Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Label Cleaning in Multimodal Public Retinal Image Datasets

Haoyu Chen
haoyuchen@cuhk.edu.hk
The Chinese University of Hong Kong https://orcid.org/0000-0003-0676-4610
Yongleng Fang
tian lin
Tingkun Shi
Joint shantou international eye center https://orcid.org/0000-0003-2382-817X
Qi Zhang
zhangq@jsiec.org
Zhuangling Lin
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science
Xixuan Ke
zhangq@jsiec.org
Guiqin He
CHAOXIN ZHENG

Article

Keywords: Retinal image, Public datasets, Label noise, Dataset cleaning, Cleanlab, Confident learning

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10775714/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: There is no conflict of interest

Abstract

Background

Public datasets are widely used to develop and validate artificial intelligence models, but label noise may compromise model performance and benchmark reliability. This study systematically assessed and corrected label noise in public multimodal retinal image datasets using automated screening and structured expert verification.

Methods

Nine public retinal image datasets spanning color fundus photography, optical coherence tomography, and ultra-widefield fundus imaging were included. After preprocessing, 34 924 images were analyzed using Cleanlab's Datalab module to identify near-duplicates, suspected label issues, and outlier samples. All flagged images underwent a four-stage manual verification process.

Results

Perceptual hash verification confirmed 638 near-duplicate images across six datasets. Cleanlab flagged 4 434 images (12.70%) for expert review, of which 1 884 (5.39% of all images) were confirmed as label errors. Dataset-level positive predictive values for label error detection ranged from 1.50% to 57.14%. Agreement between Cleanlab-suggested and expert-verified labels ranged from 51.95% to 100.00%, exceeding 60.00% in seven datasets. Among confirmed errors, 1 435 were in-distribution corrections, 103 required multi-label assignment, and 400 were out-of-distribution (OOD) cases, including rare diseases and poor-quality images. After accounting for overlap, 449 cases (23.83%) could not be represented within the original single-label classification framework.

Conclusions

Label noise is prevalent in public retinal image datasets and includes a substantial proportion of OOD and multi-label cases that cannot be fully addressed within conventional single-label classification frameworks. Cleanlab effectively reduced manual review burden and provided useful correction suggestions within the source taxonomy. This workflow enabled the creation of cleaner multimodal retinal datasets.

Summary Box

What was known before:

1. Public retinal image datasets are widely used for artificial intelligence development and benchmarking, but label errors, duplicates, out-of-distribution cases, and mismatches between single-label taxonomies and real-world retinal findings can undermine model evaluation.

2. Previous studies of retinal image label quality have mainly focused on individual datasets or specific types of label noise. The performance and limitations of automated quality control across different retinal imaging modalities have not been well established.

What this study adds:

1. This study systematically quantified and corrected multiple forms of label noise across nine public multimodal retinal image datasets using automated screening combined with structured expert verification.

2. This study provides cleaner multimodal retinal image datasets that can support more reliable downstream model training, validation, and benchmarking in ophthalmic artificial intelligence research.

1. Introduction

Retinal diseases, including diabetic retinopathy, age-related macular degeneration, and myopic retinopathy, are among the leading causes of irreversible visual impairment worldwide. The Global Burden of Disease Study estimated that 43.3 million people were blind and 295 million had moderate or severe vision impairment in 2020, with numbers expected to rise as populations continue to age[1, 2]. Retinal imaging plays an important role in the screening, diagnosis and monitoring of these diseases. Three modalities are most commonly used. Color fundus photography (CFP) provides two-dimensional visualization of the posterior pole and is widely used in both screening programs and clinical settings[3, 4]. Optical coherence tomography (OCT) enables high-resolution cross-sectional images of retinal structures[5, 6]. Ultrawide field (UWF) imaging captures up to 100° or more of the retina in a single acquisition, facilitating the detection of peripheral retinal abnormalities[7, 8].

