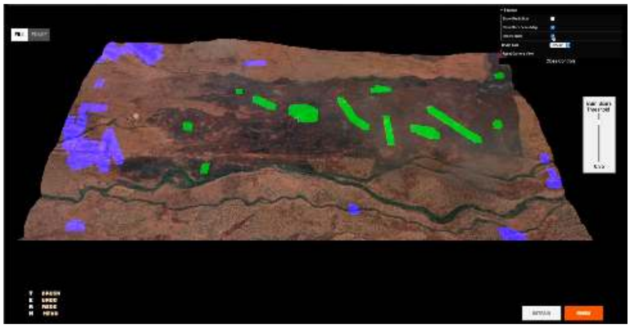


Figure 4. This figure shows the annotation tool interface. This tool was initially developed for flood annotation and we adapted it for burn scar annotation. This tool allows visualization of satellite image in 3D and perform the annotation. For annotation, we can either add labels using brush or erase the wrong labels. In the figure, green marks represents burn scar annotation and blue marks represent not burn scar annotation. These annotation labels are used to train the underlying U-Net model and the prediction can be visualized in the tool itself.

Step 2. Correct Labels: Human annotators will review the model's predictions using a web-based front-end annotation tool [3]. The annotator will correct any errors by adding the label ”burn scars” to the region or removing the label ”not burn scars” from regions, as shown in the sample screen in Fig 4

Step 3. Retrain based on Corrected Labels: The corrected labels, which include the model's predictions that humans do not modify, will be fed back into the trained model. The process will then be retrained for three epochs, returning to inference the model. This round will continue to enhance the model's performance through an active learning loop. We do this round until accuracy reaches 90%.

4. Evaluation

4.1. User Study

To evaluate the effectiveness of the proposed active learning approach, we conducted a user study. This study compared user performance on an annotation task under two conditions: annotation with active learning and annotation without active learning (baseline).

4.1.1. Participants

Four graduate students who are taking computer vision class and one Computer Science PhD student were initially recruited for the study, comprising four students who participated in this paper. Data from one participant had to be excluded due to exported errors identified in their assigned dataset during the baseline condition, leaving us with data from 4 participants.

4.1.2. Evaluation Protocol

Four participants, who also took a computer vision class, underwent manual annotation (baseline annotation) on the first day. After a period of more than one day, we conducted the annotation process using an active learning approach. Each participant annotated two regions, and the annotation time was manually recorded by each individual using the stopwatch function on their phone.

For one Ph.D. student, we began by explaining the purpose of our tasks, followed by an initial annotation of two regions. We then asked the participant to annotate the same regions again using active learning. We manually recorded the time using the stopwatch function on our phone. After completing the annotations, the participant contributed by asking questions about the insights of the annotation tools.

4.1.3. Datasets

A common machine learning approach for annotating burn scars is to divide a large, high-resolution image into smaller patches. Since ground truth labels are unavailable for this new dataset, we annotated a small subset of regions. We use the data sets from the California fire [4] and the Maui fire [5] illustrated in table 1. These new images differ from the training data [6].

Table 1. Database Image Source

Region	Number of images
Maui	Total: 6, Use: 6
California	Total: 5, Use: 4

4.1.4. Resources

We utilized the M-series Apple MacBook Air or MacBook Pro, as training each round requires CPU power.

4.1.5. Evaluation Metrics

We evaluate the model's performance iteratively, before and after each round of active learning. We used initial training as a ground truth and measured the accuracy of each round of the active learning by using Precision, Recall, and F1-score.

4.1.6. Additional Evaluation Details

In order to asses active learning labels data's effectiveness, we trained the U-Net again using the annotated labels generated through active learning.

