CleanPatrick: A Benchmark for Image Data Cleaning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Robust machine learning depends on clean data, yet current image data cleaning benchmarks rely on synthetic noise or narrow human studies, limiting comparison and real-world relevance. We introduce CleanPatrick, the first large-scale benchmark for data cleaning in the image domain, built upon the publicly available Fitzpatrick17k dermatology dataset. We collect 496,377 binary annotations from 933 medical crowd workers, identify off-topic samples (4%), near-duplicates (21%), and label errors (22%), and employ an aggregation model inspired by item-response theory followed by expert review to derive high-quality ground truth. CleanPatrick formalizes issue detection as a ranking task and adopts typical ranking metrics mirroring real audit workflows. Benchmarking classical anomaly detectors, perceptual hashing, SSIM, Confident Learning, NoiseRank, and SelfClean, we find that, on CleanPatrick, self-supervised representations excel at near-duplicate detection, classical methods achieve competitive off-topic detection under constrained review budgets, and label-error detection remains an open challenge for fine-grained medical classification. By releasing both the dataset and the evaluation framework, CleanPatrick enables a systematic comparison of image-cleaning strategies and paves the way for more reliable data-centric artificial intelligence.

Benchmark: github.com/Digital-Dermatology/CleanPatrick
Data & Dataset Card: huggingface.co/datasets/Digital-Dermatology/CleanPatrick

1 Introduction

The quality of training data is a cornerstone of effective machine learning (ML), with recent trends increasingly emphasizing data-centric approaches to boost model performance [1, 2]. The quality of evaluation data is equally crucial, as contamination directly impacts how progress in the field is measured and the conclusions drawn from it [3, 4]. Currently, the evaluation of cleaning strategies, which could be used to resolve such issues, heavily relies on synthetic corruption of assumed to be clean datasets, e.g., by artificially introducing noise or mislabeling samples [3–5]. Although these controlled, synthetic setups offer repeatability, they often lack standardization since different studies adopt varied corruption protocols, making it challenging to compare results directly and benchmark progress across the literature. Furthermore, it is unclear how much a synthetic benchmark can mimic the nuances of real-world noise instead of solely favoring the authors' methods.

Several recent works have extended beyond synthetic methods to evaluate cleaning strategies on real-world contamination [3–6]. However, such evaluations tend to fall short in scope. Typically, these approaches repurpose annotations initially collected for other tasks (e.g., multi-annotator assignments), while others rely on limited-sample human evaluations. Consequently, despite these efforts, a research gap remains in establishing a robust, universally applicable benchmark for data cleaning in the image domain, and thus a clear understanding of how well these approaches perform outside of their original setting.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In contrast to the general image domain, comprehensive benchmarks exist for cleaning structured data. For example, Abdelaal et al. [7] introduced REIN, a benchmark framework for data cleaning in ML pipelines, while Li et al. [8] explored the impact of data cleaning on classification performance in their CleanML study. These initiatives have not only standardized evaluation methods but have also driven rapid progress by providing clear, comparable metrics [7, 8]. Similar endeavors in data-centric AI, exemplified by benchmarks such as DCBench [9] and DataPerf [10], highlight the potential of well-defined benchmarks in driving advancements across diverse ML tasks.

Addressing these limitations for unstructured data, we introduce CleanPatrick, the first dedicated benchmark for data cleaning in the image domain featuring exhaustive annotation for three data quality issues. CleanPatrick originates from the Fitzpatrick17k dataset [11], a collection of 16,577 dermatological disease images collected from online dermatology lexicons. By focusing on medical imaging, we create both a challenging and sustainable benchmark for data cleaning, including all peculiarities of the medical domain, such as fine-grained and long-tail distribution of labels, and importance of fine-grained details (e.g., textures). For this benchmark, the original images were repurposed and comprehensively annotated for data quality issues. Specifically, the dataset was reviewed by medical crowd workers who identified off-topic samples, near duplicates, and label errors, following terminology established by recent works in data-centric ML [4]. To achieve meaningful results for near duplicates, we selected pairs of samples based on a carefully engineered iterative procedure that requires at most as many annotations as the size of the dataset. Across the three issue types, we collected 496,377 annotations from 933 unique annotators. The resulting data-cleaning benchmark provides a realistic representation of contamination as it occurs in practice, especially when obtained through a semi-automatic procedure, moving beyond the constraints of synthetic evaluations.

By standardizing the evaluation procedure with CleanPatrick, we aim to facilitate a fair and detailed comparison of different cleaning strategies. Our benchmark not only builds on the lessons learned from structured data cleaning but also addresses the unique challenges inherent to image data, where real-world contamination is often more nuanced and complex than what synthetic corruptions can capture. Ultimately, CleanPatrick lays the foundation for future innovations in curation approaches by providing a comprehensive resource that bridges the gap between traditional synthetic evaluations and the demands of real-world applications.

When evaluating existing methods for detecting different types of data quality issues, we found that while near duplicates are relatively easy for human experts to detect, as reflected in the high inter-annotator agreement, they are challenging for existing methods, especially when they result from a part-whole relationship. Off-topic samples are difficult for current approaches to detect, but they are relatively easy for human experts to identify when given precise instructions. Label errors are both difficult for human experts to detect, as reflected by the lower inter-annotator agreement compared to the other issue types, and for current approaches, likely due to the challenging nature of the dataset. Overall, while current approaches already achieve promising results, there is still much room for improvement, especially in detecting context-dependent data quality issues, such as those present in uncurated medical imaging datasets.

In summary, the main contributions are: 1) The release of the first data cleaning benchmark for images obtained from 496,377 annotations from medical crowd workers and verified by medical experts. 2) The outline of a standardized procedure for evaluating data cleaning methods. 3) The comparison of existing approaches for detecting diverse data quality issues and an analysis of their failure cases.

2 Related work

Traditionally, the evaluation of data cleaning methods in computer vision has been carried out by introducing synthetic corruptions into believed-to-be clean datasets. For instance, frameworks such as SelfClean [4] and Confident Learning [3, 5] simulate realistic noise, mislabeled samples, or other forms of data contamination using artificial perturbations. These evaluation strategies are beneficial since they can be programmatically generated for any dataset and produce many variations. However, they inherently rely on assumptions about the nature of contamination that may not fully capture the complexity of real-world data errors.

In the domain of dermatology, several studies have reported challenges arising from real contamination in clinical image datasets. Analyses of widely used dermatological image datasets have uncovered

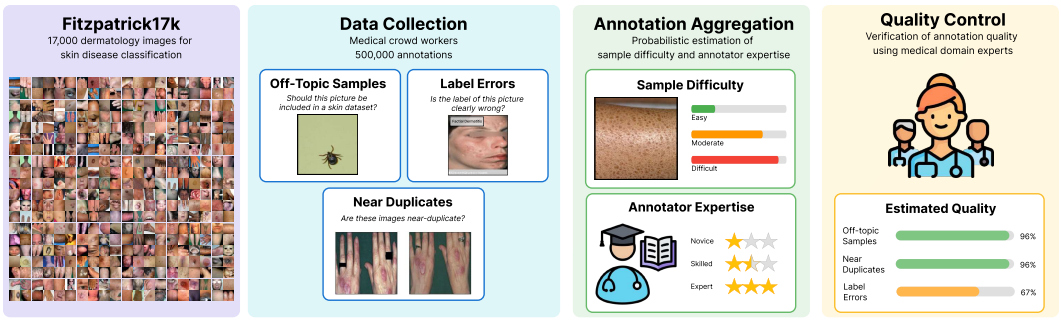

Figure 1: Process of acquiring and curating the CleanPatrick benchmark. We started by collecting annotations from medical crowd workers for three types of data quality issues. This was followed by a probabilistic estimation of sample quality and annotator expertise, used to aggregate the collected annotations. Finally, a group of medical domain experts judged the quality of a subsample of the dataset.

significant issues such as data leakage across training and testing splits, the presence of near-duplicate images, off-topic samples, and non-standardized or erroneous diagnosis annotations [12, 13]. These studies, though valuable, often rely on limited-scale human evaluation or the indirect reuse of annotations, and instead aim to release new and improved versions of established benchmarks rather than evaluating the methods used during the procedure. Illustrating the problem of using the annotations of these studies, since they rely on existing tools or heuristics to speed up data cleaning, which in turn makes a true, unbiased evaluation virtually impossible. In this paper, we instead do not use any tools or methods for performing the data selection but only rely on exhaustive annotation and a large pool of medical crowd workers to obtain unbiased annotations.

