Universiteit Leiden

Understanding Label Errors and Ambiguous Samples in Earth Observation Benchmarks: A Case Study on EuroSAT

Hao Chen (s3990788)

Supervisors:
Julia Wąsala & Mitra Baratchi

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 22, 2026

Abstract

Benchmark datasets play an important role in evaluating Earth Observation (EO) image classification models, but their reliability depends on the quality of their labels. In EO datasets, label issues may arise not only from annotation errors, but also from visually ambiguous samples that contain mixed land-use patterns or unclear class boundaries. This thesis investigates label errors and ambiguous samples in the EuroSAT benchmark dataset and examines how different label handling strategies affect classification performance.

A confident learning-based workflow using out-of-sample prediction probabilities was used to identify candidate label issues. MobileNet-V3-Large was selected as the main classification model after comparison with ResNet18 and EfficientNet-B0. Using 5-fold stratified cross-validation, out-of-sample probabilities were generated for all 27,000 EuroSAT samples. The label issue detection method identified 425 suspicious samples, corresponding to 1.57% of the dataset. This relatively small proportion suggests that the detected issues are unlikely to substantially affect overall training performance by themselves, but they provide a focused subset for analysing different types of label quality problems. These samples were therefore manually reviewed and assigned to three categories: acceptable, ambiguous, and clear error.

The reviewed samples represented several different cases rather than a single type of label problem. Only 56 samples were confirmed as clear annotation errors, while 149 were classified as ambiguous and 220 were still considered acceptable after visual inspection. This indicates that low self-confidence does not necessarily imply an incorrect label. Label handling experiments further showed that different treatments of these samples can affect benchmark performance. Across three random seeds, removing both clear errors and ambiguous samples achieved the highest average macro-F1 score, while fully automatic removal of all Cleanlab-flagged samples performed worse than the original baseline on average.

These findings suggest that automatically detected label issues should not be treated as a single category. For EO image classification benchmarks, distinguishing between clear errors, ambiguous samples, and acceptable low-confidence samples is important for reliable model evaluation and dataset cleaning.

Contents

1 Introduction **1**

2 Related work **3**
2.1 Label Errors in Machine Learning Datasets . . . 3
2.2 Confident Learning and Label Issue Detection . . . 4
2.3 Ambiguity and Human Uncertainty . . . 4
2.4 Label Quality in Earth Observation Datasets . . . 5

3 Methods **6**
3.1 Problem Statement and Workflow . . . 6
3.2 Confident Learning-Based Label Issue Detection . . . 6
3.3 Annotation Taxonomy . . . 7
3.4 Ambiguity-Aware Handling Strategies . . . 7

4 Experimental Setup **9**
4.1 Research Questions . . . 9
4.2 Dataset . . . 9
4.3 Model Selection . . . 10
4.4 Training and Out-of-Sample Probability Generation . . . 10
4.5 Human Validation Procedure . . . 11
4.6 Evaluation Metrics . . . 12

5 Results and Discussion **13**
5.1 Cleanlab Candidate Label Issue Detection . . . 13
5.2 Human Validation of Candidate Label Issues . . . 16
5.2.1 Review Category Distribution . . . 16
5.2.2 Acceptable and Ambiguous Samples . . . 16
5.2.3 Clear Annotation Errors . . . 17
5.3 Impact of Label Handling Strategies . . . 17
5.3.1 Initial Strategy Comparison under Seed 42 . . . 17
5.3.2 Additional Analysis Across Random Splits . . . 19
5.3.3 Effect of Removing Ambiguous Samples . . . 20
5.3.4 Effect of Fully Automated Cleanlab Removal . . . 21
5.3.5 Sensitivity to Ambiguous Sample Weights . . . 21
5.4 Class-wise Performance Analysis . . . 23
5.4.1 Class-wise F1 Comparison . . . 23
5.4.2 Class-wise Changes Relative to Baseline . . . 23
5.4.3 Classes Most Affected by Label Handling . . . 24
5.5 Additional Cleanlab Removal Experiment . . . 25
5.5.1 Progressive Removal Setup . . . 25
5.5.2 Comparison of Removal Levels . . . 26
5.6 Broader Implications for EO Benchmark Reliability . . . 27

6 Limitations and Future Work **29**
6.1 Limitations 29
6.2 Future Work 29

7 Conclusion **30**

Acknowledgements **31**

Code and Data Availability **31**

A Baseline Model Selection **31**

References **34**

1 Introduction

Benchmark datasets play a central role in evaluating Earth Observation (EO) image classification models. EO imagery is widely used to support applications such as environmental monitoring, urban planning, agriculture, disaster management, and climate research [3, 17]. With the increasing availability of satellite imagery, machine learning and deep learning methods have become important tools for automatically classifying land-use and land-cover patterns in large-scale EO datasets [17]. Datasets such as EuroSAT are therefore commonly used as benchmarks for land-use and land-cover classification [5]. However, the reliability of these models is often assessed using benchmark datasets whose labels are treated as ground truth. This makes label quality an important condition for fair and meaningful model evaluation.

In EO image classification, label quality is not always straightforward. Some samples may contain clear annotation errors, where the assigned label is inconsistent with the visual content of the image. Other samples may be visually ambiguous rather than simply incorrect, because a single image patch can contain multiple land-use types, unclear boundaries, or visually similar scene elements. This is especially relevant for EO benchmarks, where satellite image patches often represent complex real-world land-cover patterns rather than isolated objects. Multi-label EO benchmarks such as BigEarthNet and BigEarthNet-MM also reflect this issue, since individual image patches may be associated with multiple land-cover classes [13, 14]. As a result, some samples may be difficult to assign to one single class even when the annotation is not clearly wrong.

Label issues can affect benchmark reliability in different ways. Incorrect labels may distort performance estimates and affect benchmark rankings [7]. More broadly, noisy labels can also influence model training and generalization in deep learning [12]. Ambiguous samples, on the other hand, can introduce uncertainty into model evaluation without necessarily representing annotation mistakes. If label errors and ambiguous samples are treated as the same type of problem, dataset cleaning decisions may become misleading. Therefore, understanding the difference between clear annotation errors and intrinsic ambiguity is important for improving the interpretation of EO benchmark results.

Previous work has shown that label errors are common in widely used machine learning datasets and can affect model evaluation and benchmark rankings [7]. Confident learning provides a method for identifying candidate label issues by using out-of-sample prediction probabilities and estimating whether the observed label is consistent with model predictions [8]. At the same time, studies on human uncertainty and dataset cartography suggest that low-confidence or difficult samples are not always mislabelled, since some samples may be inherently ambiguous or receive different plausible labels from different annotators [10, 15]. Similar annotation challenges have also been reported in EO datasets, where complex scenes, visually similar classes, and boundary regions can make annotation difficult [1].

Despite these advances, existing studies have often focused either on detecting label errors or on studying uncertainty and ambiguity, while less attention has been given to explicitly separating clear annotation errors from intrinsically ambiguous samples in EO image classification benchmarks. Confident learning-based methods can identify samples that deserve further inspection, but they do not directly determine whether a suspicious sample is a true labeling mistake, an ambiguous case, or a difficult but still acceptable example. This leaves the relationship between label quality, ambiguity, and benchmark reliability insufficiently understood in the EO classification setting.

This thesis addresses this gap through a case study on the EuroSAT benchmark dataset. It

combines confident learning-based label issue detection with human validation to examine whether automatically detected low-confidence samples correspond to clear errors, ambiguous samples, or acceptable samples. It further evaluates how different ambiguity-aware handling strategies influence classification performance and benchmark interpretation.

The main contributions of this thesis are as follows:

- This thesis provides a reproducible workflow for identifying and validating candidate label issues in EO image classification benchmarks. Unlike a purely automatic label-cleaning approach, the workflow explicitly distinguishes between clear annotation errors, ambiguous samples, and acceptable low-confidence samples.
- This thesis presents an empirical analysis of the candidate label issue subset of EuroSAT and shows that automatically detected suspicious samples are heterogeneous. Only a minority are clear annotation errors, while many are ambiguous or remain acceptable after human inspection.
- This thesis evaluates several label handling strategies and shows that the treatment of suspicious samples can lead to small but observable changes in classification performance and class-wise benchmark behaviour. The results suggest that automatically detected label issues should be interpreted carefully before they are used for dataset cleaning or benchmark evaluation.

2 Related work

This chapter reviews the literature relevant to this thesis in four parts. First, it discusses label errors and noisy labels in machine learning datasets, which motivates the need to inspect benchmark annotations. Second, it introduces Confident Learning as a method for detecting candidate label issues from model predictions. Third, it reviews ambiguity and human uncertainty, which motivates the distinction between difficult samples and incorrectly labelled samples. Finally, it discusses label quality challenges in EO datasets, where complex scenes and mixed land-cover patterns make this distinction especially relevant.

2.1 Label Errors in Machine Learning Datasets

Label quality has received increasing attention in recent years, as machine learning models have become highly dependent on large benchmark datasets whose annotations are often treated as ground truth [7, 12]. Several studies have shown that annotation issues can influence both model development and evaluation, motivating research on identifying and understanding problematic samples [7, 15]. Northcutt et al [7]. investigated label quality in ten widely used machine learning benchmark datasets and found that label errors are both common and widespread. Their study estimated an average of at least 3.3% mislabelled samples across the examined datasets, with the ImageNet validation set containing at least 6% label errors. Furthermore, they demonstrated that correcting these errors could significantly affect reported model performance and, in some cases, alter the ranking of competing models.

More broadly, label noise has been widely studied in deep learning, as noisy labels can degrade generalization performance and motivate the development of robust training methods [12]. However, the presence of a problematic sample does not necessarily imply that its label is incorrect. Some samples may be difficult to classify because of intrinsic ambiguity rather than annotation mistakes. To better understand dataset quality, Swayamdipta et al [15]. proposed Dataset Cartography, a framework that analyzes training dynamics using metrics such as confidence and variability. Their work categorized samples into easy-to-learn, hard-to-learn, and ambiguous groups. Easy-to-learn samples typically exhibit high confidence and low variability during training, whereas hard-to-learn samples show consistently low confidence. In contrast, ambiguous samples are characterized by high variability, meaning that model predictions fluctuate considerably throughout training. An important observation from Dataset Cartography is that ambiguous samples are not necessarily mislabelled. Instead, they may represent instances that are inherently difficult to interpret or that contain multiple plausible explanations.

These works are relevant to this thesis because they show that dataset quality problems should not be reduced to a binary distinction between correct and incorrect labels. While work on label errors motivates the detection of suspicious samples, work on dataset cartography suggests that some suspicious or difficult samples may be ambiguous rather than wrong. This thesis follows this distinction by explicitly separating clear annotation errors from ambiguous samples in an EO image classification benchmark.

2.2 Confident Learning and Label Issue Detection

Identifying label errors in large datasets is challenging because manually inspecting every sample is often impractical. Large-scale benchmark datasets can contain tens of thousands or even millions of images, so exhaustive manual inspection is costly and time-consuming. For example, Northcutt et al [7]. used automated detection followed by human validation to study label errors across multiple benchmark datasets, illustrating the need for methods that can first reduce the search space to a smaller set of suspicious samples. To address this type of problem, Northcutt et al [8] proposed Confident Learning, a framework for estimating uncertainty in dataset labels and identifying potentially mislabelled samples.