The widespread use of retinal imaging has facilitated the development of deep learning models for automated disease detection and classification[9, 10]. Recent studies have shown that artificial intelligence (AI) systems can achieve specialist-level performance across retinal disorders[11–13]. Public datasets are the primary resources for deep learning training and evaluation. Some of them, such as EyePACS, ODIR, and IDRiD have been cited in hundreds of articles[14]. Despite their widespread use, dataset labels are rarely systematically evaluated before downstream research.

Label noise, defined as incorrect, ambiguous, or inconsistent class annotations, represents a major challenge in medical image analysis[15]. In medical imaging, label noise may arise from inter-observer variability, ambiguous category boundaries, annotation fatigue during large-scale grading, and errors introduced during data collection and curation[14, 16]. Even relatively low levels of label noise can substantially affect model performance. For example, a label error rate of only 1.2% in a national diabetic retinopathy screening dataset was reported to alter model sensitivity by 12.5%, potentially changing the clinical

interpretation of benchmark results[17]. Previous investigations have identified label noise rates ranging from 0.15% to 38.5% across commonly used retinal datasets, with some estimates suggesting error rates approaching 40% in EyePACS[18]. Because these datasets are frequently reused across multiple studies, undetected annotation errors may propagate throughout the literature and influence the development, validation, and comparison of AI models[19, 20].

Cleanlab is an open-source data-centric AI framework based on the theory of confident learning, designed to identify label issues and improve dataset quality[21]. Its Datalab module detects suspected label errors, near-duplicates, and outliers. In previous experiments involving CFP and OCT datasets with synthetically injected label noise ranging from 0% to 70%, Cleanlab improved label quality and restored downstream RETFound classification performance by as much as 52.9%[18]. On the public EyePACS and APTOS-2019 datasets, Cleanlab outperformed alternative data-cleaning approaches in identifying mislabeled samples[18]. However, these evaluations were largely limited to a single disease domain and relied primarily on artificially generated label noise. In real-world datasets, annotation errors are often structured, class-dependent, and affected by dataset imbalance[22, 23]. Consequently, the effectiveness of Cleanlab in detecting naturally occurring label errors across diverse retinal imaging modalities remains insufficiently understood.

In this study, we applied Cleanlab to nine publicly available retinal image datasets encompassing CFP, OCT, and UWF imaging. All images flagged by the algorithm underwent a four-stage expert verification process to establish reference labels and characterize error types. Our objectives were threefold: (1) to quantify the prevalence of label noise in real-world multimodal retinal imaging datasets; (2) to evaluate the performance of Cleanlab in identifying and correcting annotation errors; and (3) to investigate the pattern of label error.

2. Materials and methods

2.1 Dataset Selection

This study incorporated nine public retinal image datasets, which were selected to provide broad coverage across three imaging modalities (CFP, OCT, UWF), diverse disease categories and dataset sizes. Selection criteria prioritized datasets that were publicly accessible, widely used in retinal artificial intelligence research, and suitable for single-label classification tasks. The CFP modality comprised the Ocular Disease Intelligent Recognition dataset (ODIR)[24], the Multi-label Retinal Diseases dataset (MuReD)[25], and the Eye Diseases Classification dataset (EDC)[26]. The OCT modality included the OCT2017 dataset (OCT2017)[12], the OCT Image Database (OCTID)[27], and the OCT Deep Learning dataset (OCTDL)[28]. The UWF datasets consisted of the UWF Fundus Images of Intraocular Tumors dataset (TumorUWF)[29], the Open Ultra-Widefield Fundus dataset (OpenUWF)[30], and the Tsukazaki Hospital UWF dataset (TsukazakiUWF)[31, 32]. Dataset characteristics are summarized in Supplementary Table 1.