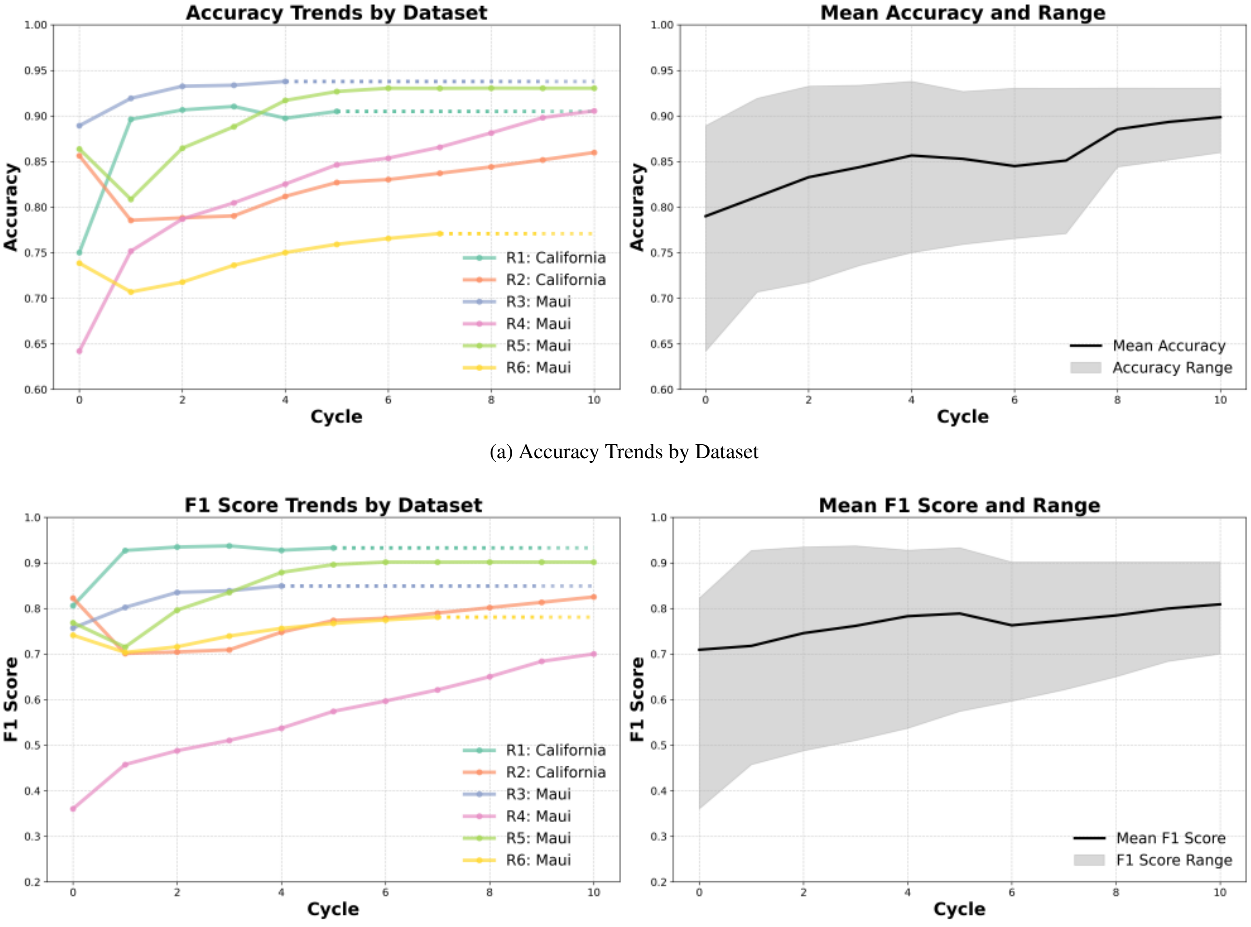
5. Results

5.1. Quantitative Performance Evaluation

First, we quantitatively evaluated the impact of active learning on the performance of the burn scar detection model. Fig 5 shows that, as the active learning cycles progressed, both the accuracy and F1 score steadily improved across all experimental regions (California and Maui). Most experimental groups reached over 90% accuracy, and even those starting with lower performance, particularly in the Maui region, exhibited consistent gains. These trends suggest that the iterative correction of imperfect predictions and retraining through active learning allowed the model's decision boundary to converge more closely with the true burn scar regions.