The central idea of Confident Learning is to estimate the joint distribution between observed noisy labels and latent true labels using predicted class probabilities. This estimated confident joint can then be used to identify samples whose observed labels are likely to be inconsistent with the model evidence. In practice, Confident Learning can operate using the original dataset labels and model-predicted probabilities. Out-of-sample predicted probabilities are useful in this setting because they reduce the risk of using predictions from a model that has already been trained on the same sample. Samples whose observed labels receive low prediction confidence can therefore be treated as candidate label issues for further inspection.

For this thesis, Confident Learning is important as a candidate detection method rather than as a fully automatic label correction method. It can help identify samples where the observed label and model predictions disagree, but it does not directly determine whether a suspicious sample represents a genuine annotation error, an intrinsically ambiguous case, or a difficult but still acceptable sample. Consequently, human validation is required to determine the nature of the detected issues before different label handling strategies can be evaluated.

2.3 Ambiguity and Human Uncertainty

While label errors represent one source of dataset quality issues, not all problematic samples are necessarily mislabelled. In many classification tasks, some samples are inherently difficult to interpret and may reasonably receive different labels from different annotators. This observation has motivated research on human uncertainty and ambiguity in machine learning datasets [10, 15].

Peterson et al. investigated the role of human uncertainty in image classification and introduced CIFAR-10H [10], a dataset containing multiple human annotations for each image in the CIFAR-10 test set. Their study showed that annotator disagreement often reflects genuine uncertainty rather than annotation mistakes. Some images consistently received different labels from different annotators, suggesting that ambiguity can be an inherent property of the data. Furthermore, Peterson et al. demonstrated that incorporating human uncertainty into model training can improve both generalization and robustness. Instead of treating all labels as equally certain, they represented labels as distributions reflecting the opinions of multiple annotators. Their results showed that models trained with these uncertainty-aware labels performed better on out-of-distribution data and exhibited greater robustness to adversarial perturbations.

The distinction between annotation errors and intrinsic ambiguity is particularly relevant for EO image classification. Satellite images often contain mixed land-use patterns, visually similar classes, or boundary regions, making some samples difficult to assign to a single correct category [1, 13, 14]. Therefore, ambiguous samples should not automatically be treated as label errors, and

different handling strategies may be required for these two types of label issues. This thesis uses this distinction as the basis for separating clear annotation errors from ambiguous samples during human validation.

2.4 Label Quality in Earth Observation Datasets

Label quality issues are not limited to general machine learning benchmarks and have also been observed in Earth Observation (EO) datasets [1]. The annotation of EO imagery is particularly challenging because satellite images often contain complex scenes, mixed land-use patterns, and visually similar categories. As a result, obtaining accurate and consistent labels can be difficult, even for experienced annotators.

Aybar et al. [1] investigated annotation quality in CloudSEN12, a benchmark dataset for cloud semantic segmentation. Through manual inspection of the dataset, they identified various annotation problems, including incorrect labels and inconsistent annotations. Their analysis showed that some issues resulted from genuine annotation mistakes, while others arose from the inherent difficulty of distinguishing visually similar classes and boundary regions. The challenge of assigning a single label to EO image patches is also reflected in multi-label remote sensing benchmarks such as BigEarthNet [13] and BigEarthNet-MM [14], where image patches may be associated with multiple land-cover classes. This supports the idea that ambiguity in EO imagery may come from the visual content of the image itself rather than from annotation mistakes alone.

These findings suggest that label quality problems in EO datasets may originate from multiple sources and cannot always be attributed to simple labeling errors. Similar to observations in general machine learning datasets, problematic samples in EO benchmarks may reflect both annotation errors and intrinsic ambiguity. This directly motivates the approach used in this thesis: suspicious samples are first detected using a confident learning-based method and are then manually reviewed to distinguish between acceptable samples, ambiguous samples, and clear annotation errors.

Taken together, the reviewed literature shows that label quality is important for benchmark reliability, but also that suspicious samples should not automatically be interpreted as mislabelled. Previous work has studied label errors, confident learning, human uncertainty, and EO annotation challenges, but relatively little attention has been given to explicitly distinguishing clear annotation errors from intrinsically ambiguous samples in EO image classification benchmarks. This thesis addresses this gap by combining confident learning-based label issue detection with human validation in EuroSAT and by evaluating how different handling strategies affect classification performance and benchmark interpretation.

3 Methods

This section describes the methodological workflow used in this thesis. The workflow is designed to move from automatic detection of suspicious samples to human interpretation and then to ambiguity-aware label handling. The section first defines the label quality problem addressed in this study, then explains the confident learning-based detection step, the annotation taxonomy used for human interpretation, and the label handling strategies evaluated later in the experiments.

3.1 Problem Statement and Workflow

This thesis investigates label quality in an EO image classification benchmark. Given a dataset of images (x_i) with observed dataset labels $(\tilde{y}_i)$, the goal is not to assume that every low-confidence sample is mislabelled, but to determine what kind of label issue it may represent. In this setting, a suspicious sample may correspond to a clear annotation error, an intrinsically ambiguous image, or a difficult but still acceptable example. Distinguishing between these cases is important because they have different implications for dataset cleaning and benchmark interpretation.

The methodological problem is therefore twofold. First, the full dataset is too large to inspect manually, so a method is needed to identify a smaller set of candidate label issues. Second, automatic detection alone does not determine whether a candidate issue is a genuine labeling mistake or an ambiguous case. For this reason, this thesis combines confident learning-based label issue detection with human validation. The automatic step is used to prioritize suspicious samples, while the human validation step is used to interpret the nature of these samples.

The workflow consists of four main steps. First, a confident learning-based detection method is used to assign label quality scores to the dataset samples and identify a ranked subset of candidate label issues. Second, these candidate samples are reviewed manually and assigned to one of three categories: acceptable, ambiguous, or clear error. Third, this taxonomy is used to define ambiguity-aware label handling strategies. Finally, the resulting strategies are evaluated to examine whether separating clear errors from ambiguous samples affects classification performance and benchmark interpretation.

The following subsections describe these steps in more detail. Section 3.2 explains the confident learning-based detection step. Section 3.3 defines the annotation taxonomy used to distinguish acceptable, ambiguous, and clear-error samples. Section 3.4 then describes the label handling strategies evaluated in this thesis.

3.2 Confident Learning-Based Label Issue Detection

The first step of the workflow is to identify a smaller set of candidate label issues from the full dataset. This is done using a confident learning-based approach, which compares the observed dataset labels with model-predicted class probabilities to estimate whether the assigned labels are supported by the model evidence [8]. In this thesis, the purpose of this step is not to automatically correct labels but to prioritize samples that require human inspection.

For each sample (x_i) with observed label $(\tilde{y}_i)$, a self-confidence score is computed as the predicted probability assigned to the observed label:

$$SC_i = P(\tilde{y}_i \mid x_i).$$

A high self-confidence score indicates that the model strongly supports the observed label, whereas a low score indicates disagreement between the model prediction and the dataset annotation. Samples with lower self-confidence are therefore treated as more suspicious and are ranked higher for manual review. The self-confidence criterion was used because it provides an interpretable sample-level score for prioritizing candidate label issues.

The output of this step is a ranked list of candidate label issues. Importantly, these candidates are not treated as confirmed annotation errors. A low self-confidence score may indicate that the original label is wrong, but it may also result from visual ambiguity or from a difficult example that the model fails to classify confidently. Therefore, the ranked candidates are passed to the human validation step described in Section 4.5.

3.3 Annotation Taxonomy

Low-confidence samples identified by a confident learning-based method do not necessarily correspond to annotation errors. Some samples may have incorrect labels, while others may be visually ambiguous or difficult for the model despite having a reasonable original label. Therefore, a taxonomy is needed to distinguish between different types of suspicious samples before defining label handling strategies.

This thesis uses three review categories: **acceptable**, **ambiguous**, and **clear error**. A sample is classified as **acceptable** when the original label remains visually reasonable after review, even if the model assigns higher confidence to another class. A sample is classified as **ambiguous** when multiple labels are plausible or when the image does not provide enough visual evidence for a single clear category. This may occur when an image patch contains mixed land-use patterns, visually similar classes, or unclear boundaries. A sample is classified as a **clear error** when the image content is clearly inconsistent with the original label and another class is evidently more appropriate.

This distinction is important because clear errors and ambiguous samples have different meanings. A clear error indicates a likely annotation mistake, whereas an ambiguous sample does not necessarily contain a wrong label. Treating these cases separately enables a more detailed analysis of label quality and makes it possible to evaluate handling strategies that remove, retain, or down-weight different types of suspicious samples.

3.4 Ambiguity-Aware Handling Strategies

After the candidate samples have been assigned to the three review categories, the next step is to examine whether different ways of handling these categories affect classification performance and benchmark interpretation. Several label handling strategies are evaluated, as summarized in Table 1. The original training set and fully automatic removal strategy are used as reference conditions, while the strategies that distinguish between clear errors and ambiguous samples are treated as ambiguity-aware variants proposed and evaluated in this thesis.

1. **Original training set.** This strategy keeps the original training data unchanged. It serves as the baseline condition against which all other strategies are compared.

2. **Remove clear errors.** This strategy removes only samples classified as clear annotation errors during human validation. It evaluates whether removing confirmed label mistakes

improves classification performance while retaining ambiguous and acceptable samples.

3. **Remove all detected candidates.** This strategy removes all samples identified by the confident learning-based detection step, regardless of their human review category. It represents a fully automatic dataset cleaning approach and is included as a reference condition to test whether automatic removal alone is sufficient.

4. **Remove clear errors and ambiguous samples.** This ambiguity-aware strategy removes both clear annotation errors and ambiguous samples. It is based on the idea that, under a single-label classification setup, ambiguous samples may introduce uncertain supervision even when they are not clearly mislabelled.

5. **Remove clear errors and down-weight ambiguous samples.** This ambiguity-aware strategy removes clear annotation errors while retaining ambiguous samples with a reduced training weight of 0.5. The motivation is that ambiguous samples may still contain useful information, but should have less influence on model optimization than samples with more certain labels.

Table 1: Label handling strategies evaluated in this study.

Strategy	Acceptable	Ambiguous	Clear errors
Original training set	Keep	Keep	Keep
Remove clear errors	Keep	Keep	Remove
Remove all detected candidates	Remove	Remove	Remove
Remove clear errors and ambiguous samples	Keep	Remove	Remove
Remove clear errors and down-weight ambiguous samples	Keep	Weight = 0.5	Remove

In addition to these main strategies, an exploratory progressive removal experiment is conducted to examine whether the ranking produced by the confident learning-based detection step is useful as a standalone pruning criterion. In this experiment, samples are sorted by self-confidence, and the top 50, top 200, and all detected candidates are used as removal pools. This experiment evaluates whether removing increasingly many low-confidence samples improves performance, or whether human validation is necessary before using the detected candidates for dataset cleaning.

4 Experimental Setup

This section describes how the methodological workflow introduced in the previous section is implemented and evaluated. While Section 3 defined the conceptual workflow for detecting, interpreting, and handling suspicious samples, this section specifies the empirical setting used in the experiments. It introduces the research questions, the EuroSAT dataset, the model selection procedure, the training and out-of-sample probability generation setup, the human validation procedure, and the evaluation metrics.