2.2 Data Preprocessing

Preprocessing ensured compatibility with Cleanlab and reliable single-label classification. In ODIR and MuReD, images assigned to the other category were excluded (ODIR: n = 708; MuReD: n = 148), as this category aggregates heterogeneous conditions that lack consistent diagnostic definitions and are therefore unsuitable for single-label classification. In TsukazakiUWF, images labeled AO and DM were removed (n = 1 071) because these categories were not clearly defined in the original dataset documentation. The OCT2017 contains multiple serial images acquired from individual patients. To avoid overrepresentation of patients and maintain consistency with previous studies, only the horizontal foveal cross-sectional image from each patient was retained[33], excluding 77 864 images. Following preprocessing, a total of 34 924 images were included in subsequent analyses.

2.3 Model Training

The Swin Transformer model[34] was trained independently on each dataset to compute predicted class probabilities and extract feature embeddings, which serve as the primary input to the confident learning algorithm[21]. Parameters followed Cleanlab defaults: batch size 32, learning rate 0.0001, 10 epochs, 3 folds, and patience 2. Model optimization was performed using the Adam optimizer with cross-entropy loss. All code was implemented in Python (version 3.9) using the PyCharm Community Edition (version 2024.3).

2.4 Automated Dataset Screening Using Cleanlab's Datalab Module

Following model training, predicted class probabilities and feature embeddings were imported into Cleanlab's Datalab module. Three quality checks were performed for each dataset[21].

The near-duplicate detection module measures pairwise similarity in feature space and identifies image pairs or groups with exceptionally small feature-space distances. Images exceeding dataset-specific thresholds were flagged as near-duplicates.

The Label Issue module estimates the confident joint, a matrix representing the joint distribution between observed noisy labels and latent true labels, and identifies images for which the predicted probability of an alternative class exceeds the per-class threshold. Each flagged image is accompanied by a suggested corrected label with an associated confidence score.

The Outlier module calculates an outlier score based on the distribution of feature embeddings within each assigned class. Images with scores exceeding a dataset-specific threshold are flagged as potential outliers. Unlike the Label Issue module, the Outlier module does not propose alternative class labels.

2.5 Manual Verification Workflow

Near-duplicate image groups flagged by Cleanlab were independently verified using perceptual hashing (pHash) implemented through the ImageHash library. Groups with hash distances below a predefined threshold (Hamming threshold ≤ 5) were classified as true duplicates. Only one representative image from each confirmed group was retained.

All images flagged by the Label Issue and Outlier modules underwent a four-stage expert review process. In the first stage, two ophthalmologists (Reviewer A and Reviewer B) independently assessed each flagged image. Reviewers were provided with the original dataset label and, for images identified by the Label Issue module, Cleanlab's suggested corrected label. They were blinded to the corresponding confidence score and to each other's assessments. Reviewers could retain the original label, accept the Cleanlab-suggested label, or assign an alternative diagnosis based on their clinical judgment. In the second stage, cases with discordant assessments between Reviewers A and B were referred to a third senior ophthalmologist (Reviewer C), who independently evaluated each case while remaining blinded to the prior reviews. In the third stage, the final consensus label was determined by majority opinion among the three reviewers. For cases in which all three reviewers assigned different labels, the assessment of Reviewer C was adopted as the final decision. In the fourth stage, a senior consultant ophthalmologist reviewed all cases assigned a revised label and confirmed the final classification.

Following expert review, revised labels were categorized into three groups: (1) in-distribution corrections, in which an image was reassigned to another predefined dataset class; (2) multi-label cases, in which more than one retinal condition was identified; (3) out-of-distribution (OOD) single-label cases, including diseases not represented in the original dataset taxonomy, and poor-quality or ungradable images, for which a reliable diagnosis could not be established. (Fig. 1)

2.6 Performance Metrics and Statistical Analysis

The positive predictive value (PPV) of label error was defined as the proportion of Label Issue module and Outlier module-flagged images confirmed as true label errors by expert review. Cleanlab label correction agreement rate was defined as the proportion of confirmed label errors for which Cleanlab's suggested label matched the expert-verified final label. As the Outlier module does not generate predicted correction labels, it was treated as supplementary to the Label Issue module for label error PPV analysis. Cohen's kappa (κ) was calculated for Stage 1 independent assessments, with 95% confidence intervals (CI) derived by bootstrap resampling over 1 000 iterations. Cleanlab confidence scores were compared between confirmed label errors and false-positive flags using the Mann-Whitney U test. Fisher's exact test was used to compare OOD and rare disease detection rates between modules. All tests were two-sided, with $P < 0.05$ considered statistically significant. All statistical analyses were performed in Python (version 3.9) using the PyCharm Community Edition (version 2024.3).