In addition, as illustrated in Figure 6, we compared the performance metrics of the initial baseline U-Net with those of the model after active learning. The results showed improvements not only in accuracy but also in other metrics (precision, recall, and F1 score) - specifically, precision increased from 0.80 to 0.90, recall from 0.69 to 0.79, and the F1 score from 0.71 to 0.83 for burn scar regions. This indicates that the model not only made more correct predictions overall, but also better captured the critical burn scar areas.

A detailed comparison of the performance between six test regions (California and Maui) is provided from Table 2. On all datasets, active learning consistently outperformed baseline U-Net in terms of overall accuracy and class-specific performance. Notably, even in more challenging cases, such as R4 (Maui), where the baseline model has a burn scar F1 score of just 0.36, active learning improves this metric to 0.72. These results confirm the generalizability and effectiveness of active learning in improving model performance across a variety of regional characteristics. Table 3 clearly points the benefits of the proposed approach by aggregating performance metrics regionally and globally. On average, active learning improves the overall accuracy from 0.79 to 0.886 and the F1 score for burn scars from 0.709 to 0.835. This consistent improvement in both California and Maui regions further strengthens the conclusion that active learning is a reliable and efficient method for



(a) Accuracy Trends by Dataset

(b) F1-Score Trends by Dataset

Figure 5. Performance progression of the burn scar detection model over active learning cycles. ((a): trend in accuracy, (b): F1-score trends) Both metrics demonstrate steady improvements with each cycle, indicating that active learning effectively enhances model performance across diverse regions.

improving burn scar detection models with limited manual labeling.

Precision-recall curve of Figure 7 further supports this improvement. The curve of the model trained with active learning consistently outperformed the baseline model, and the Area Under the Curve (AUC) also increased from 0.849 (baseline U-Net) to 0.871 (active learning).

5.2. Qualitative Performance Evaluation

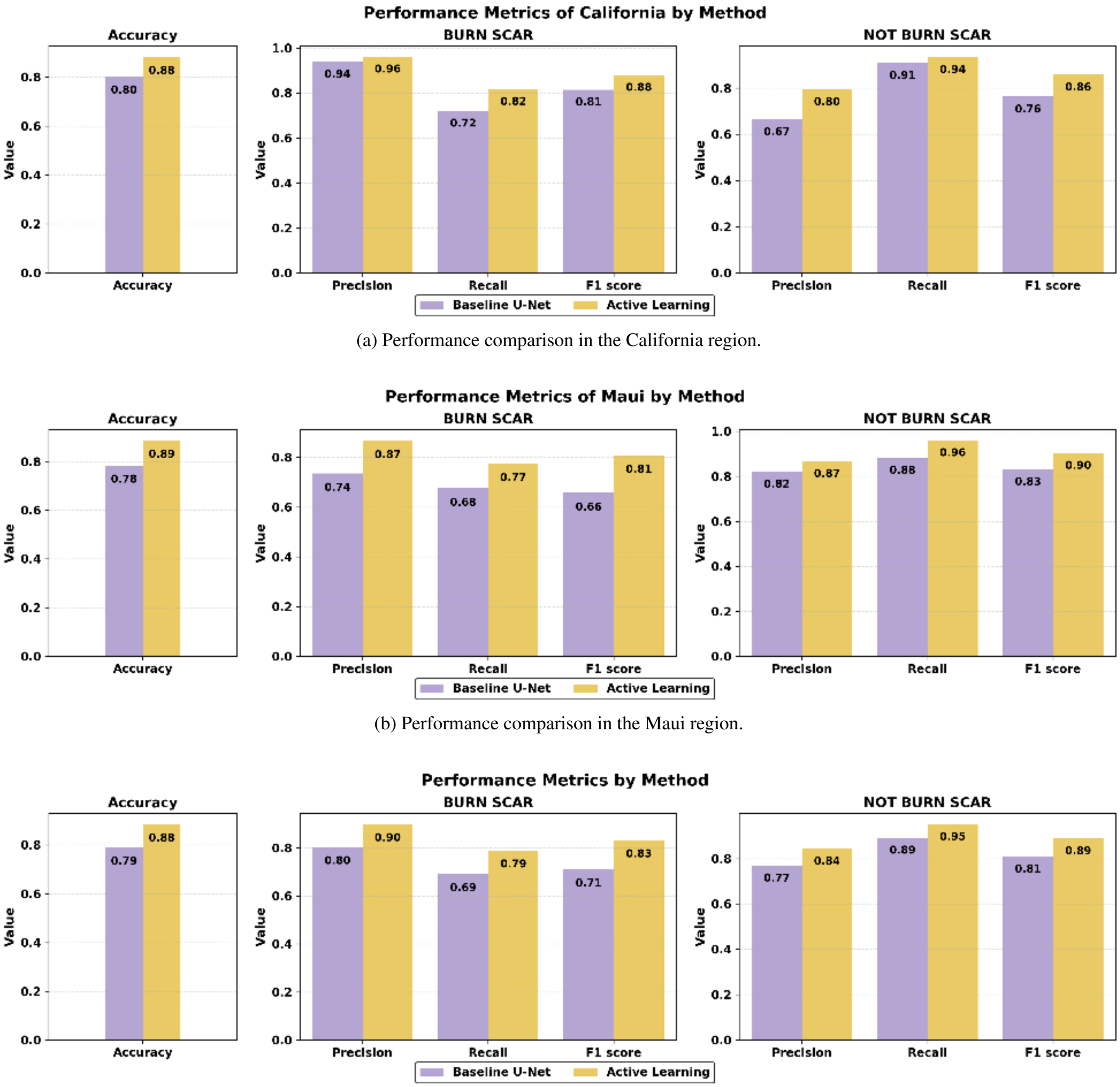
In addition to quantitative metrics, we also qualitatively analyzed the outputs of baseline U-Net and active learning reinforcement models. As shown in Figure 9, baseline models often created predictions that were highly influenced by the visual characteristics of images, including their brightness and texture. For example, dark-colored areas such as oceans, rivers, and shadows were often incorrectly classified as burn scars due to their similar hues to actual burn areas.