While the image domain has lagged behind in standardized evaluation procedures, the field of structured data cleaning has seen considerable progress in recent years. Benchmarks such as REIN by Abdelaal et al. [7] and CleanML by Li et al. [8] offer comprehensive frameworks and quantitative metrics for comparing cleaning methods. Similarly, tools integrated within data-centric AI platforms, such as DCBench [9] and DataPerf [10], provide a systematic evaluation environment that has spurred rapid advancements in data quality for structured data. These benchmarks have not only provided a level playing field for method comparison but have also driven progress by clearly highlighting the impact of data quality on downstream model performance. However, extending these insights to the unstructured image domain remains challenging due to the inherently different nature of visual data and the difficulty of designing detailed instructions for annotation tasks.

For label error detection, the AQuA benchmark [14] injects seven synthetic noise patterns (uniform, asymmetric, class-dependent, instance-dependent, dissenting label, dissenting worker, and crowd majority) across 14 vision, text, and tabular datasets, creating a fully controlled testbed for cleaning algorithms. Human re-annotation efforts include CIFAR-10H [15], which provides 511,400 crowd labels for the CIFAR-10 test set, yielding a soft ground-truth distribution widely used to evaluate label quality. For near-duplicate detection, Morra and Lamberti [16] released an unsupervised benchmark with verified duplicate pairs that cover common web transformations. To evaluate off-topic or out-of-distribution images, robustness benchmarks such as ImageNet-O, featuring 2,000 categories absent from ImageNet-1k, measure false-positive rates when irrelevant classes appear at test time [17].

While these specialized resources have driven progress for their error types, they assess methods in isolation and often on general-domain imagery. Furthermore, none of the existing benchmarks obtain labels directly for data quality issues. Obtaining labels by re-labeling instead of asking annotators to find quality issues leads to very different outcomes. Instead the outlined benchmark corresponds to human assessment of the issues themselves. CleanPatrick advances the field by providing expert-verified annotations for all three major data-quality problems (label errors, near duplicates, and off-topic images) within a single, clinically realistic dataset, enabling holistic evaluation and direct cross-method comparison.

3 The CleanPatrick benchmark

This section details how we transform the Fitzpatrick17k collection into CleanPatrick, a rigorously annotated benchmark for data cleaning research. We proceed as follows: Section 3.1 revisits the provenance and characteristics of Fitzpatrick17k, Section 3.2 describes the large-scale annotation campaign with medical crowd-workers, Section 3.3 explains how we aggregate the (noisy) votes with item-response theory, Section 3.4 reports the independent quality-control performed by medical domain experts, and finally, Section 3.5 formalizes the three tasks and their evaluation metrics.

3.1 Fitzpatrick17k

Fitzpatrick17k is a public collection of 16,577 clinical photographs covering 114 distinct skin disease diagnoses and labeled with Fitzpatrick skin types (I–VI) [11]. Images were obtained from two open-access dermatology atlases, DermNet (12,672 images) and Atlas Dermatológico (3,905 images), and are released under a CC BY-NC-SA 3.0 license. Compared to other dermatology datasets, it is relatively large, more diverse in both disease spectrum and skin tone distribution, and can be considered weakly supervised, as labels were extracted from atlas captions rather than obtained from structured clinical metadata. The original taxonomy groups diseases into 114 classes, although the dataset features two additional, coarser-grained levels that were obtained during post-processing. For CleanPatrick, we retain the finest granularity to preserve compatibility with prior work while keeping the data as close as possible to the originally obtained collection. The dataset has been subject to thorough analysis [18, 4, 13, 19], in which previous studies estimated between 16%–30% problematic images. This real-world noise and challenges of a medical dermatological dataset (i.e., diverse skin tones and unbalanced classes) motivated our decision to start with Fitzpatrick17k.

3.2 Annotation process

We decomposed the annotation process of the data-quality issues into three independent tasks according to their type, i.e., off-topic samples, near duplicates, and label errors, and deployed each as a separate labeling task on Centaur Labs[1], a platform that screens contributors for medical knowledge and thus has access to a large collective of medical crowd workers. Precise instructions, including examples, were formulated in collaboration with dermatologists and annotation specialists from Centaur Labs (see Appendix E). Additionally, we collected a few hundred gold standard samples for each labeling task and used them to obtain immediate feedback on the accuracy of the collection, educate annotators, and filter inattentive or adversarial raters. These gold standards were obtained from unanimous agreement of domain experts for off-topic samples and near duplicates, and from three board-certified dermatologists for label errors.

The following paragraphs summarize the labeling task description for each issue type, as provided to both crowd workers and expert annotators. Their exact formulations, including screenshots of the labeling platform, can be found in Appendix E.

Off-topic samples. The task of the annotators was to determine if a picture was included in a dataset of human skin lesions by mistake and is, therefore, off-topic in the dataset's context. For each sample, we asked the annotators *should this picture be included in a dataset of skin condition images?* where they were expected to answer with *yes* if the image correctly showed a skin condition and *no* if the image did not meet the criteria for inclusion in the dataset. Reasons to not include the image in the dataset were, for example, that the image is from a different modality (e.g., an X-ray or a PowerPoint slide with mostly text) or that the image did not focus on a skin condition (e.g., it shows a fully clothed patient without any visible skin disease). Pictures should be included if they are photos of human skin diseases. In the instruction, we additionally showed examples of typical images from other skin condition datasets and some examples of images that were likely included by mistake.

Near duplicates. The task of the annotators was to determine if two pictures, shown side by side, were near duplicates. We thus asked the annotators *are these images near-duplicates?* and asked them to answer *no* if the images were not related and *yes* if these images are transformations of one another (e.g., rotations, flips, image edits), or nearly identical because they were taken within seconds of each other, or some other reason which created a relationship among the samples. In the instruction, we additionally showed examples of both near duplicates and non-duplicate pairs.

[1] `https://www.centaurlabs.com/`, accessed on 9th of May 2025.

To avoid the prohibitive $\mathcal{O}(N^2)$ effort of exhaustively judging all $\binom{N}{2}$ image pairs in a dataset with N samples, we introduce a *fast-duplicates* procedure (see Appendix G) which speeds up the annotation process by relying on batch-wise annotation. Specifically, each image x_i is embedded with an encoder that was pre-trained on ImageNet with self-supervision (i.e., DINO [20]). Its nearest neighbor $n(x_i)$ is then retrieved, and only the at most N unordered pairs $\{x_i, n(x_i)\}$ are sent to annotation by crowd workers. This process is then repeated after the annotation is finished for the current batch of positively annotated (i.e., near duplicate) samples. Under the *fast-cleaning* assumption that every near duplicate of x_i is closer to x_i than any non-duplicate, this strategy discovers all duplicate cliques in at most $\lfloor \log_2 K \rfloor + 1$ rounds, with K being the size of the largest clique, while requiring no more than $2N$ pairwise judgements in total (see Lemma 1). In the case at hand, the procedure stopped after nine batches with duplicates and one with all negative responses, and the batch size dropped exponentially as expected (see Figure 7 in the Appendix).