3. Results

Cleanlab's Datalab module flagged 1 424 (4.08%) images belonging to 689 near-duplicate groups, with group sizes ranging from 2 to 78 images. Verification using pHash confirmed 638 images from 635 groups as true duplicates and subsequently removed. Some examples of duplicates are shown in Supplementary Fig. 1. Most confirmed duplicates originated from TsukazakiUWF (n = 488) and OCTDL (n = 103). No duplicates were identified in MuReD, OCTID, or OpenUWF. Dataset-specific counts are summarized in Table 1.

Table 1
Summary of Cleanlab-identified data quality issues across nine retinal

Modality	Dataset	Total images	Duplicates			Label error detection and verification					
			Cleanlab		Verified	Cleanlab detection (% of total images)				Manual verification detected)	
Image			Group	Duplicate removed	Label Issue	Outlier	Total		Label Issue	Outlier	Total
CFP	ODIR	5684	4	2	2	1301 (22.89%)	18 (0.32%)	1310 (23.05%)		512 (39.35%)	7 (38.89
	MuReD	2060	/	/	/	479 (23.25%)	12 (0.58%)	484 (23.50%)		231 (48.23%)	2 (16.67
	EDC	4217	19	9	3	133 (3.15%)	82 (1.94%)	208 (4.93%)		68 (51.13%)	25 (30.49
OCT	OCT2017	5620	14	7	7	355 (6.32%)	30 (0.53%)	379 (6.74%)		133 (37.46%)	11 (36.67
	OCTID	572	/	/	/	7 (1.22%)	/	7 (1.22%)		4 (57.14%)	/
	OCTDL	2064	205	102	103	97 (4.70%)	16 (0.78%)	107 (5.18%)		33 (34.02%)	4 (25.00
UWF	TumorUWF	2031	176	68	35	44 (2.17%)	102 (5.02%)	133 (6.55%)		1 (2.27%)	1 (0.98%
	OpenUWF	700	/	/	/	103 (14.71%)	35 (5.00%)	131 (18.71%)		41 (39.81%)	5 (14.29
	TsukazakiUWF	11976	1006	501	488	1248 (10.42%)	599 (5.00%)	1675 (13.99%)		714 (57.21%)	189 (31.55
Total	/	**34924**	**1424**	**689**	**638**	**3767 (10.79%)**	**894 (2.56%)**	**4434 (12.70%)**		**1737 (46.11%)**	**244 (27.29**

Note: Cleanlab label correction agreement rate was defined as the proportion of confirmed label errors for which Cleanlab's suggested label matched the ex tomography; UWF, ultra-widefield fundus imaging; OOD, out-of-distribution.

Across all nine datasets, the Label Issue module flagged 3 767 (10.79%) images and the Outlier module flagged 894 (2.56%) images as suspected label errors. After accounting for overlap, 4 434 suspected label errors (12.70% of 34 924 images) were submitted for expert review.

The processes of manual verification are shown in Supplementary Table 2. In Stage 1, inter-rater agreement was moderate for both modules (Label Issue: $\kappa = 0.34$, 95% CI 0.31–0.36; Outlier: $\kappa = 0.46$, 95% CI 0.40–0.52).

Following expert review, 1 737 images flagged by the Label Issue module and 244 images flagged by the Outlier module were confirmed as true label errors, corresponding to PPVs of 46.11% and 27.29%, respectively. Accounting for overlap between the two modules, 1 884 unique images were ultimately verified as mislabeled, representing 42.49% of flagged images and 5.39% of all images. Dataset-level PPVs ranged from 1.50% in TumorUWF to 57.14% in OCTID (Table 1). Notably, 147 of the 244 label errors identified by the Outlier module (60.25%) were not detected by the Label Issue module, demonstrating the complementary value of outlier-based screening.