Active learning models, on the other hand, showed robustness to this visual ambiguity. By iteratively correcting prediction errors during training, the model learned to better distinguish burn scars based on contextual and spatial patterns without relying solely on pixel strength. Examples from Figure 9 show that the refined model can correct false positives and better localize burn scar areas, especially in complex environments such as coastal areas or mountainous terrain.

These qualitative results confirm the importance of human-in-the-loop corrections in reducing visually biased predictions, reinforcing the effectiveness of active learning that leads the model to more semantically accurate inference.

5.3. Annotation Efficiency and Effort Reduction

To further evaluate the practical benefits of active learning approaches, we analyzed their impact on annotation effi-



(a) Performance comparison in the California region.

(b) Performance comparison in the Maui region.

(c) Overall performance averaged across all test regions.

Figure 6. Bar charts comparing the performance of the baseline U-Net and the active learning model across three evaluation scopes: **(a)** California, **(b)** Maui, and **(c)** overall average. Each chart presents accuracy, precision, recall, and F1 score for both burn scar and non-burn scar classes. The active learning model consistently outperforms the baseline across all regions and evaluation metrics.

ciency from temporal and spatial perspectives.

First, we measured the annotation time required for each method. As summarized in Table 4, the average time taken to manually annotate a sample using traditional supervised learning was 19.75 minutes. On the other hand, the active learning approach took only 10.22 minutes, including additional steps such as reviewing model prediction, correcting segmentation labels, and retraining. This reduces the time by nearly 50%, indicagting that using active learning can significantly reduce human annotation tasks.

Second, we investigated the percentage of annotated pixels in all active learning cycles. As shown in Figure 8, the average corrected or labeled percentage of image pixels was only 18.5%. This result indicates the ability of active learning to achieve high performance while significantly reducing the annotation burden.