Label errors. The task of the annotators was to determine if a picture was wrongly annotated. We thus showed a single picture along with its originally assigned diagnosis and asked the annotators, *is the label of this picture clearly wrong?*. The annotators were expected to answer *yes* if the diagnosis was clearly wrong for the given image and *no* if the diagnosis was not a clear label error. In the instruction, we additionally showed examples of clearly incorrect and correctly annotated samples. We explicitly mentioned that if the diagnosis for a skin lesion is likely incorrect but could be correct under special circumstances, the annotation is not clearly wrong and should not be considered as a label error, as the goal was to find errors rather than unlikely or ambiguous cases. Furthermore, since some of the diagnoses can be rare or difficult to assess, we recommended the experts to consult dermatological online atlases, such as DermWeb, if they were unsure about the condition.

With this procedure, we collected in total 496,377 binary votes from 933 medical crowd workers, with some annotating as many as 15,630 samples and some as few as 1. For each sample, we collected an average of 10 votes, with some samples having as many as 225 and others only 1. The raw annotation data, i.e., the vote of each unique annotator, can be found in the released dataset. Appendix F contains a detailed analysis of the annotations.

3.3 Label aggregation

To best leverage the wealth of annotations from medical crowd workers, we need to consider that annotators differ widely in skill and commitment. Some will be novices in dermatology, whereas others are experts, and since we have limited influence on recruitment, we should consider adversaries, i.e., annotators intentionally not solving the task. Thus, we utilize ideas from item response theory (IRT), where one can model the skill of an annotator and the difficulty of a sample, instead of assuming that all annotators and samples are equal, as typically done with majority voting. The following section describes the IRT model employed to probabilistically estimate the difficulty of a sample and the ability of an annotator, which are then used to obtain the final labels.

Let $\mathcal{Y} = \{(a, i, y_{a,i})\}$ be the set of Y noisy binary annotations collected from the medical crowd-workers through the process outlined above, where each tuple records *who* (annotator a of total A) labeled *what* (item i of total I) and the observed binary response $y_{a,i} \in \{0, 1\}$. Because most annotators label only a fraction of the items, the resulting observation matrix is sparse.

We adapt the *Generative Model of Labels, Abilities, and Difficulties* (GLAD) [21] to our setting. GLAD assumes that the probability of a correct annotation depends multiplicatively on annotator ability and item difficulty:

$$\Pr(y_{a,i} = 1 \mid c_a, b_i) = \sigma(c_a b_i), \qquad \sigma(x) = \frac{1}{1 + e^{-x}},$$

where c_a is the expertise or ability of annotator a and $b_i \in \mathbb{R}$ captures the difficulty of item i. The generative process is modeled by

$$y_{a,i} \mid c_a, b_i \sim \text{Bernoulli}\left(\sigma(c_a b_i)\right).$$

We make two key modifications to the original formulation. First, we drop the exponential parametrization of b_i such that $b_i \in \mathbb{R}$, allowing positive and negative values to encode the positive and negative latent classes, respectively. This unifies "difficulty" and "class orientation" in a single parameter, where small $|b_i|$ means difficult, and the sign of b_i reveals the latent class. Second, we choose the priors

$$c_a \sim \mathcal{N}(0, 1), \qquad b_i \sim \mathcal{N}(0, \sigma_b^2),$$

with $\sigma_b = 10^3$ giving a *vague* prior for difficulty, and the starting abilities centered around zero. Compared to the original $c_a \sim \mathcal{N}(1, 1)$, this choice reflects a more pessimistic prior due to low annotator control, since zero corresponds to chance-level performance, positive values indicate expertise, and negative values a performance below chance that possibly indicates adversarial behavior.

Variational inference. Since exact posterior inference is intractable, we use stochastic variational inference (SVI) as implemented in PYRO [22]. The mean-field variational family factorises over latent variables:

$$q(c_1, \ldots, c_A, b_1, \ldots, b_I) = \prod_{a=1}^{A} \mathcal{N}\big(c_a \mid \mu_{c_a}, \sigma^2_{c_a}\big) \prod_{i=1}^{I} \mathcal{N}\big(b_i \mid \mu_{b_i}, \sigma^2_{b_i}\big).$$

We optimize the evidence lower bound (ELBO) with Adam [23] (learning rate 0.1) for 10,000 steps, which is sufficient for convergence on all splits.

Predictive aggregation via the b_i distribution. Under the assumption that most annotators are not adversarial, the sign of b_i encodes the most likely class of sample i. To estimate the probability of a data point to belong to the positive class, we draw $M = 1{,}000$ samples $\{b_i^{(m)}\}_{m=1}^{M}$ from each b_i's variational posterior and compute

$$\bar{p}_i = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}[b_i^{(m)} > 0].$$

The distribution of the $\bar{p}_i$ for the different issue types is given in Figure 6 of the appendix. Our aggregation model is aware of uncertainties and estimates that 1.3% of near duplicates, 8.4% of label errors, and 1.3% of off-topic samples have a wrongly assigned label.

3.4 Quality control

To estimate the quality of annotations from medical crowd workers, we recruited three medical domain experts with more than five years of experience as practicing dermatologists, after they completed their medical examinations. Since a full new annotation was not possible due to resource limitations, we use a quantile-stratified random sampling scheme to ensure uniform coverage of the entire probability range $\bar{p}_i$. Specifically, we split the probabilities into 20 bins and randomly selected 20 samples from each bin, resulting in a total of 400 samples per data quality issue type. The medical experts followed the same protocol as outlined in section 3.2 and instructions as the medical crowd workers.

We computed inter-annotator reliability for the three issue types in terms of Krippendorff's α and Cohen's κ. Krippendorff's α, computed over all three raters, shows excellent agreement for near-duplicate annotations ($\alpha \approx 0.91 \pm 0.03$). It falls to a lower agreement level for off-topic samples ($\alpha \approx 0.60$ with a wide 95% CI spanning 0.20–0.85, caused by the high imbalance between problematic and non-problematic samples) and drops further for label errors ($\alpha \approx 0.42 \pm 0.06$). Consistent with Krippendorff's α, Cohen's κ values are near 0.90 for near duplicates regardless of the annotator pair, whereas they fluctuate much more for off-topic samples and exhibit very large confidence intervals for some annotator pairs. Agreement on label errors is the weakest, yet it still shows substantial agreement for such a challenging annotation task. In summary, the agreement not only shows that verification is consistent among the experts, based on standard categorization of agreement values [24], but also that the designed annotation protocol and description work as expected. This is not trivial for a task as complex as label error detection, which other studies have found to be substantially difficult [25, 26, 12]. Figure 8 in the appendix summarizes the results for the inter-annotator reliability.

In order to estimate the quality of the crowd annotations, we aggregated the expert annotations using majority voting and compared them. Experts and crowd workers agree on 96% of the verified images for off-topic samples and near duplicates, and on 67% of label errors. This further demonstrates that a carefully designed protocol is helpful to achieve high-quality annotations at scale with the help of medical crowd workers, even for complex tasks.

Additionally, we use expert annotations to estimate the threshold for obtaining the final labels for each data quality issue separately, rather than naively choosing a fixed value of 0.5. For this, we use the aggregated expert annotation and the bins we defined above, and check the distribution of labels for each one. We choose the threshold t at the bin where the distribution of positive labels starts to

increase and select the threshold as the average probability of that bin. The final label is then obtained using this estimated threshold, which is different for each issue type:

$$\hat{y}_i = \mathbb{I}[\bar{p}_i \geq t].$$

3.5 Evaluation tasks

We cast the detection of data-quality issues as a *ranking* problem. Rather than producing a binary keep/discard decision, each method must assign a real–valued score that reflects how strongly an example (or example pair) is suspected to be a data quality issue. Prior work has shown that practitioners subsequently inspect items in descending score order, making ranking the most faithful abstraction of real-world use [4, 12].