For confidence-score analysis, statistically significant differences between confirmed label errors and false-positive detections were observed in three of the nine datasets for the Label Issue module and in one dataset for the Outlier module. However, the discriminatory performance of confidence scores was limited, with area under the receiver operating characteristic curve (AUC) values ranging from 0.54 to 0.67 across all significant findings (Supplementary Table 3). In the TsukazakiUWF dataset, the Label Issue module demonstrated a significant but inverse association, whereby false-positive detections received higher confidence scores than confirmed label errors.

Among the 1 884 confirmed label errors, 1 435 images were corrected to another predefined dataset category and were classified as in-distribution corrections (Fig. 2a). An additional 103 images were reclassified as multi-label cases (Fig. 2b), while 400 images were identified as OOD cases (Fig. 2c). Because 54 images belonged to both the multi-label and OOD categories, a total of 449 unique cases (23.83% of all confirmed label errors) could not be reconciled within the original single-label classification framework. Among the OOD cases, 32 images were deemed ungradable because of poor image quality, while 13 images were diagnosed as rare retinal diseases (Fig. 3).

The Outlier module detected OOD images at a significantly higher rate than the Label Issue module among their respective flagged samples (15.88% vs 7.70%; OR = 2.26, 95% CI 1.82–2.81, $p < 0.001$). A similar pattern was observed for rare disease detection (0.78% vs 0.16%; OR = 4.95, 95% CI 1.66–14.76, $p = 0.005$).

In addition to identifying suspected label errors, the Label Issue module provided a suggested corrected label for each flagged image. Among images confirmed as mislabeled by expert review, agreement between Cleanlab's suggested label and the final expert-verified label ranged from 51.95% to 100.00%

across datasets, exceeding 60.00% in seven of the nine datasets (Table 1). No significant association was observed between dataset-level label error PPV and Cleanlab label agreement rate (Spearman's r = − 0.60, p = 0.097).

4. Discussion

This study evaluated a semi-automated label quality control workflow across nine public retinal image datasets spanning CFP, OCT, and UWF modalities. Cleanlab flagged 4 434 of 34 924 images (12.70%) for manual review, substantially reducing the need for exhaustive manual re-annotation. Following structured manual review and adjudication, 1,884 images (5.39% of the entire dataset cohort) were confirmed to contain label errors. The main findings were that Cleanlab effectively reduced manual review burden, provided clinically useful label corrections, and identified OOD samples, although its confidence scores demonstrated limited individual-level discriminative performance. Collectively, these findings suggest that label noise is prevalent in public retinal datasets and that Cleanlab can support dataset curation and quality assurance.

Near-duplicate detection identified 638 confirmed duplicate images across six datasets, with most duplicates originating from TsukazakiUWF and OCTDL. This uneven distribution likely reflects dataset construction and modality-specific acquisition characteristics. In UWF imaging, multiple overlapping or sequential images may be acquired during a single examination, increasing the likelihood of near-identical images being retained within a dataset. Similarly, OCT datasets frequently contain adjacent or repeated images with highly similar features. Duplicate and near-duplicate images are important sources of bias because they can lead to data leakage when distributed across training and testing partitions. A previous study has shown that improper OCT dataset splitting can inflate model accuracy[35]. Our findings therefore highlight the importance of duplicate screening in dataset quality control.

The PPV of Cleanlab-flagged label errors varied substantially across datasets, ranging from 1.50% to 57.14%. This variability likely reflects differences in label taxonomy, class separability, annotation quality, class imbalance, and the underlying prevalence of label errors. Importantly, the combined use of the Label Issue and Outlier modules improved overall error coverage. Among 244 label errors confirmed from the Outlier module, 147 (60.25%) were not captured by the Label Issue module, indicating that the two modules provide complementary information. Whereas the Label Issue module primarily identifies samples whose labels are inconsistent with model predictions, the Outlier module can identify atypical images outside the expected distribution of a given class.