These findings show that active learning not only improves model accuracy, but also provides a more efficient

			BURN SCAR			NOT BURN SCAR		
Dataset	**Method**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
R1: California	Baseline U-Net	0.750	0.930	0.711	0.806	0.524	0.855	0.650
	Active Learning	**0.905**	**0.959**	**0.909**	**0.933**	**0.785**	**0.895**	**0.836**
R2: California	Baseline U-Net	0.856	0.946	**0.728**	0.823	**0.808**	0.965	0.879
	Active Learning	**0.860**	**0.962**	0.723	**0.825**	0.807	**0.976**	**0.883**
R3: Maui	Baseline U-Net	0.889	0.765	0.750	0.757	0.926	0.931	0.928
	Active Learning	**0.938**	**0.958**	**0.763**	**0.850**	**0.933**	**0.990**	**0.961**
R4: Maui	Baseline U-Net	0.642	0.241	0.713	0.360	0.930	0.630	0.752
	Active Learning	**0.913**	**0.665**	**0.780**	**0.718**	**0.963**	**0.935**	**0.935**
R5: Maui	Baseline U-Net	0.864	**0.954**	0.644	0.769	0.836	**0.983**	0.903
	Active Learning	**0.930**	0.897	**0.907**	**0.902**	**0.949**	0.943	**0.946**
R6: Maui	Baseline U-Net	0.738	**0.982**	0.595	0.741	0.588	**0.981**	0.736
	Active Learning	**0.771**	0.980	**0.650**	**0.781**	**0.622**	0.977	**0.760**

Table 2. Per-dataset performance comparison between the baseline U-Net and active learning. The table reports accuracy, precision, recall, and F1 score separately for burn scar and non-burn scar classes across six test regions (California and Maui). Active learning shows consistent performance improvements in most cases, particularly in challenging regions such as R4.

			BURN SCAR			NOT BURN SCAR		
Region	**Method**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
California	Baseline U-Net	0.803	0.938	0.720	0.814	0.666	0.910	0.764
	Active Learning	**0.882**	**0.960**	**0.816**	**0.879**	**0.796**	**0.935**	**0.860**
Maui	Baseline U-Net	0.783	0.735	0.676	0.657	0.820	0.881	0.830
	Active Learning	**0.888**	**0.875**	**0.775**	**0.813**	**0.867**	**0.961**	**0.901**
All	Baseline U-Net	0.790	0.803	0.690	0.709	0.769	0.891	0.808
	Active Learning	**0.886**	**0.903**	**0.789**	**0.835**	**0.843**	**0.953**	**0.887**

Table 3. Aggregated performance comparison across California, Maui, and all test regions. The results summarize the average accuracy and per-class metrics, demonstrating that active learning significantly improves model performance with less manual labeling. Consistent improvements across regions confirm the generalizability of the approach.

and scalable alternative to fully manual annotation, especially in domains where expert labeling is time-consuming and expensive.

Method	**Manual Annotating**	**Active Learning**
Annotation Time (min)	19.75	10.22

Table 4. Average annotation time per sample required by manual annotation and active learning approaches. The active learning framework reduces annotation time by nearly half, despite incorporating additional steps such as prediction review and model updates.

5.4. User Experiment for Active Learning Consistency Verification

In order to verify the effectiveness of the active learning technique more quantitatively, a user study was conducted to evaluate the consistency of results between users. In general, active learning includes the process of modifying the prediction result by a person, and the modifier is assumed to be an expert with an understanding of the domain. However, in actual situations, the annotation criteria may differ from person to person, which can lead to different prediction performance for the same data. Accordingly, in this study, an auxiliary experiment was conducted to check whether active learning works stably between users.

Two different users (A, B) participated in the experiment.

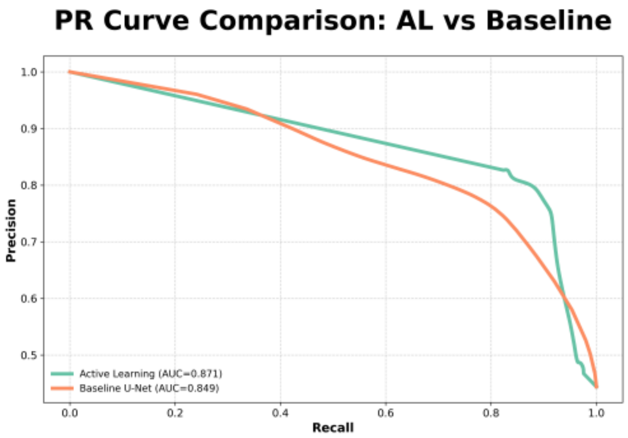


Figure 7. Precision–Recall (PR) curve comparing the baseline U-Net and the model trained with active learning. The curve illustrates that active learning yields higher precision across most recall levels, demonstrating more robust performance in identifying burn scar regions.