Tasks. The benchmark comprises three evaluation tasks, one for each data-quality issue type:

1. **Off–topic sample detection.**
   Input: a single image x_i.
   Output: an anomaly score $s(x_i) \in \mathbb{R}_{\geq 0}$, where larger scores indicate a higher likelihood that the image does *not* depict a skin condition.
   Positive criterion: the image is off–topic as identified by the medical crowd workers.
2. **Near-duplicate detection.**
   Input: an unordered image pair (x_i, x_j).
   Output: a similarity score $s(x_i, x_j) \in \mathbb{R}_{\geq 0}$ reflecting the confidence that the pair is a near duplicate.
   Positive criterion: the pair belongs to the same near-duplicate component identified. Note that we only compare the annotated samples, i.e., we do not treat the unannotated pairs as negative to have a more reflective performance estimation of the methods.
3. **Label-error detection.**
   Input: an image x_i together with its original diagnosis y_i.
   Output: a confidence score $s(x_i) \in \mathbb{R}_{\geq 0}$ that the assigned label y_i is *incorrect*.
   Positive criterion: medical crowd workers judged the label to be "clearly wrong".

Metrics. Methods are evaluated using standard ranking metrics, such as P@k, R@k, area under the receiver operating characteristic curve (AUROC), and average precision (AP). AUROC and AP are the primary metrics, while P@k and R@k illustrate the practical trade-offs between effort and gain for review budgets of $k \in \{100, 500, 1000\}$ images. Additionally, we report the proportion of positive samples p^+, which corresponds to the baseline AP.

4 Results

4.1 Data quality issues

Our extensive annotation process with medical crowd workers revealed numerous data quality issues in the Fitzpatrick17k dataset, as illustrated in Figure 2.

Off-topic samples constitute 613 images (4%) of the full dataset. These problematic samples fall into two main categories: *unrelated* content, including non-dermatological images such as laboratory equipment, diagrams, and completely unrelated photographs, and *low information content* images that, while potentially skin-related, lack sufficient clarity or focus to be diagnostically meaningful. These include severely blurred images, extreme close-ups with minimal context, or images where the skin condition is barely visible.

Near duplicates form a substantial portion of the dataset, with 3,556 instances (21%) out of the 15,306 annotated samples. These appear primarily as: *thumbnails*, where identical images exist at different resolutions or with minor cropping differences, and *multiple viewpoints* of the same skin condition from slightly different angles or captured moments. This redundancy artificially inflates certain diagnostic categories and may introduce data leakage between the training and test splits.

Label errors represent the most prevalent issue, affecting 3,666 images (22%) of the full dataset. These errors manifest as: *mislabelings*, where the assigned diagnostic label clearly contradicts the

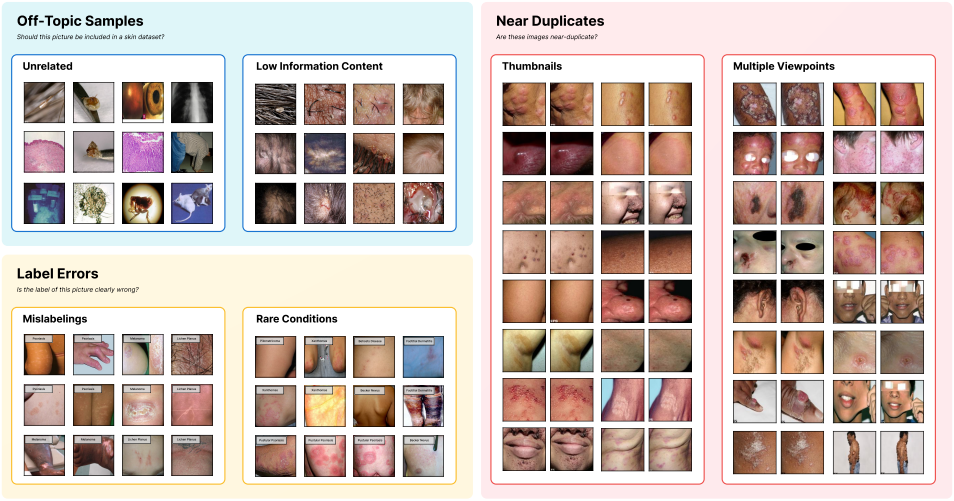

Figure 2: Examples of data quality issues identified in the Fitzpatrick17k dataset.[2] **Off-topic samples** include unrelated content (e.g., laboratory equipment, diagrams) and images with insufficient diagnostic value. **Near duplicates** comprise identical images at different resolutions (thumbnails) and multiple photographs of the same condition from different angles. **Label errors** show both clear mislabelings and rare conditions that were incorrectly classified or assigned. These naturally occurring issues form the foundation of the CleanPatrick benchmark, providing a realistic test scenario for evaluating data cleaning algorithms across varying levels of detection difficulty.

visible condition, and *rare conditions* that were incorrectly classified, likely due to their uncommon presentation or similarity to more common conditions.

The significant prevalence and diversity of these naturally occurring issues make the Fitzpatrick17k dataset an ideal foundation for a medical data cleaning benchmark. Unlike synthetic corruptions that artificially introduce noise following predetermined patterns, these issues represent authentic challenges that data cleaning algorithms must address in real-world applications. The distribution of issue types (4% off-topic, 21% near duplicates, 22% label errors) provides a comprehensive test bed that spans the spectrum of common data quality problems. This natural distribution is particularly valuable for benchmarking, as it reflects one example of contamination encountered in practice rather than artificially balanced scenarios. Furthermore, having ground truth for these issue types enables precise evaluation of detection algorithms across varying difficulty levels, from the relatively straightforward identification of off-topic samples to the more nuanced task of detecting label errors in specialized medical imagery. This comprehensive characterization of data quality issues establishes CleanPatrick as a robust, realistic benchmark for advancing data cleaning methodologies in the image domain.

4.2 Benchmark results

Figure 3 and Table 2 present the main performance metrics for each method and issue type, including AUROC, AP, and precision/recall at review budgets $k = \{100, 500, 1000\}$. Below, we describe these results and discuss their implications for real-world data-cleaning workflows.

Off-topic sample detection. Classical anomaly detectors, such as IForest, HBOS, and ECOD, achieve similar overall rankings (AUROC: 0.76–0.77, AP: 0.15–0.16). In contrast, SelfClean, a dedicated data cleaning strategy, attains an AUROC of 0.67 and AP of 0.15 but exhibits higher precision for the top 100 candidates (P@100 = 0.52) compared to the other methods. This indicates that while SelfClean's global scores are less calibrated, its highest-confidence predictions are more reliable under limited review budgets compared to classical anomaly detectors.

Near-duplicate detection. Perceptual hashing and SSIM perform near chance (AUROC $\approx$ 0.50, AP marginally above 0.31), reflecting their difficulty in capturing the subtle duplicates present in CleanPatrick. SelfClean, by leveraging self-supervised embeddings, achieves an AUROC of 0.92 and

[2] Note that the categorization is solely used for visualization and not part of the released benchmark.

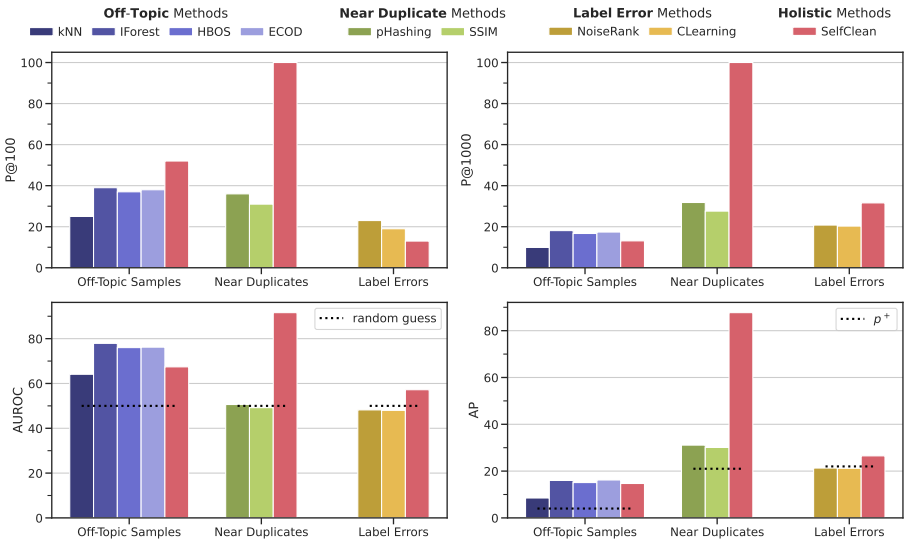

Figure 3: Performance of different data cleaning approaches (represented in colors) for the three quality issues investigated for different ranking metrics (P@100, P@1000, AUROC, and AP). Methods are separated into data quality issue-specific ones and holistic methods able to detect multiple issues. The dotted lines refer to the uninformed baseline, which randomly shuffles the ranking.

an AP of 0.88, with P@100–P@1000 = 1.00. This gain highlights the effectiveness of representation learning in duplicate detection, consistent with previous findings [27, 4].