4.1 Research Questions

Based on the methodological workflow described in Section 3, this study investigates how automatically detected candidate label issues should be interpreted and handled in an EO image classification benchmark. The focus is not only on detecting suspicious samples but also on distinguishing clear annotation errors from intrinsically ambiguous samples and acceptable low-confidence samples.

The main research question is:

> **RQ:** How do label errors and ambiguous samples affect evaluation in EO image classification benchmarks, and how does ambiguity-aware handling influence classification performance and benchmark interpretation?

This question is addressed through the following sub-questions:

1. **RQ1:** To what extent does a confident learning-based detection method identify candidate label issues in EuroSAT, and how many of these candidates correspond to clear annotation errors, ambiguous samples, or acceptable low-confidence samples after human validation?
2. **RQ2:** How do different label handling strategies, including automatic removal, removing clear errors, removing ambiguous samples, and down-weighting ambiguous samples, influence classification performance and benchmark interpretation?

4.2 Dataset

This study uses EuroSAT [5] as a case study for detecting and analyzing label errors and ambiguous samples in an EO image classification benchmark. EuroSAT is suitable for this purpose because it is a standard land-use and land-cover classification benchmark, while its image patches may still contain mixed land-use patterns, visually similar categories, and unclear class boundaries. These properties make it relevant for studying whether suspicious samples are true annotation errors or intrinsically ambiguous cases.

EuroSAT was introduced by Helber et al. [5] and is constructed from Sentinel-2 satellite imagery provided by the European Space Agency (ESA). Sentinel-2 is part of the Copernicus Earth observation programme and provides high-resolution multispectral imagery of the Earth's surface [3]. EuroSAT was designed to support the development and evaluation of machine learning methods for land-use and land-cover classification tasks.

The dataset contains 27,000 georeferenced image patches of size (64×64) pixels. These images are evenly distributed across ten land-use and land-cover categories: Annual Crop, Forest,

Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea/Lake. The balanced class distribution makes the dataset suitable for benchmarking image classification models while reducing the influence of class imbalance.

Although Sentinel-2 provides thirteen spectral bands, this study uses only the RGB representation of each image. Restricting the input to RGB imagery follows a common benchmark setting and makes the results easier to compare with standard image classification experiments. It also supports the manual review stage, since RGB images can be visually inspected without requiring additional spectral analysis tools.

Because benchmark datasets are often treated as ground truth references, label quality directly affects the reliability of reported model performance. For this reason, EuroSAT provides an appropriate case study for investigating how clear label errors and ambiguous samples influence EO benchmark evaluation.

4.3 Model Selection

A preliminary model comparison was conducted to select a suitable classification model for the subsequent label quality experiments. Three convolutional neural network (CNN) architectures were evaluated: ResNet18, EfficientNet-B0, and MobileNet-V3-Large. These models were chosen because they are widely used image classification architectures and represent different trade-offs between model depth, efficiency, and classification performance.

ResNet18 [4] uses residual connections to support the training of deep neural networks and is commonly used as a baseline architecture in computer vision. EfficientNet-B0 [16] improves efficiency through compound scaling, which jointly scales network depth, width, and input resolution. MobileNet-V3-Large [6] is designed to provide a favourable balance between computational efficiency and classification accuracy by combining lightweight convolutional operations with neural architecture search.

All three models were trained and evaluated using the same experimental protocol. Based on this preliminary comparison, MobileNet-V3-Large achieved the best overall performance and was selected as the main model for the remaining experiments. Therefore, the subsequent out-of-sample probability generation, label issue detection, and label handling experiments were conducted using MobileNet-V3-Large.

4.4 Training and Out-of-Sample Probability Generation

All experiments were implemented in PyTorch [9]. Unless stated otherwise, experiments were conducted using a fixed random seed of 42. Only minimal preprocessing was applied to the EuroSAT images: the RGB images were converted to tensors and used at a resolution of 64×64 pixels. No data augmentation was used in order to keep the experimental setting controlled and to avoid confounding the comparison between different label handling strategies.

The models were initialized with ImageNet-pretrained weights [2]. Training was performed using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 64, and a training duration of 5 epochs. Cross-entropy loss was used as the training objective. The main training configuration is summarized in Table 2.

For the preliminary model comparison and the label handling strategy experiments, the dataset was divided into training, validation, and test subsets using a 70%/15%/15% split. The same split

Table 2: Main training configuration used throughout the study.

Parameter	Value
Framework	PyTorch
Input size	(64×64) RGB
Batch size	64
Epochs	5
Optimizer	Adam
Learning rate	10^{-3}
Loss function	Cross-entropy
Initial weights	ImageNet pretrained
Default random seed	42

was used across all label handling strategy experiments to ensure a fair comparison. Model selection was based on validation accuracy, and the checkpoint with the highest validation accuracy was retained for final evaluation on the test set.

Out-of-sample probabilities were generated using a separate 5-fold stratified cross-validation procedure. Stratification ensured that the class distribution remained approximately consistent across folds. In each fold, a new MobileNet-V3-Large model was trained on four folds and used to predict the held-out fold. This process was repeated five times so that every image served as held-out data exactly once. As a result, each image received one out-of-sample probability vector from a model that had not been trained on that image; predictions from different folds were not averaged. The resulting probability matrix had dimensions $(27{,}000 \times 10)$, with one row per image and one probability value per EuroSAT class. This matrix was then used as input to the confident learning implementation for candidate label issue detection.

The main experiments were conducted using seed 42. An additional split-sensitivity analysis using seeds 42, 43, and 44 was conducted for the label handling strategies and is reported in the results section.

4.5 Human Validation Procedure

After candidate label issues had been detected, a manual review was conducted to determine the type of label issue represented by each sample. This step was necessary because the automatic detection method can identify suspicious samples but cannot determine whether they correspond to clear annotation errors, intrinsically ambiguous cases, or difficult but acceptable examples.

For each candidate sample, the review interface displayed the RGB image, the original dataset label, and the model-predicted class with the highest probability. The review was based only on visual inspection of the RGB image. No external geographic information, metadata, multispectral bands, or additional annotations were used. During inspection, the original label was considered acceptable when the visually dominant scene content was consistent with the assigned class, or when the assigned class remained a reasonable interpretation of the image. A sample was considered a clear error when the original label was visually unsupported and another EuroSAT class was evidently more appropriate, such as a sample labelled as *River* but visually corresponding more clearly to *Highway*. A sample was considered ambiguous when multiple classes were visually plausible, when the image contained mixed land-use patterns, or when the visual evidence was

insufficient to assign a single confident label. Examples include image patches containing both river and forest regions or scenes located near class boundaries.

Each candidate sample was assigned to one of the three categories defined in Section 3.3: **acceptable**, **ambiguous**, or **clear error**. All reviews were performed by the author. The purpose of the manual review process was not to immediately correct labels but to distinguish between different types of suspicious samples. The resulting human validation labels were then used to analyze the composition of the detected candidate issues and to evaluate the label handling strategies described in Section 3.4.

4.6 Evaluation Metrics

The effect of different label handling strategies was evaluated using accuracy and macro-F1 score, which are commonly used metrics for classification tasks [11]. Accuracy was used to measure the overall proportion of correctly classified samples. However, accuracy alone may hide class-specific changes, especially when different label handling strategies affect some classes more strongly than others.

For this reason, macro-F1 was used as the primary evaluation metric. For each class, the F1 score combines precision and recall and is defined as

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

The macro-F1 score is then computed by averaging the F1 scores across all (C) classes:

$$Macro - F1 = \frac{1}{C} \sum_{c=1}^{C} F1_c.$$

Macro-F1 assigns equal weight to each class and is therefore useful for comparing label handling strategies beyond overall accuracy. In this study, accuracy is reported as a complementary measure of overall performance, while macro-F1 is used as the main metric for interpreting differences between strategies.

5 Results and Discussion

This section reports and interprets the main findings of the study. The results are organised around two parts. First, the Cleanlab-based candidate detection and human validation results are used to examine what kinds of samples were identified as potential label issues. This addresses RQ1 by analysing how many candidate issues were detected and how they were distributed across acceptable, ambiguous, and clear-error categories. Second, the label handling experiments evaluate how different treatments of these samples affected classification performance and class-wise benchmark behaviour, addressing RQ2.

MobileNet-V3-Large was used as the classifier for all experiments reported in this section. It was selected based on a preliminary comparison with ResNet18 and EfficientNet-B0, together with its stable training behaviour. The full model selection results are reported in Appendix A.

5.1 Cleanlab Candidate Label Issue Detection

Figure 1 shows that the self-confidence distribution is strongly skewed towards high values, with most EuroSAT samples concentrated near a score of 1. The corresponding summary statistics are reported in Table 3. The median self-confidence across the full dataset was 0.9999, while the mean was 0.942. Together, the figure and table indicate that the model prediction was consistent with the observed label for the majority of the dataset.

At the same time, the distribution contains a small low-confidence tail. In total, 251 samples had a self-confidence score below 0.01, and 882 samples had a score below 0.20. Cleanlab identified 425 samples as candidate label issues, corresponding to 1.57% of the dataset. This proportion is relatively small compared with the full dataset, suggesting that the detected issues are unlikely to dominate the overall training process by themselves but still provide a focused subset for manual inspection. It is also lower than the average label error rate of at least 3.3% reported by Northcutt et al. across several benchmark datasets [7]. Although the two values are not directly comparable, since the 1.57% reported here refers to Cleanlab candidate issues whereas Northcutt et al. estimated confirmed label errors after validation, the comparison suggests that EuroSAT does not exhibit an unusually large number of potentially problematic samples under this detection setup.

As shown in Table 3, the Cleanlab-flagged subset had substantially lower self-confidence than the dataset as a whole. Its median self-confidence was 0.0103, compared with 0.9999 for all samples, confirming that Cleanlab concentrated the review process on the lowest-confidence portion of the dataset rather than selecting a random subset.

Table 3: Summary of self-confidence scores for all samples and Cleanlab-flagged samples. The flagged samples form a small but clearly lower-confidence subset of EuroSAT.