The moderate inter-rater agreement observed during the initial review stage (Label Issue: $\kappa = 0.34$; Outlier: $\kappa = 0.46$) should be interpreted in the context of multiclass retinal disease classification, borderline diagnostic cases, and the inclusion of OOD categories. These findings support structured adjudication rather than single-reviewer relabeling when constructing reference standards for dataset correction.

Although Cleanlab confidence scores differed significantly between confirmed label errors and false-positive detections in several datasets, their overall discriminative performance was limited. AUC values ranged from only 0.54 to 0.67 among statistically significant findings and did not approach levels typically considered useful for individual-level decision-making. Therefore, confidence scores should not be used alone to determine whether an image is mislabeled. They are better used to prioritize manual review.

Beyond error detection, Cleanlab also generated suggested alternative labels for images flagged by the Label Issue module. Agreement between Cleanlab-generated labels and the final expert-verified labels ranged from 51.95% to 100.00% across datasets and exceeded 60% in seven of the nine datasets. Although the 100.00% agreement observed in TumorUWF was based on a single confirmed error and should be interpreted cautiously, the overall results suggest that Cleanlab-generated corrections can provide meaningful guidance during expert review. Notably, no significant association was observed between dataset-level PPV and label agreement rate (Spearman's r = − 0.60, p = 0.097). A dataset may contain a relatively low proportion of true errors among flagged samples while still allowing accurate correction of those errors once identified.

The present findings reinforce the value of data-level correction in ophthalmic artificial intelligence. While algorithm-level strategies can improve robustness to noisy labels during model training, they do not address errors embedded within the underlying datasets[19, 36, 37]. In contrast, data-level correction improves dataset quality at its source and can benefit subsequent analyses. Compared with previous studies that primarily evaluated synthetic label noise or focused on a single imaging modality[18], this work assessed label quality across nine real-world retinal image datasets spanning multiple modalities and disease categories. Furthermore, it integrated duplicate detection, label error identification, and outlier detection within a unified workflow and employed a structured expert adjudication process to establish clinically meaningful reference labels.

Several limitations should be acknowledged. First, expert review of flagged samples remained labor-intensive and may limit scalability in very large datasets. Second, Cleanlab performance depends on the quality of model-generated probability estimates, which may vary across datasets and influence error detection accuracy. Third, substantial heterogeneity among datasets limited direct comparisons and likely contributed to variability in performance metrics. Fourth, the single-label framework could not fully accommodate OOD and multi-label samples, leaving a subset of clinically valid cases difficult to represent. Finally, this study did not evaluate the downstream impact of dataset cleaning on model performance, calibration, or generalizability.

Future research should explore multi-label and open-set frameworks for real-world ophthalmic diagnoses. Semi-automated review strategies, active learning approaches, and human-in-the-loop annotation systems may further reduce expert workload while maintaining annotation quality. In addition, future work should evaluate effects on model training, calibration, fairness, and external validation.

In conclusion, Cleanlab substantially reduced the manual review burden by prioritizing a subset of images for expert assessment, while successfully identifying label errors, duplicate images, and atypical samples. These findings highlight the importance of data-centric quality control in ophthalmic artificial intelligence and support the use of Cleanlab as an effective screening tool for dataset curation.

Declarations

Funding

This research was supported by the Key Research Platforms and Projects of Guangdong Provincial Universities (Grant No. 2022LSYS006, 2023KCXTD013 and 2024ZDZX2024).

Author Contribution Statement

Yongleng Fang: conceptualization, methodology, data collection and analysis, and writing the original draft. Tian Lin: methodology, code correction, and experimental deployment. Tingkun Shi, Qi Zhang, Zhuangling Lin, Xixuan Ke, Guiqin He and Chaoxin Zheng: manual verification of labels. Haoyu Chen: conceptualization, supervision, clinical assessment and curation, expert verification, reviewing, and editing.