Label error detection. Detecting misannotations in medical images remains challenging: NoiseRank and Confident Learning both hover around random performance (AUROC: 0.48, AP: 0.21, matching the base rate of 0.22). SelfClean offers a modest improvement (AUROC: 0.57, AP: 0.27) but fails to detect any true label errors in the top predictions. This suggests that label errors might require richer, context-aware signals than those available through current cleaning methods.

The difference between global ranking metrics and top-k precision highlights a critical trade-off in data-cleaning methods, namely that methods optimized for AUROC or AP may not prioritize the most egregious errors when annotation budgets are constrained. SelfClean's holistic, representation-based approach excels at surfacing high-confidence anomalies, particularly duplicates, making it well-suited for audits with budget constraints. However, its limitations in label-error detection imply that hybrid pipelines, which combine specialized, domain-aware detectors with self-supervised models, may yield better overall coverage. The persistent challenge of label noise, however, invites future research into integrating metadata, human-in-the-loop feedback, or multi-stage detection strategies.

5 Conclusion

In this work, we introduced CleanPatrick, the first benchmark for data cleaning in the image domain. Building on the publicly available Fitzpatrick17k dataset for skin disease classification, we collected nearly 496,377 annotations from 933 medical crowd workers, which were further validated through expert review. This process helped identify data-quality issues of three types: off-topic samples, near duplicates, and label errors. We formalized each detection task as a ranking problem with standardized evaluation metrics (AUROC, AP, P@k, R@k) and provided clear protocols for annotation, aggregation via a model inspired by item-response theory, and expert-driven threshold selection. In extensive experiments, we found that, on this benchmark, near-duplicate detection benefits greatly from self-supervised representations, off-topic detection is addressed well by classical anomaly detectors, achieving higher top-k precision under limited review budgets, and label-error detection remains an open challenge, with current methods performing near chance. By releasing both the CleanPatrick dataset and an accompanying evaluation framework, we provide a realistic testbed that moves beyond synthetic corruptions and captures the nuanced, real-world contamination patterns encountered in medical imaging. Future work will include releasing a revised version of Fitzpatrick17k itself.

References

[1] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[2] Boqi Chen, Cédric Vincent-Cuaz, Lydia A Schoenpflug, Manuel Madeira, Lisa Fournier, Vaishnavi Subramanian, Sonali Andani, Samuel Ruiperez-Campillo, Julia E Vogt, Raphaëlle Luisier, et al. Revisiting Automatic Data Curation for Vision Foundation Models in Digital Pathology. *arXiv preprint arXiv:2503.18709*, 2025.

[3] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Advances in Neural Information Processing Systems*, 2021.

[4] Fabian Gröger, Simone Lionetti, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, Ludovic Amruthalingam, Labelling Consortium, Matthew Groh, Alexander A. Navarini, and Marc Pouly. Intrinsic Self-Supervision for Data Quality Audits. *Advances in Neural Information Processing Systems*, 2024.

[5] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

[6] Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I Zeki Yalniz. Noiserank: Unsupervised label noise reduction with dependence models. In *European Conference on Computer Vision*, 2020.

[7] Mohamed Abdelaal, Christian Hammacher, and Harald Schoening. Rein: A comprehensive benchmark framework for data cleaning methods in ml pipelines. *arXiv preprint arXiv:2302.04702*, 2023.

[8] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In *International Conference on Data Engineering*, 2021.

[9] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *Advances in Neural Information Processing Systems*, 2022.

[10] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, et al. Dataperf: Benchmarks for data-centric ai development. *URL https://arxiv. org/abs/2207.10062*, 2022.

[11] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[12] Fabian Gröger, Simone Lionetti, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, Matthew Groh, Roxana Daneshjou, Alexander A Navarini, Marc Pouly, Labelling Consortium, et al. Towards reliable dermatology evaluation benchmarks. In *Machine Learning for Health*, 2023.

[13] Kumar Abhishek, Aditi Jain, and Ghassan Hamarneh. Investigating the quality of dermamnist and fitzpatrick17k dermatological image datasets. *Scientific Data*, 2025.

[14] Mononito Goswami, Vedant Sanil, Arjun Choudhry, Arvind Srinivasan, Chalisa Udompanyawit, and Artur Dubrawski. Aqua: A benchmarking tool for label quality assessment. *Advances in Neural Information Processing Systems*, 2023.

[15] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *International Conference on Computer Vision*, 2019.

[16] Lia Morra and Fabrizio Lamberti. Benchmarking unsupervised near-duplicate image detection. *Expert Systems with Applications*, 2019.

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. *Conference on Computer Vision and Pattern Recognition*, 2021.

[18] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y. Zou. SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Advances in Neural Information Processing Systems*, 2022.

[19] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Gin Tan, Vincent Tang, et al. A General-Purpose Multimodal Foundation Model for Dermatology. *arXiv preprint arXiv:2410.15038*, 2024.

[20] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 2021.

[21] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, 2009.

[22] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.*, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 2013.

[25] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*, 2019.

[26] Frederik Krefting, Maurice Moelleken, Stefanie Hölsken, Jan-Malte Placke, Robin Tamara Eisenburger, Lea Jessica Albrecht, Alpaslan Tasdogan, Dirk Schadendorf, Selma Ugurel, Joachim Dissemond, et al. Comparison of visual diagnostic accuracy of dermatologists practicing in Germany in patients with light skin and skin of color. *Scientific reports*, 2024.

[27] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *IEEE International Conference on Data Mining*, 2008.

[29] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 2012.

[30] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, et al. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[31] D. Marr, E. Hildreth, and Sydney Brenner. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1997.

[32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[33] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 1999.

[34] Otakar Borůvka. O jistém problému minimálním. *Práce Mor. Prırodved. Spol. v Brne (Acta Societ. Scienc. Natur. Moravicae)*, 1926.

[35] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD International Conference on Management of Data*, 2000.

[36] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, 2002.

A Limitations

Our benchmark acknowledges three limitations: (1) By presenting annotators with only each image's nearest neighbor rather than exhaustive, global pairwise comparisons, the assumption that all near-duplicates lie closer in embedding space than any non-duplicate may lead to incomplete discovery of near-duplicate groups. (2) Calibrating ground-truth thresholds using annotations from only three board-certified dermatologists may not fully capture the spectrum of clinical judgment and could introduce biases or idiosyncrasies into our final labels, highlighting a possible future research direction. And (3) as the introduced benchmark builds on a medical imaging data collection, it is challenging and less likely to be saturated quickly. However, data-cleaning strategies performing well on CleanPatrick need to bring special traits needed for medical imaging, such as the importance of fine-grained details or long-tailed labels. These traits are likely more challenging to obtain and require sophisticated methodologies.

B Resources

Experiments were performed on an NVIDIA DGX equipped with eight V100 GPUs (32 GB each), 512 GB of system memory, and 40 CPU cores, for a total of 50 GPU hours, corresponding to roughly 5.5 kg of CO_2 emissions.

In addition to computing resources, the authors financially compensated Centaur Labs to recruit and manage the medical crowd workers who provided 496,377 binary annotations across the off-topic, near-duplicate, and label-error detection tasks. This fee covered platform access, the creation and evaluation of gold-standard samples, ongoing quality-control procedures, and participant compensation via the DiagnosUs app's contest-based prize structure.