Statistic	All samples	Cleanlab-flagged samples
Count	27,000	425
Minimum	1.35×10^{-23}	1.35×10^{-23}
Maximum	1.000	0.333
Mean	0.942	0.0247
Median	0.9999	0.0103

Table 4 further shows that most flagged samples were concentrated at very low self-confidence

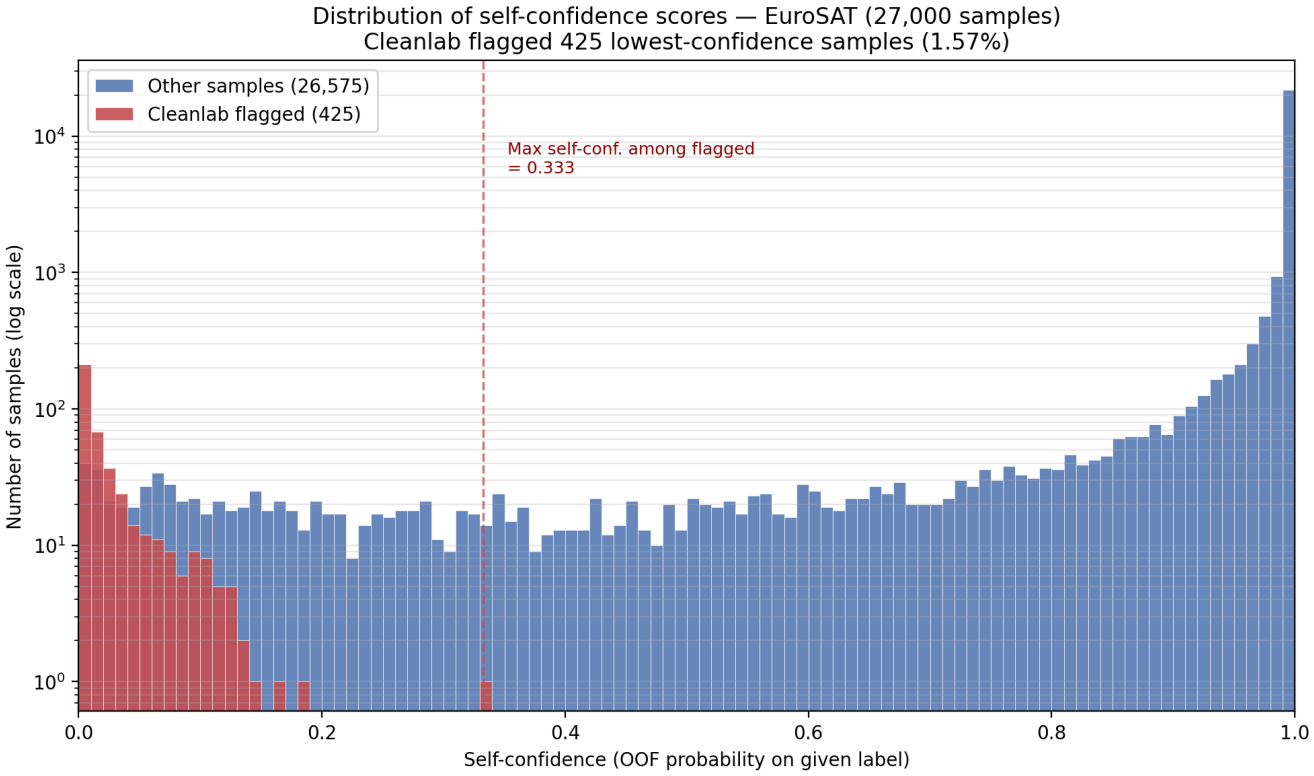

Figure 1: Distribution of self-confidence scores for all EuroSAT samples. Most samples received high confidence for their original labels, while a small low-confidence tail was selected for further inspection.

values. For example, 401 of the 425 flagged samples had self-confidence below 0.10, and 424 were below 0.20. This result is expected because the detection was based on self-confidence ranking. The more important observation is that only a small fraction of the full dataset falls into this low-confidence region, which suggests that EuroSAT is largely consistent with the model predictions, while still containing a meaningful subset of difficult or potentially problematic samples.

Table 4: Number of samples below selected self-confidence thresholds. Most Cleanlab-flagged samples are concentrated in the low-confidence tail of the full distribution.

Threshold	All samples	Cleanlab-flagged samples
< 0.01	251	211
< 0.05	489	354
< 0.10	668	401
< 0.20	882	424

Figure 2 shows the top-ranked Cleanlab detections. These examples illustrate why low self-confidence should be interpreted as a signal for inspection rather than as direct evidence of a wrong label. Some highly ranked samples were clear annotation errors, especially cases where images labelled as *River* were predicted as *Highway*. This confusion is plausible because both classes can appear as long, narrow, and visually continuous structures in RGB satellite images. At the 64×64

patch scale, roads and rivers may also occupy only part of the image, while surrounding vegetation or built-up areas can make the dominant class difficult to identify.

Figure 2: Top-ranked Cleanlab samples. Low self-confidence often reflects visual confusion between similar classes, but the examples include clear errors as well as acceptable samples.

However, not all highly ranked samples were clear errors. Some examples were reviewed as acceptable despite extremely low self-confidence scores. These cases often contain visual features that overlap with the predicted class, but the original label remains defensible when considering the whole image patch. This suggests that the model sometimes focuses on locally salient structures or visually similar patterns, while the human review may still consider the original label reasonable. Therefore, the Cleanlab output should be treated as a candidate set for human validation, not as an automatically confirmed set of mislabelled samples.

5.2 Human Validation of Candidate Label Issues

5.2.1 Review Category Distribution

All 425 samples flagged by Cleanlab were manually reviewed and assigned to one of three categories: **acceptable**, **ambiguous**, or **clear error**. The distribution of review labels is shown in Table 5 and Figure 3.

Table 5: Human review results for the 425 Cleanlab-flagged samples. Percentages are reported relative to both the flagged set and the full EuroSAT dataset.

Review category	Count	% of flagged samples	% of full dataset
Acceptable	220	51.76%	0.81%
Ambiguous	149	35.06%	0.55%
Clear error	56	13.18%	0.21%
Total	425	100%	1.57%

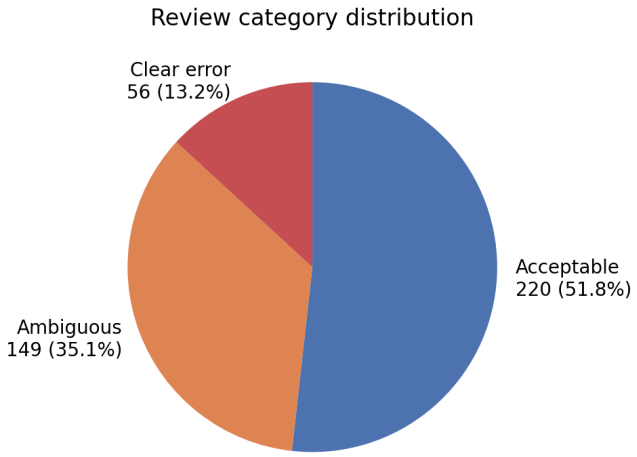

Figure 3: Distribution of human review categories among the Cleanlab-flagged samples. The candidate set is heterogeneous: most flagged samples were not confirmed as clear annotation errors.

Only 56 of the 425 flagged samples were confirmed as clear annotation errors. This corresponds to 13.18% of the flagged set and approximately 0.21% of the full EuroSAT dataset. Confirmed clear errors therefore form only a small fraction of both the candidate set and the complete dataset. In contrast, 220 flagged samples were judged to be acceptable after visual inspection, corresponding to 51.76% of the flagged set and approximately 0.81% of the full dataset. This confirms that the original label is not necessarily wrong when the model assigns it a low self-confidence score.

5.2.2 Acceptable and Ambiguous Samples

A likely explanation for the large acceptable category is that many acceptable flagged samples are difficult examples rather than annotation mistakes. These images may contain mixed land-use

patterns, visually similar structures, or only partial evidence for the labelled class. As a result, the model may assign higher confidence to another visually plausible class, while the original label remains defensible when the full image patch is considered. This interpretation is especially relevant for EO image classification, where a single 64×64 image patch can contain several land-cover elements and unclear class boundaries.

For example, a patch labelled as *SeaLake* may contain surrounding vegetation or open land that makes the model prefer *Pasture*. Similarly, a patch labelled as *River* may contain linear structures or nearby roads that make it visually similar to *Highway*. In such cases, the model may focus on locally salient patterns, while the original dataset label remains reasonable when the full image patch is considered.

Ambiguous samples formed the second largest category, with 149 samples, corresponding to 35.06% of the flagged set and approximately 0.55% of the full dataset. These samples were not clearly wrong, but multiple labels were visually plausible. Ambiguity often occurred when an image patch contained mixed land-use patterns, visually similar structures, or unclear class boundaries. For example, some samples originally labelled as *Industrial* or *Highway* were predicted as *River*, because the image contained both built structures and river-like or road-like linear patterns.

Ambiguous samples therefore constitute a substantial part of the Cleanlab-flagged set. More generally, this suggests that confident learning-based detection in EO image classification may surface not only annotation mistakes, but also visually uncertain or mixed land-cover cases. Treating all ambiguous samples as label errors would therefore oversimplify the dataset quality problem. Instead, ambiguous samples require separate consideration because they may still contain useful information for model training and benchmark evaluation.

5.2.3 Clear Annotation Errors

The clear error category contained 56 samples. These were cases where the image content was clearly inconsistent with the original dataset label and another EuroSAT class was visually more appropriate. For example, *River_2331* was originally labelled as *River*, but the visual content was more consistent with a highway scene. Similarly, *Pasture_889* was originally labelled as *Pasture*, while the image was more consistent with *Forest*. These cases show that the Cleanlab-ranked candidate set did include genuine annotation errors, although they represented only a minority of the flagged samples.

Figure 4 shows representative examples from the three human review categories. The examples illustrate why human validation was necessary after automated detection: some low-confidence samples were clear errors, while others were ambiguous or still acceptable. This heterogeneity provides the basis for the label handling strategies evaluated in the next section, where clear errors and ambiguous samples are treated differently.

5.3 Impact of Label Handling Strategies

5.3.1 Initial Strategy Comparison under Seed 42

After the human validation step, five label handling strategies were first evaluated using MobileNet-V3-Large under the original seed-42 experimental setting. All strategies used the same training configuration, validation set, and test set. Only the training set was modified according to the corresponding label handling rule. The test set contained 4,050 images for all strategies.

Figure 4: Representative examples from the human validation process. The flagged samples include acceptable, ambiguous, and clear-error cases, showing that low self-confidence does not correspond to a single type of label issue.

Table 6 summarizes the performance of the five strategies under seed 42. The baseline strategy, *train_original*, used the original training set without removing or reweighting any samples. It achieved an accuracy of 96.84% and a macro-F1 score of 96.76%.

In this initial seed-42 experiment, *prune_clear_and_ambiguous* achieved the best performance, with an accuracy of 97.23% and a macro-F1 score of 97.18%. Compared with *train_original*, this corresponds to an increase of 0.39 percentage points in accuracy and 0.42 percentage points in macro-F1. The other modified strategies performed slightly below the baseline in this setting.

Figure 5 visualizes the accuracy and macro-F1 scores of the five strategies under seed 42.

These results provide an initial comparison of how different label handling strategies affect model performance. However, the observed differences between strategies are relatively small. Therefore, an additional analysis across random splits was conducted to examine how sensitive the observed pattern was to the particular seed-42 split.

Table 6: Initial seed-42 performance comparison of label handling strategies on the EuroSAT test set.