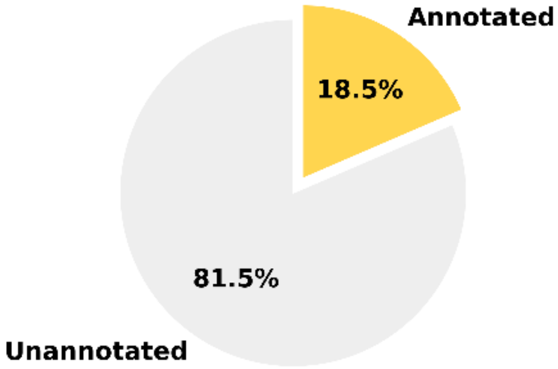


Figure 8. Proportion of annotated pixels during the active learning process. On average, only 18.5% of pixels were manually corrected or labeled across all cycles, indicating a substantial reduction in annotation effort compared to fully supervised labeling.

Among them, user B is an external participant who has no prior knowledge of the structure and background of this project, and user A is a user who has some experience participating in the project. Two users performed manual annotation (baseline) and active learning-based annotation on the same image dataset, respectively. Thereafter, the model was trained based on the labels generated from each user and the prediction results were compared.

Table 5 shows the model performance of users A and B, respectively. When both users applied active learning, they recorded improved accuracy and F1 score compared to baseline, and the difference in performance between users was insignificant. In particular, for the burn scar class, both the precision and recall showed similar levels, suggesting that the difference in annotation criteria does not significantly affect model performance.

In addition, Figure 10 presents an example of visually comparing the prediction results of users A and B. The predicted burnscar region is mostly consistent, with some differences occurring only at the boundary and fine regions. This shows that active learning-based model learning can operate reliably despite deviations between users.

These results support the universal applicability of active learning and imply that effective learning is possible even in various user environments.

6. Discussion

We conducted a user study to evaluate the effectiveness of active learning for annotating burn scars on satellite imagery. We address our research question in this section based on the study's results.

6.1. Research Question

Our active learning approach enabled the creation of high-quality labels suitable for burn scar detection compared to manual annotation (baseline annotation). For annotation efficiency, the active learning method reduced the average annotation time by approximately nine minutes compared to manual annotation. This indicates that our annotation tool can effectively reduce the overall annotation workload.