Data availability

All original datasets used in this study are publicly available from their respective sources, as cited in the manuscript. The cleaned label files and data-curation records generated in this study are available from the corresponding author upon reasonable request.

Conflict of Interest

The authors declare no competing interests.

References

1. Blindness, G.B.D., C. Vision Impairment, and S. Vision Loss Expert Group of the Global Burden of Disease. *Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study*. Lancet Glob Health. 2021; 9(2): p. e130–e143.
2. Blindness, G.B.D., C. Vision Impairment, and S. Vision Loss Expert Group of the Global Burden of Disease. *Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study*. Lancet Glob Health. 2021; 9(2): p. e144–e160.
3. Cen, L.P., J. Ji, J.W. Lin, S.T. Ju, H.J. Lin, T.P. Li, et al. *Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks*. Nat Commun. 2021; 12(1): p. 4828.
4. Liang, H., Q. Zhang, T. Lin, C. Hu, C. Zheng, X. Yao, et al. *Generation of multidisease fundus photographs with code-free platform*. Br J Ophthalmol. 2025; 109(12): p. 1382–1387.
5. De Fauw, J., J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, et al. *Clinically applicable deep learning for diagnosis and referral in retinal disease*. Nat Med. 2018; 24(9): p. 1342–1350.
6. Peng, Y., A. Lin, M. Wang, T. Lin, L. Liu, J. Wu, et al. *Enhancing AI reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis*. Cell Rep Med. 2025; 6(1): p. 101876.
7. Nagiel, A., R.A. Lalane, S.R. Sadda, and S.D. Schwartz. *ULTRA-WIDEFIELD FUNDUS IMAGING: A Review of Clinical Applications and Future Trends*. Retina. 2016; 36(4): p. 660–78.
8. Huang, B., C. Zheng, S. Chen, X. Liao, and H. Chen. *Quantifying retinal size and shape distortion in different ultra-widefield imaging systems*. BMJ Open Ophthalmol. 2025; 10(1).
9. Oganov, A.C., I. Seddon, S. Jabbehdari, O.E. Uner, H. Fonoudi, G. Yazdanpanah, et al. *Artificial intelligence in retinal image analysis: Development, advances, and challenges*. Surv Ophthalmol. 2023; 68(5): p. 905–919.
10. Cao, D., J. Yao, D.S.W. Ting, and G.S.W. Tan. *Artificial intelligence in predicting anti-VEGF treatment response in diabetic macular edema: current progress and future directions*. Visual Neuroscience. 2025; 42(1): p. e027.
11. Ting, D.S.W., C.Y. Cheung, G. Lim, G.S.W. Tan, N.D. Quang, A. Gan, et al. *Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes*. JAMA. 2017; 318(22): p. 2211–2223.
12. Kermany, D.S., M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, et al. *Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.* Cell. 2018; 172(5): p. 1122–1131 e9.
13. Wang, M., T. Lin, A. Lin, K. Yu, Y. Peng, L. Wang, et al. *Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model*. Nat Commun. 2025; 16(1): p. 5528.
14. Shi, J., K. Zhang, C. Guo, Y. Yang, Y. Xu, and J. Wu. *A survey of label-noise deep learning for medical image analysis*. Med Image Anal. 2024; 95: p. 103166.
15. Bernhardt, M., D.C. Castro, R. Tanno, A. Schwaighofer, K.C. Tezcan, M. Monteiro, et al. *Active label cleaning for improved dataset quality under resource constraints*. Nat Commun. 2022; 13(1): p. 1161.
16. Frenay, B., and M. Verleysen. *Classification in the presence of label noise: a survey*. IEEE Trans Neural Netw Learn Syst. 2014; 25(5): p. 845–69.
17. Wang, Y., X. Han, C. Li, L. Luo, Q. Yin, J. Zhang, et al. *Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study*. J Med Internet Res. 2024; 26: p. e52506.
18. Lin, T., M. Wang, A. Lin, X. Mai, H. Liang, Y.C. Tham, et al. *Efficiency and safety of automated label cleaning on multimodal retinal images*. NPJ Digit Med. 2025; 8(1): p. 10.
19. Northcutt, C.G., A. Athalye, and J. Mueller. *Pervasive label errors in test sets destabilize machine learning benchmarks*. in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021.