Strategy	Train size	Removed	Accuracy (%)	Macro-F1 (%)
train_original	18,900	0	96.84	96.76
prune_clear_error	18,861	39	96.27	96.19
prune_cleanlab_all	18,598	302	96.35	96.27
prune_clear_and_ambiguous	18,759	141	97.23	97.18
prune_clear_downweight_amb	18,861	39	96.27	96.26

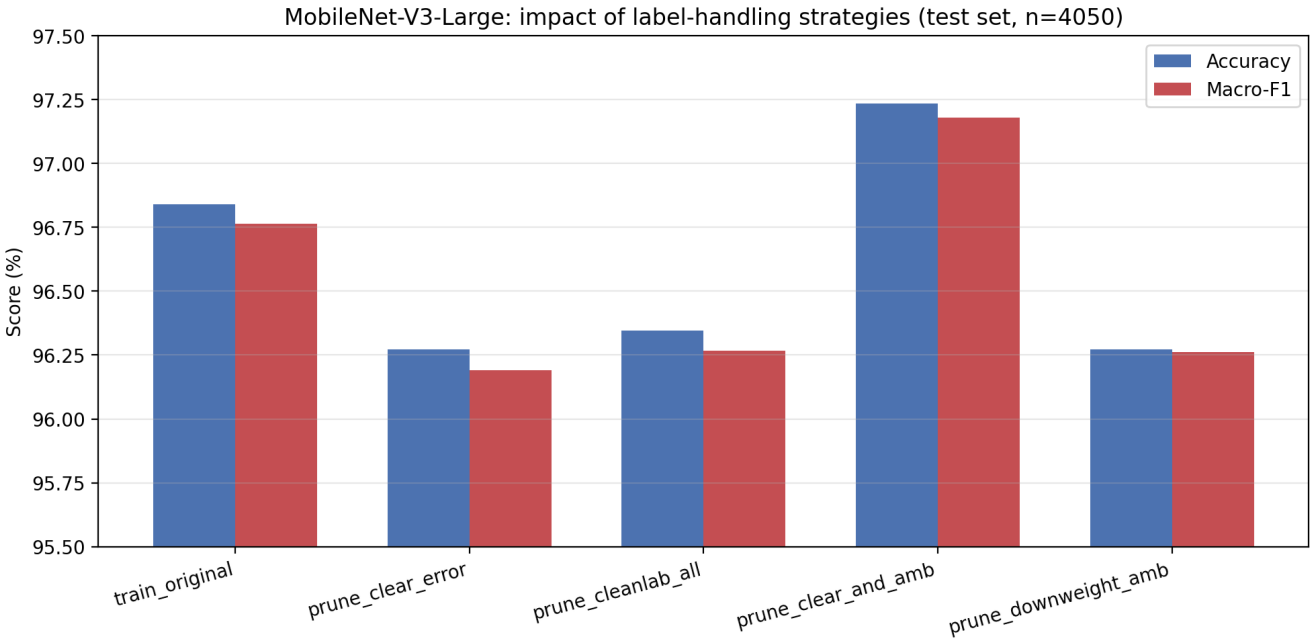

Figure 5: Initial seed-42 comparison of label handling strategies. Removing both clear errors and ambiguous samples achieved the highest performance, while fully automatic Cleanlab removal did not improve over the baseline.

5.3.2 Additional Analysis Across Random Splits

As an additional analysis, the five label handling strategies were evaluated using three random seeds: 42, 43, and 44. In these runs, the random seed affected the train-validation-test split, meaning that the test set was not identical across seeds. Therefore, these results should not be interpreted as a strict repeated-run comparison of the strategies. Instead, they provide an indication of how sensitive the observed trends are to the particular data split used in the main seed-42 experiment.

Table 7 shows the mean and standard deviation of accuracy and macro-F1 across the three random splits.

Across the three splits, *prune_clear_and_ambiguous* achieved the highest average accuracy and macro-F1 score. Its mean macro-F1 was 96.75%, compared with 95.93% for the *train_original* baseline. This pattern is consistent with the main seed-42 experiment, where the same strategy also achieved the highest performance. However, because the test sets differ across seeds and the differences between strategies are relatively small, this result should be interpreted cautiously.

Table 7: Performance of label handling strategies across three random train-validation-test splits. These results provide a split-sensitivity check rather than a strict repeated-run comparison.

Strategy	Accuracy (%)	Macro-F1 (%)
train_original	96.02 ± 0.85	95.93 ± 0.89
prune_clear_error	96.42 ± 0.55	96.35 ± 0.56
prune_cleanlab_all	95.84 ± 0.40	95.69 ± 0.44
prune_clear_and_ambiguous	96.86 ± 0.51	96.75 ± 0.57
prune_clear_downweight_amb	96.66 ± 0.31	96.57 ± 0.25

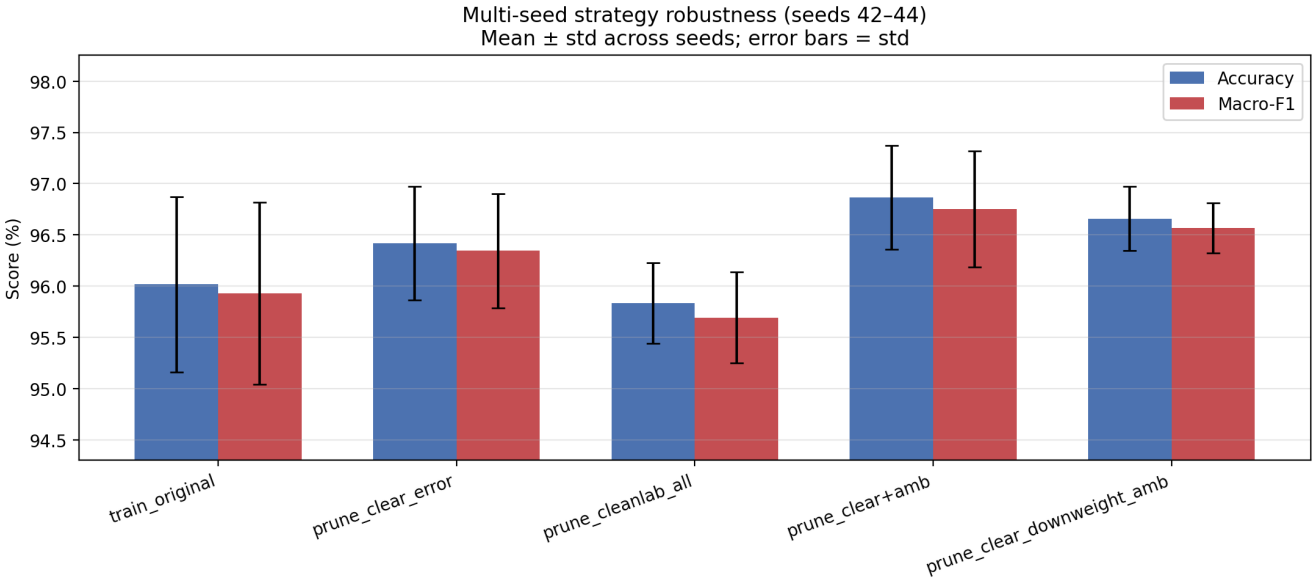

Figure 6: Mean accuracy and macro-F1 of label handling strategies across three random train-validation-test splits. Error bars show the standard deviation across splits, indicating that performance varies with the selected split.

The results were not identical across all splits. For example, seed 43 produced lower overall performance for several strategies, and prune_clear_downweight_amb outperformed prune_clear_and_ambiguous on that split. This indicates that the observed improvements are sensitive to the data split. Therefore, the main seed-42 comparison, where all strategies are evaluated on the same test set, should be treated as the primary result. The additional split analysis supports the general trend but does not provide conclusive evidence of a statistically robust improvement.

5.3.3 Effect of Removing Ambiguous Samples

The main difference between *prune_clear_error* and *prune_clear_and_ambiguous* is the treatment of ambiguous samples. In the seed-42 split, *prune_clear_error* removed only 39 clear annotation errors from the training set, whereas *prune_clear_and_ambiguous* removed 141 samples in total, including 39 clear errors and 102 ambiguous samples.

In the single-seed experiment, removing only clear errors did not improve performance compared with the original baseline. However, when ambiguous samples were also removed, performance

increased to 97.23% accuracy and 97.18% macro-F1. This indicates that the ambiguous samples had a stronger effect on the observed performance than the clear errors alone in the seed-42 setting.

The additional split analysis is consistent with this observation at the average level. Across the three random splits, *prune_clear_and_ambiguous* achieved the highest mean macro-F1 score, while *prune_clear_error* achieved a smaller improvement over the baseline. This suggests that, under the current single-label classification setup, removing visually ambiguous samples can sometimes produce a cleaner training set and improve benchmark performance.

This result should be interpreted carefully. It does not mean that all ambiguous samples are incorrect or useless. Ambiguous samples may still contain meaningful information about mixed or boundary land-cover cases. The result only shows that, in this experimental setup, removing the reviewed ambiguous samples led to the best average performance among the tested strategies.

5.3.4 Effect of Fully Automated Cleanlab Removal

The *prune_cleanlab_all* strategy removed all samples flagged by Cleanlab that appeared in the training split, regardless of their human review category. In the seed-42 split, this removed 302 training samples. These included 39 clear errors, 102 ambiguous samples, and 161 samples that were later judged as acceptable during human review.

In the seed-42 experiment, *prune_cleanlab_all* achieved 96.35% accuracy and 96.27% macro-F1, which was lower than the original baseline by 0.49 percentage points for both metrics. In the additional split analysis, this strategy also remained below the baseline on average, with a mean macro-F1 of 95.69% compared with 95.93% for *train_original*.

This result highlights an important limitation of fully automated dataset cleaning. Cleanlab was effective at identifying samples with low self-confidence, but low self-confidence did not always correspond to an incorrect label. As shown by the human validation results, more than half of the Cleanlab-flagged samples were considered acceptable after visual inspection. Removing all flagged samples therefore discarded many samples that may still have been useful for training.

Compared with *prune_clear_and_ambiguous*, the fully automated removal strategy performed worse, even though both strategies removed clear errors and ambiguous samples from the training split. The key difference is that *prune_cleanlab_all* also removed acceptable samples. This supports the use of human validation to distinguish between clear errors, ambiguous samples, and acceptable low-confidence samples.

5.3.5 Sensitivity to Ambiguous Sample Weights

The *prune_clear_downweight_amb* strategy used a weight of 0.5 for ambiguous samples. To examine whether the result depended strongly on this selected weight, an additional sensitivity experiment was conducted using seed 42. In this experiment, clear errors were removed from the training set, while ambiguous samples were retained with different loss weights: 0.25, 0.50, 0.75, and 1.00.

Table 8 summarizes the results.

Among the tested weights, 0.50 achieved the highest macro-F1 score. However, the difference between weight 0.50 and weight 1.00 was small. Weight 1.00, which is equivalent to keeping ambiguous samples with normal weight after removing clear errors, achieved the same accuracy and a slightly lower macro-F1 score.

Table 8: Sensitivity analysis for different ambiguous sample weights under seed 42.

Ambiguous weight	Accuracy (%)	Macro-F1 (%)
0.25	95.58	95.48
0.50	96.27	96.26
0.75	95.85	95.77
1.00	96.27	96.19

The sensitivity experiment shows that changing the ambiguous sample weight did not produce a clear improvement over the 0.5 setting. More importantly, all tested downweighting settings performed below *prune_clear_and_ambiguous* under seed 42. This suggests that, in this experiment, reducing the influence of ambiguous samples was less effective than removing them entirely.