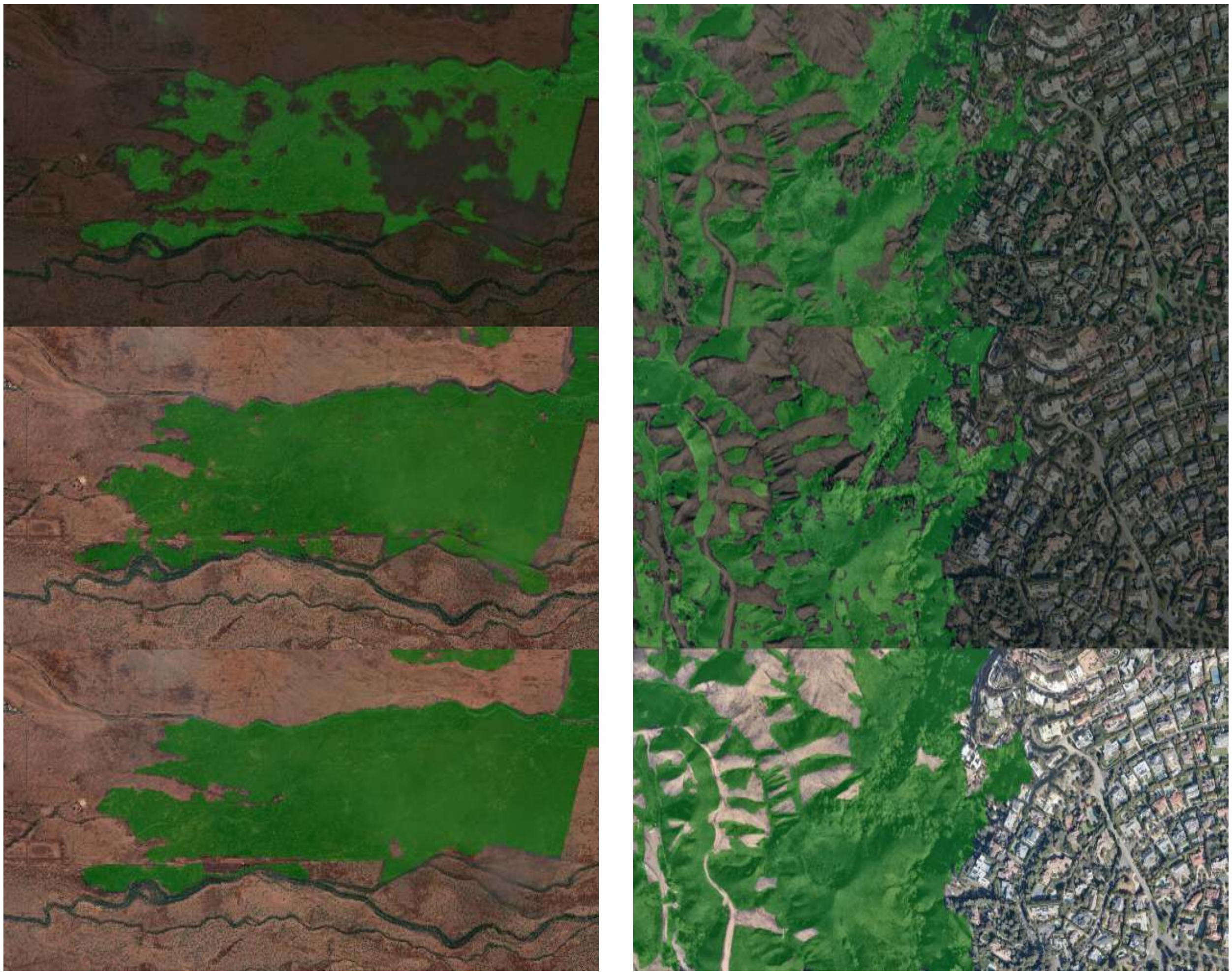
We did not have time to evaluate the answer to the third research question, so this will be addressed in future research.

It is also worth noting that we concluded the annotation process once the model reached approximately 90% accuracy. However, further iterations of active learning rounds will likely increase even higher accuracy levels.

6.2. Further research question

Our annotation tool has the potential to be applied to other domains in a broader research context. A natural follow-up research question is whether active learning can be extended to different domains and tasks to generate high-quality datasets, especially in cases where manual annotation is labor-intensive or impractical.

Since our current annotation tool is web-based, deploying it on an iPad or other tablets could enhance accuracy. Conducting a user study with a larger number of participants and incorporating a controlled experimental design may further improve the reliability of this paper. Additionally, we can refine the active learning web-based application by introducing new features, such as a model that recommends regions for re-annotation when the confidence score is low, which could significantly accelerate the annotation process.



(a) R5: Maui

(b) R2: California

Figure 9. Qualitative comparison of prediction output examples from the baseline U-Net and the active learning model across different test regions. Each presents three vertically stacked images (top to bottom): baseline prediction, active learning prediction and ground truth overlay. The active learning model demonstrates reduced false positives and improved localization of burn scars.