C Integration of novel methods

The ReadMe of the GitHub repository[3] details how novel methods can be integrated into the existing benchmark. The integration requires following a minimal integration of the existing data cleaning interface, which features methods for detecting the respective quality issues. We encourage people to create pull requests with novel methods to ensure a fair and transparent benchmark.

D Evaluated approaches

We evaluated different approaches to detect each of the three data quality issue categories, i.e., off-topic samples, near duplicates, and label errors. Some of these methods require encoding images in a low-dimensional latent space. For this projection, we used a vision transformer tiny pre-trained with supervision on ImageNet throughout the paper. In this section, we briefly summarize each evaluated approach, referring, however, to the original paper for more details. All hyperparameters for the evaluated approaches were kept to the default value.

D.1 Approaches for off-topic samples

Isolation Forest (IForest) isolates observations by randomly selecting a feature and splitting the value between the minimum and maximum of the selected feature. The number of splits required to isolate a sample corresponds to the path length from the root node to the leaf node in a tree [28]. This path length, averaged over a forest of random trees, is a measure of normality, where noticeably shorter paths are produced for anomalies.

Histogram-based outlier detection (HBOS) is an efficient unsupervised method that creates a histogram of the feature vector for each dimension and then calculates a score based on how likely a particular data point is to fall within the histogram bins for each dimension [29]. The higher the score, the more likely the data point is an outlier, i.e., a feature vector coming from an anomaly will occupy unlikely bins in one or several of its dimensions and thus produce a higher anomaly score.

Empirical Cumulative Distribution Functions (ECOD) is a parameter-free, highly-interpretable unsupervised outlier detection algorithm [30]. It estimates an empirical cumulative distribution function (ECDF) for each variable in the data separately. To generate an outlier score for an observation, it computes the tail probability for each variable using the univariate ECDFs and multiplies them together. This calculation is done in log space, accounting for each dimension's left and right tails.

[3] github.com/Digital-Dermatology/CleanPatrick, accessed on 9th of May 2025.

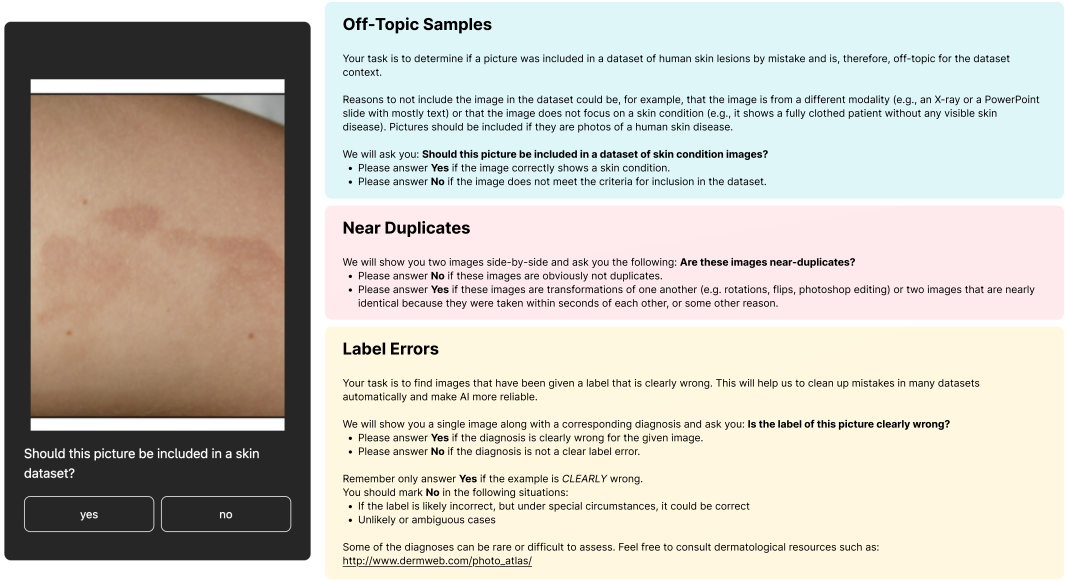

Figure 4: Left shows a screenshot of the labeling interface shown to the medical crowd workers. Right shows the instructions given to the annotators for the respective labeling tasks for the data quality issues. Along with each set of instructions, the annotators were given some example images of both the positive and negative responses.

D.2 Approaches for near duplicates

Perceptual Hash (pHashing) is a type of locality-sensitive hash, which is similar if features of the sample are similar [31]. It relies on the discrete cosine transform (DCT) for dimensionality reduction and produces hash bits depending on whether each DCT value is above or below the average value. In this paper, we use pHash with a hash size of 8.

Structural Similarity Index Measure (SSIM) is a type of similarity measure to compare two images with each other based on three features, namely luminance, contrast, and structure [32]. Instead of applying SSIM globally, i.e., all over the image at once, one usually applies the metrics regionally, i.e., in small sections of the image, and takes the mean overall. This variant of SSIM is often called "Mean Structural Similarity Index". In this paper, we apply SSIM locally to 8x8 windows but still refer to the method as SSIM for simplicity.

D.3 Approaches for label errors

Confident Learning (CLearning) is a data-centric approach that focuses on label quality by characterizing and identifying label errors in datasets based on the principles of pruning noisy data, counting with probabilistic thresholds to estimate noise, and ranking examples to train with confidence [5]. It builds upon the assumption of a class-conditional noise process to directly estimate the joint distribution between noisy (given) and uncorrupted (unknown) labels, resulting in a generalized learning process that is provably consistent and experimentally performant. In this study, we use AdaBoost [33] as a classifier on top of pre-trained representations to estimate probabilities. We did not observe any significant performance difference when using different classifiers similarly to Northcutt et al. [5].

NoiseRank (Noise) is a method for unsupervised label noise detection using Markov Random Fields [6]. It constructs a dependence model to estimate the posterior probability of an instance being incorrectly labeled, given the dataset, and then ranks instances based on this probability.

D.4 Approaches for multiple issue types

SelfClean leverages context-aware self-supervised embeddings learned on the contaminated dataset and employs simple distance-based indicators in that latent space, i.e., clustering for off-topic detection, nearest-neighbor distances for near-duplicates, and class-wise distance comparisons for label errors, to rank and score samples for inspection [4]. The methodology is intended to be used with a human in the loop, where top-ranked issues are validated. However, it can also be used fully automatically by thresholding based on estimated contamination.

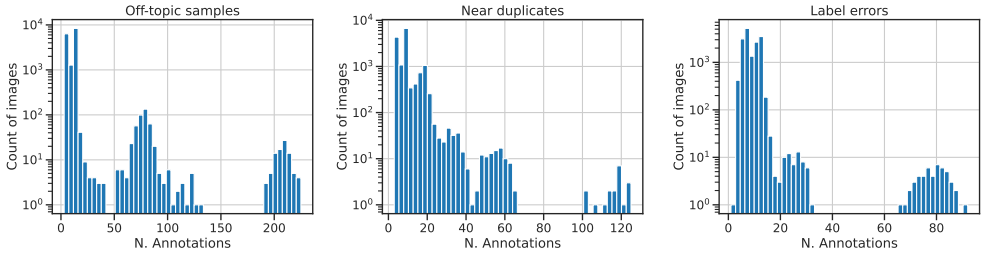

Figure 5: Histograms showing the number of annotations from medical crowd workers per image sample for each data quality issue.

E Details to Annotation Platform

Figure 4 shows a screenshot of the annotation platform of Centaur Labs, specifically of the DiagnosUs app[4], used for obtaining annotations. Additionally, it also shows the instructions given to the medical crowd workers and expert annotators.

DiagnosUs is a free app where annotators voluntarily opt in to contests, where medical image annotations are completed by labelers competing in these challenges. Labelers' submissions are scored based on their accuracy against a set of gold standard cases. Labelers who achieve high accuracy and are placed on the leaderboard are compensated with monetary prizes. Prize amounts vary depending on the contest structure, ranging from approximately $0.50 to $20 per prize.