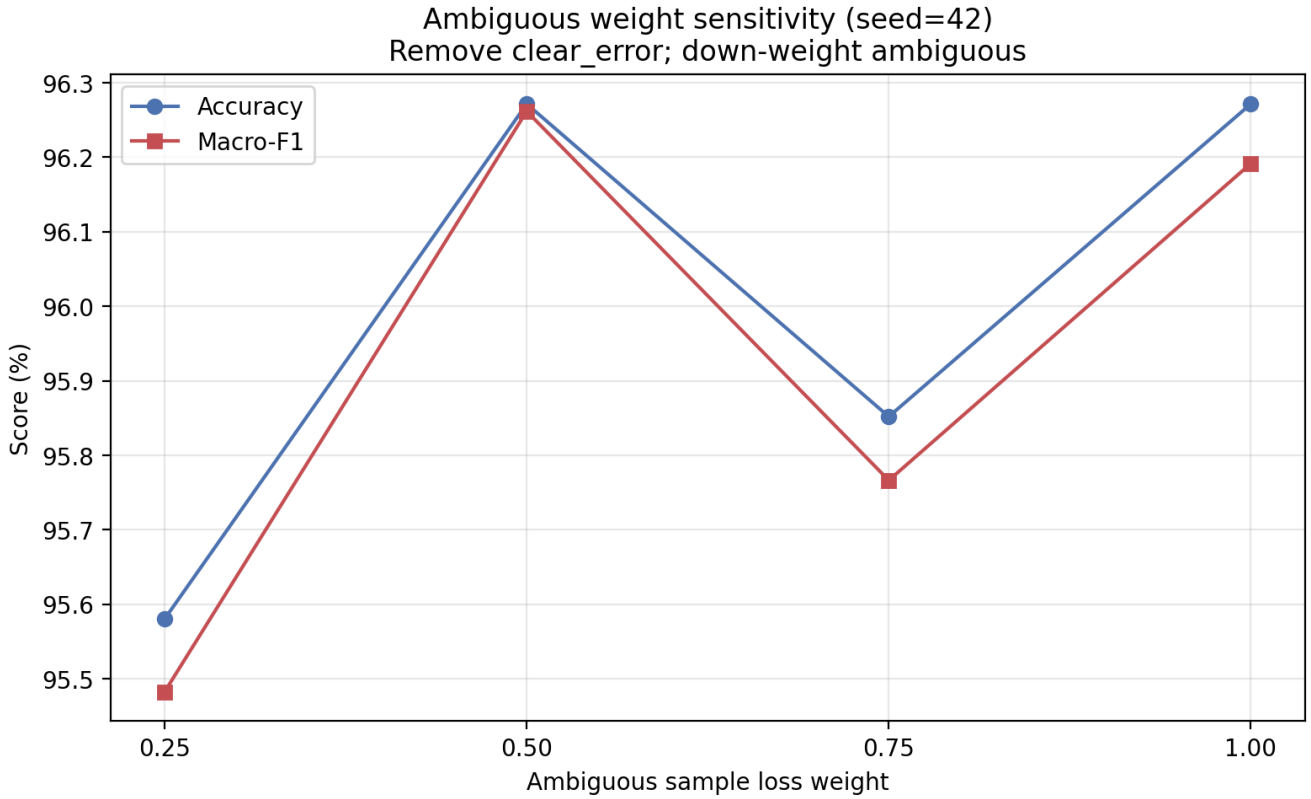

Figure 7: Accuracy and macro-F1 under different ambiguous sample weights. Down-weighting ambiguous samples did not outperform removing ambiguous samples entirely under the seed-42 setting.

This finding is central to the interpretation of the results: a low-confidence Cleanlab detection is not equivalent to a confirmed labelling mistake. More than half of the Cleanlab-flagged samples were judged to be acceptable, and about one third were judged to be ambiguous. In other words, Cleanlab was useful for narrowing the review set, but the human labels were needed to decide how each sample should be treated.

5.4 Class-wise Performance Analysis

5.4.1 Class-wise F1 Comparison

In addition to overall accuracy and macro-F1, class-wise performance was analysed to examine whether label handling strategies affected some EuroSAT classes more strongly than others. The per-class metrics were obtained from the classification reports generated for each strategy. In these reports, per-class recall corresponds to class-wise accuracy, since it measures the fraction of samples from a given class that were correctly classified.

Figure 8 shows the F1 score of each EuroSAT class under the evaluated strategies.

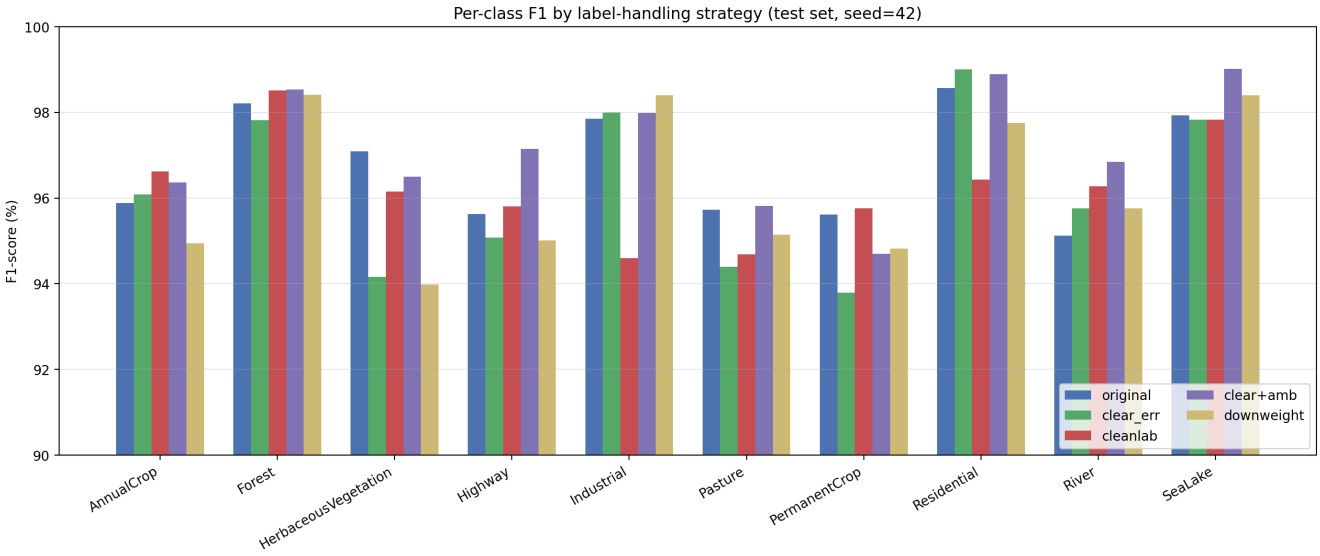

Figure 8: Per-class F1 scores for different label handling strategies.

The class-wise results show that the effect of label handling was not uniform across all classes. Some classes, such as *River*, *Highway*, and *SeaLake*, improved under the *prune_clear_and_ambiguous* strategy. Other classes, such as *PermanentCrop* and *HerbaceousVegetation*, showed small decreases compared with the original baseline.

This indicates that the improvement in overall macro-F1 was not caused by equal improvements across all categories. Instead, the performance changes were concentrated in a subset of classes.

5.4.2 Class-wise Changes Relative to Baseline

To better understand the effect of each label handling strategy, the change in per-class F1 score was computed relative to the *train_original* baseline. Figure 9 shows the F1-score difference for each non-baseline strategy.

The most important changes were observed under the *prune_clear_and_ambiguous* strategy, which was also the best-performing strategy in terms of overall macro-F1. The largest improvements occurred for *River*, *Highway*, and *SeaLake*. These classes were also visually involved in several of the Cleanlab-flagged and manually reviewed examples. In particular, *River* and *Highway* can both appear as long, narrow, and visually continuous structures in RGB satellite patches. Similarly, *SeaLake* images may contain surrounding vegetation or open land, making them visually similar to

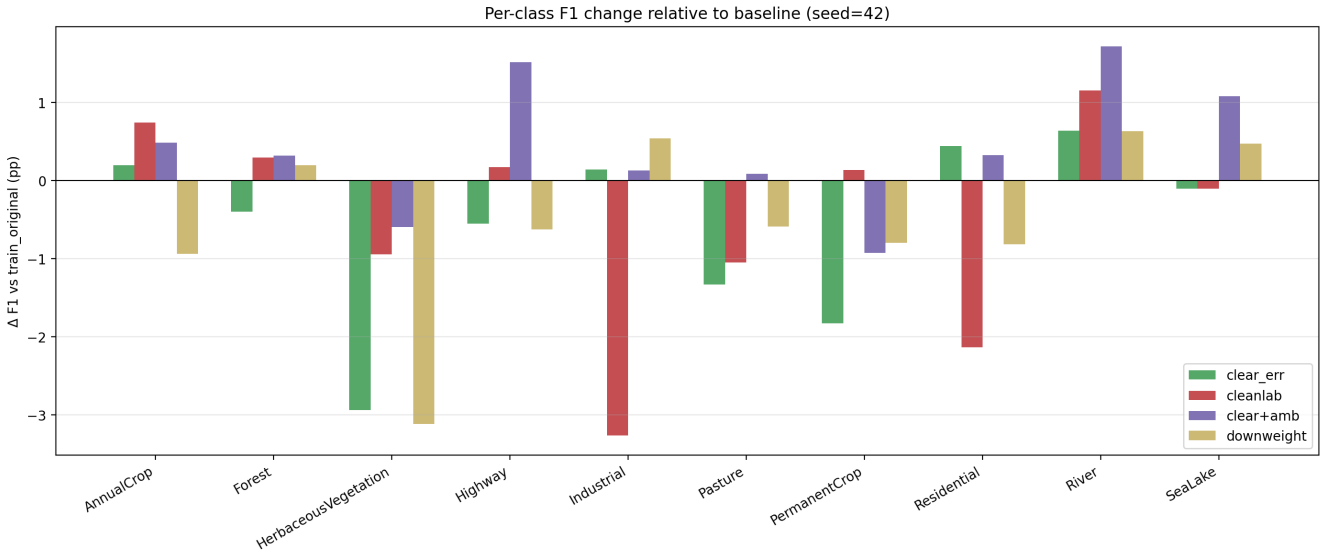

Figure 9: Change in per-class F1 score relative to the original baseline. The largest improvements under *prune_clear_and_ambiguous* occur for classes that are visually involved in several ambiguous or confused samples, especially *River*, *Highway*, and *SeaLake*.

classes such as *Pasture*. This suggests that label handling mainly affected classes where the visual boundary between categories is less clear, rather than improving all classes uniformly.

For *River*, the F1 score increased from 95.12% to 96.84%, corresponding to an improvement of 1.72 percentage points. *Highway* increased from 95.63% to 97.15%, and *SeaLake* increased from 97.93% to 99.01%. The improvement for *River* was mainly driven by recall, which increased from 90.93% under the original baseline to 98.13% under *prune_clear_and_ambiguous*. This indicates that, after removing clear errors and ambiguous samples from the training set, the model correctly identified more test images belonging to the *River* class.

However, not all classes improved. Under *prune_clear_and_ambiguous*, *PermanentCrop* decreased by 0.92 percentage points in F1 score, and *HerbaceousVegetation* decreased by 0.59 percentage points. These decreases were smaller than the gains observed for *River*, *Highway*, and *SeaLake*, but they show that label handling can shift performance between classes. This supports the interpretation that removing ambiguous samples changes class boundaries rather than simply improving the classifier across all categories.