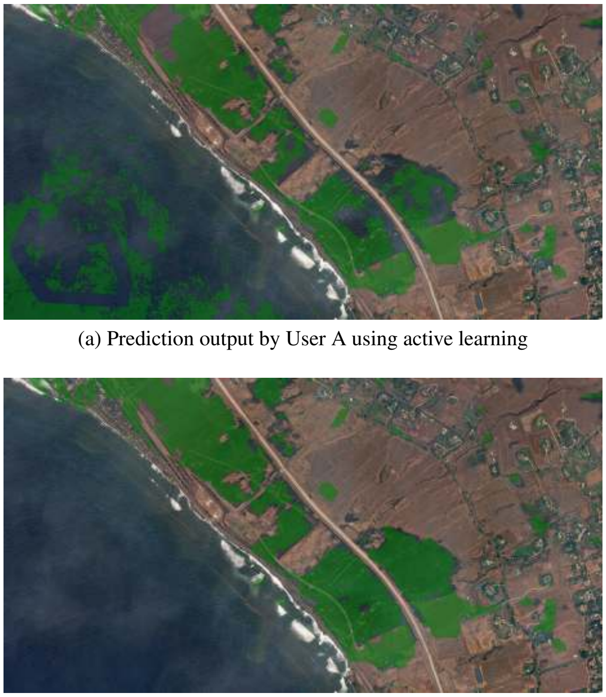
		BURN SCAR			**NOT BURN SCAR**		
User	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
A	0.9056	0.6354	0.7796	0.7001	0.9623	0.9263	0.944
B	0.9066	0.6363	0.7797	0.7007	0.9627	0.9273	0.9447

Table 5. Performance metrics comparison between two independent users (A and B) applying active learning on the same dataset. Both users achieved nearly identical accuracy and F1 scores, demonstrating the consistency and robustness of the active learning framework across annotators.

6.3. Limitations

One potential issue is that, except for one participant, all of the participants are involved in this project as a class project. As a result, some participants may have remembered the data through the team project.

Additionally, we compared the time required for initial training and annotation using active learning; however, our time measurements were inconsistent. We did not measure time programmatically, so the amount of time recorded may vary based on human involvement. The lack of a controlled



(a) Prediction output by User A using active learning

(b) Prediction output by User B using active learning

Figure 10. Visual comparison of model predictions generated by two independent annotators (User A and User B) using active learning on the same region.

environment means that our time measurements may not accurately reflect true performance.

Each participant used a Mac laptop, either a MacBook Air or a MacBook Pro, both equipped with an M-series chip CPU. However, there are differences in the CPUs; the M2 has an 8-core CPU, while the M2 Pro has either a 10-core or a 12-core CPU [2]. The PC mechanism differs between the fanless MacBook Air and the fan-equipped MacBook Pro. These variations can impact active learning times since each round's three epochs of active learning are run on each participant's computer.

7. Conclusion

This paper makes burn scar labeling faster and more accurate by combining machine learning with human expertise. The active learning approach reduces the effort needed for manual labeling while improving the model's performance. After 2–3 rounds of training and correction, we get a high-quality set of burn scar labels that can be used to train larger models for more reliable burn scar detection and mapping. This work makes it easier to track the aftermath of wildfires and plan future fire prevention and recovery efforts.

8. Work division

- Yudai: Preprocessing, UNet-initial training, code organization, final report, except result section, User Study (with one PhD student)
- Saugat: Annotation tool refinement, architecture building, Final report(related work)
- Yongjun: Poster designing, organizing, initial data preparing
- Jiyeong: Analysis and result, initial data preparation, User Study (with one PhD student), Final report (result)
- Everyone: Annotation,

References

[1] National Aeronautics and Space Administration. Wildfires, 2025. 1

[2] Apple Inc. Compare mac models. https://www.apple.com/mac/compare/, 2025. Accessed: 2025-04-25. 9

[3] L. Dyken, S. Adhikari, P. Poudel, S. Petruzza, D. Yan, W. Usher, and S. Kumar. Enabling quick, accurate crowd-sourced annotation for elevation-aware flood extent mapping. *arXiv preprint arXiv:2408.05350*, 2024. [Online]. Available: https://arxiv.org/abs/2408.05350. 1, 2

[4] California Fire Imagery. California fire imagery, 2025. 1, 3

[5] Maui Fire Imagery. Maui fire imagery, 2023. 3

[6] C. Phillips, S. Roy, A. Kumar, and R. Ramachandran. Hls foundation burnscars dataset, 2023. [Online]. Available: https://huggingface.co/datasets/ibm-nasa-geospatial/hls_burn_scars. 2, 3

[7] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):40, 2022. 1

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, 2015. 1

[9] United States Geological Survey. Wildland fire datasets, 2025. 1