F Detailed analysis of data quality issues

Annotation counts. Figure 5 illustrates the distribution of annotation counts per sample across the three issue types: off-topic samples, near duplicates, and label errors. On average, each image received 10 medical crowd worker votes, with extremes ranging from a single annotation to as many as 225. The vast majority of samples fall between 5 and 20 annotations.

Notably, *only one sample* ended up with a single annotation. This occurred because a medical crowd worker mistakenly flagged the image early in the process, causing it to be excluded from subsequent annotation rounds. To ensure no gap in quality, the authors manually reviewed this outlier in full and confirmed its correct classification in the final benchmark.

Near-duplicate components. Beyond per-sample vote counts, we also examined how near-duplicate samples group into connected components under our fast-duplicate detection procedure. We discovered 2,389 separate components of size ≥ 2. Their size distribution is shown in Table 1. This distribution shows that small components (pairs and triplets) dominate the duplicate structure, while only a handful of larger clusters (size ≥ 10) exist.

Table 1: Counts of near-duplicate components by size

Component Size	Number of Components
2	1997
3	169
4	151
5	19
6	26
7	8
8	9
10	4
11	2
12	2
25	1
30	1

[4] `https://www.diagnosus.com/`, accessed on 9th of May 2025.

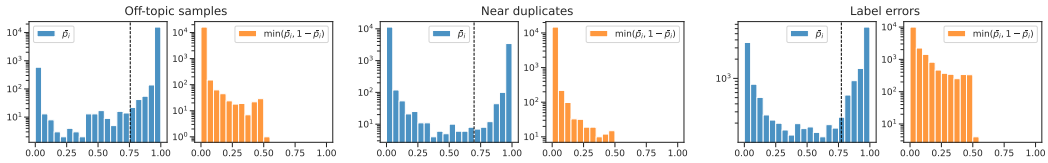

Figure 6: Distribution of per-sample annotation uncertainty, measured as $\min\big(\bar{p}_i,\, 1-\bar{p}_i\big)$, for the three issue types: off-topic samples (left), near duplicates (center), and label errors (right). Here, $\bar{p}_i$ is the estimated probability that sample i belongs to the positive class under the GLAD model. Vertical dashed lines indicate the expert-calibrated thresholds $t_{\mathrm{OT}} = 0.76$, $t_{\mathrm{ND}} = 0.70$, and $t_{\mathrm{LE}} = 0.77$ used to produce the final labels.

G Fast near duplicates

Near duplicates. Let $\mathcal{D} = \{1, 2, \ldots, |\mathcal{D}|\}$ be a dataset with samples labeled by consecutive integers. A sample pair $\{i, j\}$ corresponds to an edge in a graph with vertices $\mathcal{D}$. Verifying all near-duplicate pairs in $\mathcal{D}$ is equivalent to annotating each edge in the complete graph, and yields the subgraph $\mathcal{D}_\sim$ induced by the binary near-duplicate relation $\sim$.

Automatic cleaning. Consider a function $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ that associates a unique weight $d(i, j)$ to each edge, $d(i, j) \neq d(k, l)$ for any $i, j, k, l \in \mathcal{D}$. To keep the language and notation intuitive, we assume d is a distance as is the case in this work, but one can still obtain a similar procedure *mutatis mutandis* if this is not the case. Near-duplicate detection could be easily performed exactly if the following property were to hold.

Assumption 1 (Automatic Cleaning) *All near-duplicate pairs have a distance which is less than any non-near-duplicate pair, i.e.*

$$\forall i, j, k, l \in \mathcal{D} \mid i \sim j \wedge k \not\sim l, \quad d(i, j) < d(k, l). \tag{1}$$

In this case, there is a threshold d_* such that all near-duplicate pairs have a smaller distance, and any pair with larger distance is not a near duplicate. In practice, such a perfect ranking is very difficult to find, and one has to resort to hybrid methods which require human verification. This generates the burden of annotating all sample pairs, which grow quadratically with the dataset size.

Fast cleaning. To determine which samples are potentially related to each other, it is sufficient to partition the samples according to which subgraph they belong to, without necessarily knowing every pairwise relation. For this task, the poor scaling of manual verification can be significantly alleviated with the help of a function that satisfies the following, weaker condition.

Assumption 2 (Fast cleaning) *Near duplicates of a sample are closer to it than other samples, i.e.*

$$\forall i, j, k \in \mathcal{D} \mid i \sim j \wedge i \not\sim k, \quad d(i, j) < d(i, k). \tag{2}$$

As a heuristic side note, we observe that this assumption is substantially more *local* than assumption 1, as it only requires the distance to correctly sort near duplicates in the neighborhood of the specific sample i.

To exploit the fast cleaning assumption, one may proceed analogously to Borůvka's algorithm for minimal spanning trees [34]. For every sample $i \in \mathcal{D}$, find its nearest neighbor $n(i) \in \mathcal{D} \setminus \{i\}$ to build the set of neighbor pairs $\mathcal{N} = \{i, n(i)\}_{i \in \mathcal{D}}$. Checking all of them takes $|\mathcal{N}|$ annotations and gives the set of near duplicates $\mathcal{P}$. By virtue of assumption 2, all samples which do not appear in $\mathcal{P}$ have no near duplicates. The pairs in $\mathcal{P}$, instead, are edges that belong to the subgraph $\mathcal{D}_\sim$. The connected components of $\mathcal{P}$ partition its vertices in a set of clusters $\mathcal{D}_1$. Because of assumption 2, two such subsets belong to the same connected component of $\mathcal{D}_\sim$ if and only if their two elements with the smallest distance are near duplicates. Therefore, it is now sufficient to annotate the nearest-neighbor pairs $\mathcal{N}_1$ within $\mathcal{D}_1$ and iterate the procedure. This is guaranteed to exactly identify the connected components of $\mathcal{D}_\sim$ once no more near duplicates are found.

Cleaning complexity.

When d is symmetric and there are no ties, the number of connected components in the subgraph of nearest neighbors $\mathcal{N}$ is exactly equal to $|\mathcal{D}| - |\mathcal{N}|$, i.e., the number of duplicated nearest-neighbor edges. Indeed, each sample appears in at least one edge, so it belongs to a component which connects it to its nearest neighbor, then to the nearest neighbor thereof, and so on. However, there can be no cycles in the nearest neighbor subgraph $\mathcal{N}$, else the first sample would have connected to the last instead of the second. Any such tree therefore terminates with two samples which are reciprocally the closest to each other. The number of undirected edges that appear twice in the list $\{i, n(i)\}_{i \in \mathcal{D}}$ is therefore the number of connected components in $\mathcal{N}$, and can be expressed as $|\mathcal{D}| - |\mathcal{N}|$.

After the i-th iteration, one can scan the verified near-duplicate clusters $\mathcal{D}_i$ obtained in the last step (which have size larger than 2^i), and find the set $\mathcal{N}_i$ of the closest sample pairs that belong to different clusters. The number of nearest neighbors pairs to verify is $|\mathcal{N}_i|$ and leaves $|\mathcal{D}_{i+1}| \leq |\mathcal{D}_i| - |\mathcal{N}_i|$ new clusters that require another iteration. This terminates when the k-th iteration generates no new subsets, $|\mathcal{D}_k| = 0$. Since the size of the new subgraphs $\mathcal{D}_i$ is at least double at each iteration, one has $k \leq \lfloor \log_2 K \rfloor + 1$ where K is the size of the largest subgraph and clearly $K \leq |\mathcal{D}|$. Manipulating inequalities to have the $|\mathcal{N}_i|$ terms on the left side, the total number of annotations clearly satisfies

$$|\mathcal{N}| + |\mathcal{N}_1| + |\mathcal{N}_2| + \cdots + |\mathcal{N}_{k-1}| \leq (|\mathcal{D}| - |\mathcal{D}_1|) + (|\mathcal{D}_1| - |\mathcal{D}_2|) + \cdots + |\mathcal{D}_{k-1}| = |\mathcal{D}|. \tag{3}$$

We thus have the following guarantee:

Lemma 1 *Finding all near duplicate clusters under assumption 2 requires annotating at most $|\mathcal{D}|$ sample pairs in at most $\lfloor \log_2 K \rfloor + 1$ iterations.*

Comment on transitivity One may be tempted to think that near duplicates correspond to an equivalence relation as follows.