5.4.3 Classes Most Affected by Label Handling

The classes most affected by label handling were not simply the classes with the largest number of Cleanlab-flagged samples. For example, *Pasture*, *HerbaceousVegetation*, and *PermanentCrop* contained many flagged samples, but they did not show the largest improvements under *prune_clear_and_ambiguous*. In contrast, *River*, *Highway*, and *SeaLake* showed the largest F1 improvements, even though they had fewer or more moderate numbers of flagged samples.

The number of flagged samples alone therefore does not explain the class-wise changes. What matters is also whether the flagged samples are clear errors, ambiguous boundary cases, or difficult but still valid examples. In the manual review examples, River and Highway appeared in several

ambiguous or confused cases. This is consistent with the visual similarity between these classes: both can appear as long, narrow structures in RGB satellite patches. Removing ambiguous samples involving such visually similar categories may therefore make the class boundary cleaner, which helps explain why *River* and *Highway* improved under *prune_clear_and_ambiguous*.

The improvement for *SeaLake* should be interpreted slightly differently. There were relatively few flagged *SeaLake* samples, and most were not clear annotation errors. Its improvement may therefore be an indirect effect of changes in neighbouring decision boundaries, for example by removing ambiguous samples from visually related classes such as *Pasture*, *River*, or vegetation-related categories. This again shows that the effect of label handling is not only determined by the original class of the removed samples, but also by how these samples influence the model's separation between classes.

The fully automated strategy, *prune_cleanlab_all*, had a different effect. It most strongly reduced the F1 scores of *Industrial*, *Residential*, and *Pasture*. This is consistent with the human validation results, because many Cleanlab-flagged samples were later judged to be acceptable or ambiguous rather than clear errors. Removing all flagged samples therefore also removes visually difficult but still reasonable training examples. For classes such as Industrial and Residential, fully automatic removal may discard useful difficult examples when flagged samples are acceptable or ambiguous rather than clear errors.

Overall, the class-wise analysis shows that label handling strategies affect EuroSAT classes differently. The best-performing strategy mainly improved classes involved in visually ambiguous or confused cases, such as *River*, *Highway*, and *SeaLake*, while fully automated Cleanlab removal harmed classes where many flagged samples were not confirmed as clear errors. This provides a direct link between the human review categories and the class-wise performance changes: clear errors, ambiguous samples, and acceptable low-confidence samples have different effects on the classifier and should therefore not be treated as a single type of label issue.

5.5 Additional Cleanlab Removal Experiment

5.5.1 Progressive Removal Setup

In addition to the human-validated label handling strategies, an additional experiment was conducted to evaluate the effect of removing samples directly according to the Cleanlab ranking. Unlike the main strategy comparison, this experiment did not use the human review categories. Instead, samples were removed only based on their Cleanlab rank, where lower self-confidence indicated a more suspicious sample.

Three removal levels were tested: removing the top 50 ranked samples, removing the top 200 ranked samples, and removing all 425 Cleanlab-flagged samples. These settings are referred to as *skip_50*, *skip_200*, and *skip_all*, respectively. The baseline condition used the original training set without removing any samples.

The same seed-42 split was used for all runs in this experiment. The validation and test sets were kept unchanged, and only the training set was filtered. Since some Cleanlab-flagged samples were located in the validation or test split, the number of samples actually removed from training was smaller than the ranked pool size. For example, *skip_all* used all 425 Cleanlab-flagged samples as the removal pool, but only 302 of them appeared in the training split and were therefore removed.

This experiment was designed to test whether progressively removing more low-confidence

samples would improve classification performance. If the Cleanlab ranking directly corresponded to harmful training samples, one would expect performance to improve as the most suspicious samples were removed.

5.5.2 Comparison of Removal Levels

Table 9 summarizes the results of the progressive Cleanlab removal experiment.

Table 9: Performance of progressive removal based on Cleanlab ranking.

Experiment	Rank pool	Removed	Train size	Accuracy (%)	Macro-F1 (%)
baseline	0	0	18,900	96.84	96.76
skip_50	50	33	18,867	96.12	96.04
skip_200	200	137	18,763	96.47	96.47
skip_all	425	302	18,598	96.10	96.02

None of the Cleanlab-only removal settings improved performance over the baseline. The baseline achieved 96.84% accuracy and 96.76% macro-F1. Removing the top 50 ranked samples decreased macro-F1 by 0.72 percentage points. Removing the top 200 samples produced the closest result to the baseline, with a macro-F1 decrease of 0.29 percentage points. Removing all 425 flagged samples resulted in a macro-F1 decrease of 0.74 percentage points.

Figure 10 visualizes the accuracy and macro-F1 scores for the different removal levels.

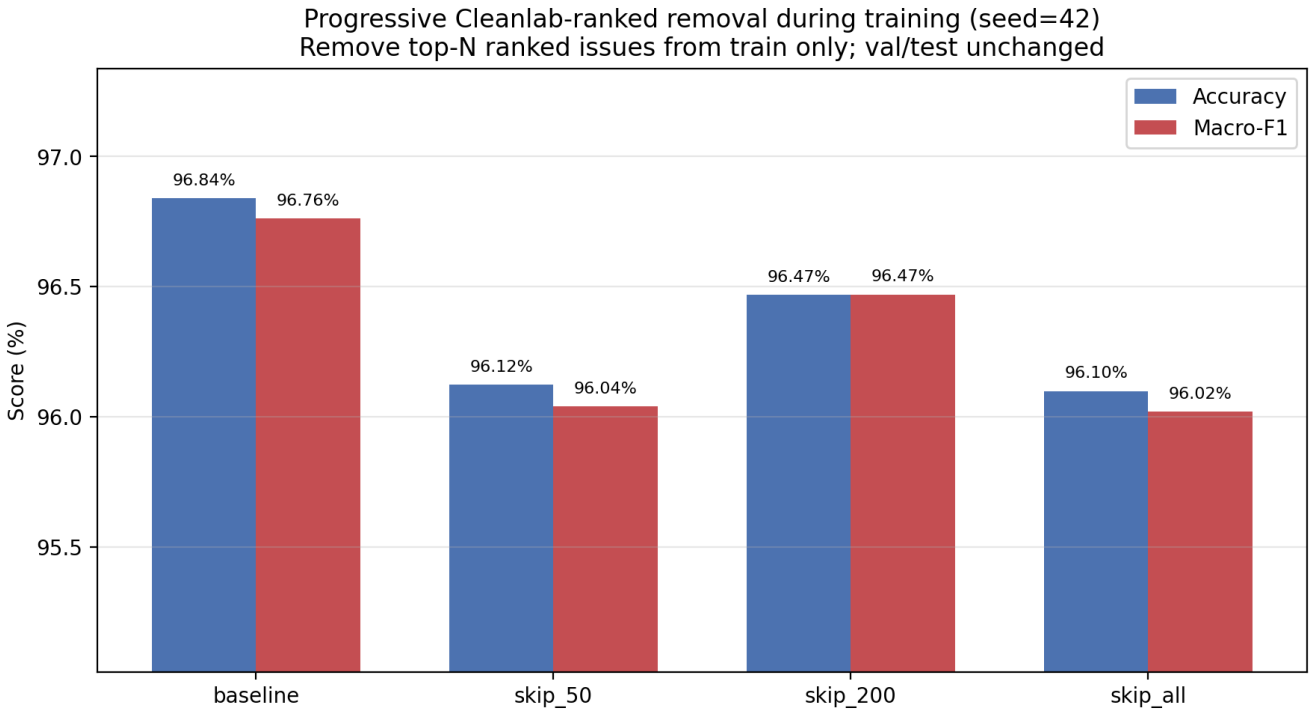

Figure 10: Accuracy and macro-F1 under progressive removal of Cleanlab-ranked samples.

The results show that removing more Cleanlab-flagged samples did not lead to a consistent improvement in performance. In fact, all three removal settings performed below the baseline. The

observed pattern indicates that Cleanlab ranking is useful for prioritizing samples for review, but is not sufficient as a standalone removal criterion.

This finding is consistent with the human validation results. Many Cleanlab-flagged samples were not clear annotation errors; more than half were reviewed as acceptable, and a substantial portion were ambiguous. Therefore, directly removing samples based only on low self-confidence can discard useful or still valid training examples.

Overall, the progressive removal experiment supports the main conclusion of the strategy comparison. Cleanlab is useful for identifying suspicious samples, but automatic removal based only on Cleanlab ranking does not necessarily improve benchmark performance. Human validation provides additional information that helps distinguish between clear errors, ambiguous samples, and acceptable low-confidence samples.

5.6 Broader Implications for EO Benchmark Reliability

The results have broader implications for how label quality should be interpreted in EO image classification benchmarks. The findings suggest that label quality problems are not limited to clear annotation errors. A low-confidence Cleanlab detection may correspond to a clear error, an ambiguous image, or a difficult but still acceptable sample. Therefore, automatic label issue detection should not be treated as equivalent to automatic error correction.

This distinction is important for the use of confident learning in EO datasets. Cleanlab was useful for narrowing down the set of samples that required inspection, but the human review showed that the same automatic signal can have different meanings. In this study, many flagged samples were not confirmed as clear errors. This means that Cleanlab should be interpreted as a candidate detection tool rather than a fully automatic dataset cleaning method.

The results also show that ambiguity is an important part of EO benchmark quality. In scene-level remote sensing classification, a single image patch can contain multiple land-cover types. As a result, some samples may be visually valid for more than one class. Such cases are not necessarily annotation mistakes; they may reflect the mixed and continuous nature of real-world land-cover scenes. Treating these samples as simple label errors would therefore oversimplify the dataset quality problem.

At the same time, ambiguous samples can still affect model training under a single-label classification setup. When a model is trained with hard labels and standard cross-entropy loss, visually ambiguous samples may introduce inconsistent supervision. This helps explain why removing ambiguous samples improved performance in the tested setup. However, this should not be interpreted as evidence that ambiguous samples are useless. These samples may contain useful information about class boundaries and mixed land-use regions, which could be better represented using soft labels, multi-label annotations, or uncertainty-aware learning methods.

These findings also matter for benchmark reliability. Benchmark datasets are often used to compare models based on small differences in accuracy or macro-F1. In this study, different treatments of the same suspicious samples led to different performance outcomes. Human-validated removal of clear errors and ambiguous samples improved performance under the tested setup, while fully automatic Cleanlab removal did not improve over the baseline. This shows that benchmark conclusions can depend not only on which samples are detected as suspicious, but also on how those samples are interpreted and handled.

Overall, benchmark reliability should not be evaluated only by counting label errors or reporting

a single overall accuracy score. The type of label issue matters. Clear errors, ambiguous samples, and acceptable low-confidence samples can affect the classifier in different ways. Human validation, class-wise analysis, and ambiguity-aware annotation schemes are therefore important for understanding how label quality affects EO benchmark evaluation.

6 Limitations and Future Work

6.1 Limitations

This study has several limitations. First, the human validation was performed by a single reviewer. As a result, inter-annotator agreement could not be measured. Since the distinction between acceptable, ambiguous, and clear-error samples can be subjective, future work should include multiple reviewers to obtain more reliable validation labels.

Second, the label handling experiments were conducted using MobileNet-V3-Large as the selected model. Although ResNet18, EfficientNet-B0, and MobileNet-V3-Large were compared as baseline models, the label handling strategies were not repeated across all model architectures. Therefore, the results show how label handling affected the selected model, but they do not fully establish whether the same trends would hold for other model architectures.

Third, the additional seed analysis should be interpreted as a split-sensitivity check rather than a strict robustness analysis. The random seed affected the train-validation-test split, meaning that the test set was not identical across seeds. Therefore, these results cannot be used as a formal repeated-run comparison on the same evaluation set. A stronger robustness analysis would require a fixed test set and repeated training runs with different random initialisations.

Finally, no formal statistical significance test was performed. The observed performance differences should therefore be interpreted as empirical trends rather than statistically proven effects.

6.2 Future Work

Future work could extend this study in several directions. First, the human validation process could be repeated with multiple independent reviewers. This would allow inter-annotator agreement to be measured and would provide a more reliable estimate of which samples are truly erroneous or ambiguous.

Second, future experiments could repeat the label handling strategies using multiple model architectures and fixed test-set repeated runs. This would help determine whether the observed trends are specific to MobileNet-V3-Large and the selected data splits, or whether they generalise across models and training runs.

Third, a separate line of future work could examine how the review outcome changes when additional information, such as multispectral Sentinel-2 bands, geographic metadata, or high-resolution reference maps, is available. This would answer a different question from the present study, which intentionally restricted human review to the RGB information available to the model.

Fourth, ambiguity-aware learning methods should be explored. Multi-label and multi-modal EO benchmarks such as BigEarthNet-MM suggest that richer annotation schemes may better represent complex remote sensing scenes [14]. Instead of removing ambiguous samples, future work could use soft labels, multi-label annotations, or uncertainty-aware loss functions. These methods may preserve useful information from ambiguous samples while reducing the negative effect of forcing a single hard label.

Finally, similar analyses could be conducted on other EO benchmark datasets. This would help determine whether the distinction between clear label errors and ambiguous samples is a general issue in EO image classification benchmarks or specific to EuroSAT.

7 Conclusion

This thesis investigated the role of label errors and ambiguous samples in the EuroSAT benchmark dataset for Earth Observation image classification. The main goal was to examine whether automatically detected label issues correspond to true annotation errors, ambiguous samples, or still acceptable samples, and to evaluate how different handling strategies affect classification performance and benchmark reliability.

A workflow based on out-of-sample prediction probabilities and Cleanlab was used to identify candidate label issues. MobileNet-V3-Large was selected as the main model after comparison with ResNet18 and EfficientNet-B0. Using 5-fold stratified cross-validation, out-of-sample probability estimates were generated for all 27,000 EuroSAT samples. Cleanlab then identified 425 suspicious samples, corresponding to 1.57% of the dataset.

The human validation results showed that the Cleanlab-flagged samples were not a homogeneous group. Among the 425 reviewed samples, only 56 samples were classified as clear annotation errors. In contrast, 149 samples were classified as ambiguous, and 220 samples were judged to be acceptable after visual inspection. This finding shows that low self-confidence does not directly imply that a sample is mislabelled. Instead, Cleanlab should be interpreted as a tool for identifying candidate label issues that require further inspection.

The label handling experiments further showed that the way in which suspicious samples are treated can influence benchmark performance. In the additional split analysis, removing both clear errors and ambiguous samples achieved the highest average macro-F1 score among the tested strategies. However, because the random seed also changed the train-validation-test split, this result should be interpreted as a split-sensitivity check rather than as conclusive robustness evidence. Fully automatic removal of all Cleanlab-flagged samples performed worse than the original baseline on average, mainly because it also removed many samples that were later judged acceptable. The progressive Cleanlab removal experiment showed a similar pattern: removing samples purely based on Cleanlab ranking did not improve performance over the baseline.

The central conclusion of this thesis is that suspicious samples identified by an automated method require further interpretation before they are used for dataset cleaning. Clear annotation errors, ambiguous samples, and acceptable low-confidence samples have different meanings and may require different handling strategies. For EO image classification benchmarks, this distinction is especially important because satellite image patches often contain mixed land-use patterns or visually similar categories.

Overall, this study shows that benchmark reliability depends not only on detecting suspicious samples but also on understanding the nature of those samples. Combining automated label issue detection with human validation provides a more informative approach to dataset quality analysis than relying on automatic detection alone. Future work could extend this approach by involving multiple reviewers, evaluating more model architectures, and exploring ambiguity-aware learning methods such as soft labels or multi-label classification.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Julia Wąsala and Mitra Baratchi, for their guidance, constructive feedback, and support throughout this thesis. Their suggestions helped me refine the research questions, experimental design, and interpretation of the results.

I would also like to thank my girlfriend (Yinchi Zhao) for discussing and helping to refine the criteria used in the human review process. Her input contributed to making the review framework clearer and more consistent.

I would also like to thank the Leiden Institute of Advanced Computer Science for providing a supportive academic environment throughout this project. Finally, I am grateful to my family and friends for their continued encouragement and support during my studies.

Code and Data Availability

The source code, experimental configurations, generated results, and manually reviewed annotations used in this study are publicly available at:

https://github.com/3000weichen/EuroSAT-Label-Quality

The repository contains the complete experimental pipeline for model training, out-of-sample probability generation, Cleanlab-based candidate detection, human review, and label handling experiments. In particular, the file `outputs/manual_review_labels.csv` provides the full list of the 425 manually reviewed candidate samples, including their original labels, model-predicted labels, review categories, and sample identifiers.

The original EuroSAT image dataset is not redistributed in the repository. Instructions for downloading and preparing the dataset are provided in the repository README.

A Baseline Model Selection

Before analysing label issues, a preliminary model comparison was conducted to select the classification model used in the subsequent experiments. This step was not intended as a main contribution of the thesis but as a controlled model selection step to ensure that the label issue detection and label handling experiments were based on a strong baseline classifier.

Table 10 shows the performance of ResNet18, EfficientNet-B0, and MobileNet-V3-Large on the EuroSAT test set.

Table 10: Baseline model performance on the EuroSAT test set.

Model	Accuracy (%)	Macro-F1 (%)
ResNet18	94.02	93.87
EfficientNet-B0	96.74	96.70
MobileNet-V3-Large	96.84	96.76

EfficientNet-B0 and MobileNet-V3-Large achieved very similar test-set performance, both exceeding 96% in accuracy and macro-F1. ResNet18 performed noticeably lower, with 94.02%

accuracy and 93.87% macro-F1. MobileNet-V3-Large obtained the highest test-set scores, although its advantage over EfficientNet-B0 was small. These test-set results are reported for completeness and were not used directly for model selection.

MobileNet-V3-Large was selected for the remaining experiments based on its validation performance and stable training behaviour during the preliminary comparison. Because the performance difference between MobileNet-V3-Large and EfficientNet-B0 was small, this selection should not be interpreted as evidence that MobileNet-V3-Large was substantially superior. Rather, it provided a strong and stable baseline for the subsequent label issue detection and label handling experiments. Using the same architecture throughout these experiments ensured that later performance differences could be attributed to the label handling strategies rather than to differences between model architectures.

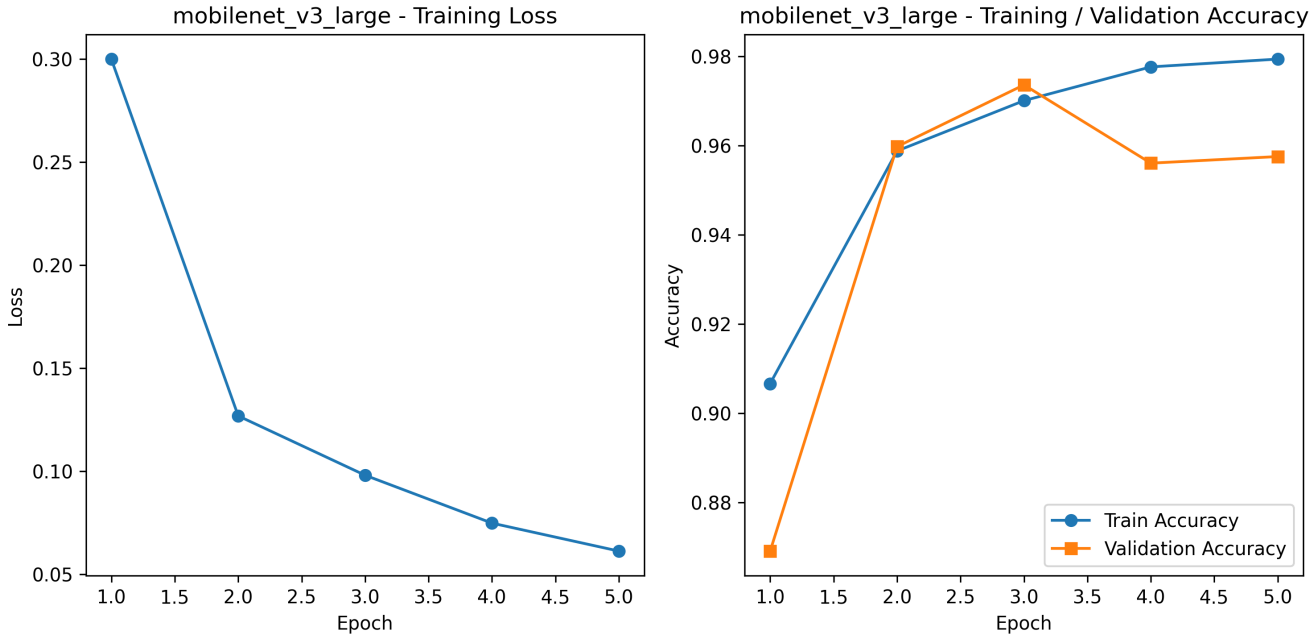

Figure 11: Training loss and accuracy curves of MobileNet-V3-Large on EuroSAT. Validation accuracy reached its highest value around the third epoch and decreased slightly afterwards, while training accuracy continued to increase.

Figure 11 shows that the training loss decreased steadily over the five epochs, while training accuracy increased throughout training. Validation accuracy reached its highest value in the third epoch and decreased slightly in the final two epochs. This pattern indicates mild overfitting after the third epoch, although the decline in validation performance remained limited. The checkpoint with the highest validation accuracy was therefore retained for test-set evaluation. The same validation-based checkpoint selection procedure was used in the subsequent experiments.

References

[1] Cesar Aybar, David Montero, Gonzalo Mateo-García, and Luis Gómez-Chova. Lessons learned from cloudsen12 dataset: Identifying incorrect annotations in cloud semantic segmentation datasets. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 3063–3066, 2023.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[3] Matthias Drusch, Umberto Del Bello, Sebastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, Aline Meygret, Fabio Spoto, Omar Sy, Francesco Marchese, and Pierluigi Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.

[7] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[8] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[10] Joshua C. Peterson, Richard M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9617–9626, 2019.

[11] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[12] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

[13] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5901–5904, 2019.

[14] Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mário Caetano, Beg"um Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.

[15] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.

[16] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[17] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.