Assumption 3 (Near-duplicate equivalence) *The near-duplicate relation $\sim$ satisfies*

1. *$i \sim i$ (reflexive)*
2. *$i \sim j \Rightarrow j \sim i$ (symmetric)*
3. *$i \sim j \wedge j \sim k \Rightarrow i \sim k$ (transitive)*

for any $i, j, k \in \mathcal{D}$.

However, this is clearly not true in practice. A counterexample are the frames of a video that was captured without interruptions but features two very different situations at the beginning and at the end. While each two consecutive frames are near duplicates, it is a question if the first and the last frames taken alone should be considered near duplicates. For this reason, it is in general better to always consider merging clusters based on the two most similar samples, i.e., using single linkage.

H Detailed plots and results

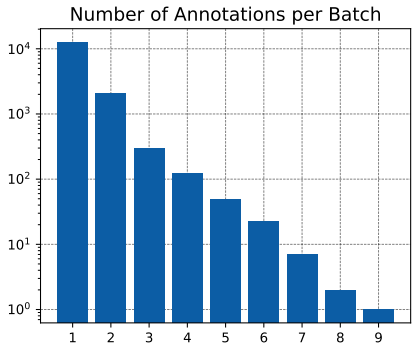

Figure 7: Number of duplicates per batch of annotation. For each batch, we select the closest pairs which has been positively identified as near duplicates in the batch before and start by taking the closest pair for every sample in the dataset.

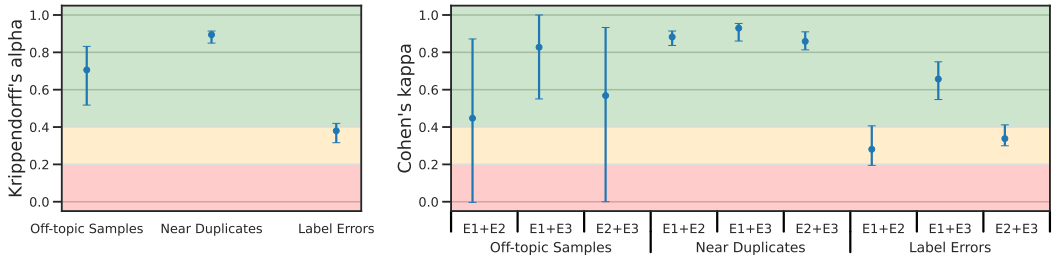

Figure 8: Inter-annotator agreement as Krippendorff's alpha among all expert annotators (left) and Cohen's kappa for all expert annotator pairs (right). Markers identify the six selected evaluation datasets, error bars are 95% confidence intervals obtained by bootstrapping annotated samples, and the background color indicates the degree of agreement [24].

Table 2: Detailed performance of evaluated approaches on the CleanPatrick benchmark. Consult appendix D for details on competing approaches. Results are given in percentages (%).

	Method	p^+	P@100	P@500	P@1000	R@100	R@500	R@1000	AUROC	AP
Off-Topic Samples	kNN [35, 36]	3.7	25.0	15.0	9.9	4.1	12.2	16.2	63.6	8.5
	IForest [28]	3.7	39.0	22.6	18.1	6.4	18.4	29.5	77.3	15.9
	HBOS [29]	3.7	37.0	20.2	16.7	6.0	16.5	27.2	75.5	15.2
	ECOD [30]	3.7	38.0	21.2	17.4	6.2	17.3	28.4	75.7	16.2
	SelfClean [4]	3.7	52.0	21.0	13.1	8.5	17.1	21.4	66.9	14.5
	Method	p^+	**P@100**	**P@500**	**P@1000**	**R@100**	**R@500**	**R@1000**	**AUROC**	**AP**
Near Duplicates	pHashing [31]	21.4	36.0	31.0	31.8	0.7	2.9	6.0	50.5	31.6
	SSIM [32].	21.4	31.0	28.0	27.6	0.6	2.7	5.2	49.1	30.5
	SelfClean [4]	21.4	100.0	100.0	100.0	1.9	9.5	19.0	91.7	87.9
	Method	p^+	**P@100**	**P@500**	**P@1000**	**R@100**	**R@500**	**R@1000**	**AUROC**	**AP**
Label Errors	NoiseRank [6]	22.1	23.0	19.0	20.8	0.6	2.6	5.7	48.4	21.3
	CLearning [5]	22.1	19.0	20.4	20.3	0.5	2.8	5.5	47.9	21.3
	SelfClean [4]	22.1	13.0	26.2	31.6	0.4	3.6	8.6	57.2	26.5

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, all claims regarding the benchmark are backed by careful evaluation of related work (see Section 2) and systematic, redundant annotations (see Section 3), and all claims regarding methods' performance are backed up by empirical results (see Section 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discusses limitations in Appendix A, and the main limitations, e.g., that evaluation of data cleaning methods applies to a single benchmark, are clearly formulated in the main body.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, the Appendix G includes the full set of assumptions and proofs of the fast near duplicate cleaning procedure, and this is the only non-empirical contribution of the work.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper discusses data collection and evaluation exhaustively in Section 3, and the evaluated approaches in detail in Appendix D. Additionally, since this paper accompanies the release of a benchmark dataset, we are releasing fully reproducible evaluation code used to obtain the results from the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, the paper contains open access to the accompanying benchmark dataset[5] and evaluation code[6].

   Guidelines:

[5] huggingface.co/datasets/Digital-Dermatology/CleanPatrick, accessed on 9th of May 2025.
[6] github.com/Digital-Dermatology/CleanPatrick, accessed on 9th of May 2025.

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper includes all details of the evaluation in Section 3.5 and features training details in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, the results do not feature error bars as no training is performed on the benchmark itself, and evaluation is unsupervised and out-of-domain of a dataset which has very likely never been seen by evaluated methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, the paper features an appendix dedicated to the computing resources required for experiments (see Appendix B).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Yes, the research is conform with every aspect of the code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The released dataset is designed to benchmark current and future image data cleaning tools, which can have positive societal impacts, especially in the medical domain. Negative societal impacts can include favoring biased tools against minority groups, which we discuss and investigate in Appendix F.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Yes, since we build on an existing dataset, namely Fitzpatrick17k, we solely release additional metadata that can be merged with the existing dataset. We do not release any new images which could potentially be misused. The risk of re-identification of crowd worker annotators is minimal, as it is impossible to trace them from binary labels without extensive knowledge about extremely specific opinions in dermatology.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, the authors of the paper are the original owners of all the assets that are released, including code and data. Code and data are both licensed under Creative Commons Attribution Non Commercial 4.0 and are labeled as such in both repositories.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, the released benchmark dataset is well documented both in the paper and the corresponding Hugging Face repository[7]. Annotations from all people in the study were anonymized, and consent was obtained from medical crowd workers through Centaur Labs, which was commissioned to obtain the annotations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

[7] huggingface.co/datasets/Digital-Dermatology/CleanPatrick, accessed on 9th of May 2025.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, the paper features a dedicated section in the appendix containing both the instructions shown to the annotators and screenshots of the labeling platform (see Appendix E). Additionally, the section details information about compensation for the annotators.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: During annotation, we solely collected answers as binary labels along with anonymized annotator identification. Thus, these annotations contain no personally identifiable information or offensive content. In discussion with experts from the institutions of the co-authors, it was concluded that this verification process does not require IRB approval because the conducted study examines publicly available datasets and does not involve human subjects beyond binary annotations.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not involve the development of any LLMs, nor their use beyond minor editing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.