20. Galanty, M., D. Luitse, S.H. Noteboom, P. Croon, A.P. Vlaar, T. Poell, et al. *Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the BEAMRAD tool*. Sci Rep. 2024; 14(1): p. 31846.
21. Northcutt, C., L. Jiang, and I. Chuang. *Confident Learning: Estimating Uncertainty in Dataset Labels*. Journal of Artificial Intelligence Research. 2021; 70: p. 1373–1411.
22. Ju, L., X. Wang, L. Wang, D. Mahapatra, X. Zhao, Q. Zhou, et al. *Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation*. IEEE Trans Med Imaging. 2022; 41(6): p. 1533–1546.
23. Wei, Y., Y. Deng, C. Sun, M. Lin, H. Jiang, and Y. Peng. *Deep learning with noisy labels in medical prediction problems: a scoping review*. J Am Med Inform Assoc. 2024; 31(7): p. 1596–1607.
24. Li, N., T. Li, C. Hu, K. Wang, and H. Kang, *A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection*, in *Benchmarking, Measuring, and Optimizing*. 2021, Springer Nature Switzerland AG: Cham. p. 177–193.
25. Rodriguez, M.A., H. Almarzouqi, and P. Liatsis. *Multi-Label Retinal Disease Classification Using Transformers*. IEEE Journal of Biomedical and Health Informatics. 2023; 27(6): p. 2739–2750.
26. Doddi, G. *Eye Diseases Classification.* 2022. https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification/.
27. Gholami, P., P. Roy, M.K. Parthasarathy, and V. Lakshminarayanan. *OCTID: Optical coherence tomography image database*. Computers & Electrical Engineering. 2020; 81: p. 106532.
28. Kulyabin, M., A. Zhdanov, A. Nikiforova, A. Stepichev, A. Kuznetsova, M. Ronkin, et al. *OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods*. Scientific Data. 2024; 11(1): p. 365.
29. Sun, J., X. Zhao, S. Chen, Y. Zhang, H. Ren, Y. Sun, et al. *An ultra-wide-field fundus image dataset for intelligent diagnosis of intraocular tumors*. Scientific Data. 2025; 12(1): p. 1521.
30. He, S., X. Ye, W. Xie, Y. Shen, S. Yang, X. Zhong, et al. *Open ultrawidefield fundus image dataset with disease diagnosis and clinical image quality assessment*. Scientific Data. 2024; 11(1): p. 1251.
31. DateCazuki. *Fundus_Diagnosis: Classifier using fundus image data set provided by Tsukazaki Hospital.* 2024. https://github.com/DateCazuki/Fundus_Diagnosis/.
32. Engelmann, J., A.D. McTrusty, I.J.C. MacCormick, E. Pead, A. Storkey, and M.O. Bernabeu. *Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning*. Nature Machine Intelligence. 2022; 4(12): p. 1143–1154.
33. Luo, Y., T. Lin, A. Lin, X. Mai, and H. Chen. *Self-supervised based clustering for retinal optical coherence tomography images*. Eye (Lond). 2025; 39(2): p. 331–336.
34. Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
35. Tampu, I.E., A. Eklund, and N. Haj-Hosseini. *Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images*. Sci Data. 2022; 9(1): p. 580.
36. Karimi, D., H. Dou, S.K. Warfield, and A. Gholipour. *Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis*. Med Image Anal. 2020; 65: p. 101759.
37. Khanal, B., T. Dai, B. Bhattarai, and C. Linte, *Active Label Refinement for Robust Training of Imbalanced Medical Image Classification Tasks in the Presence of High Label Noise*, in *Lecture Notes in Computer Science*. 2024. p. 37–47.

Figures

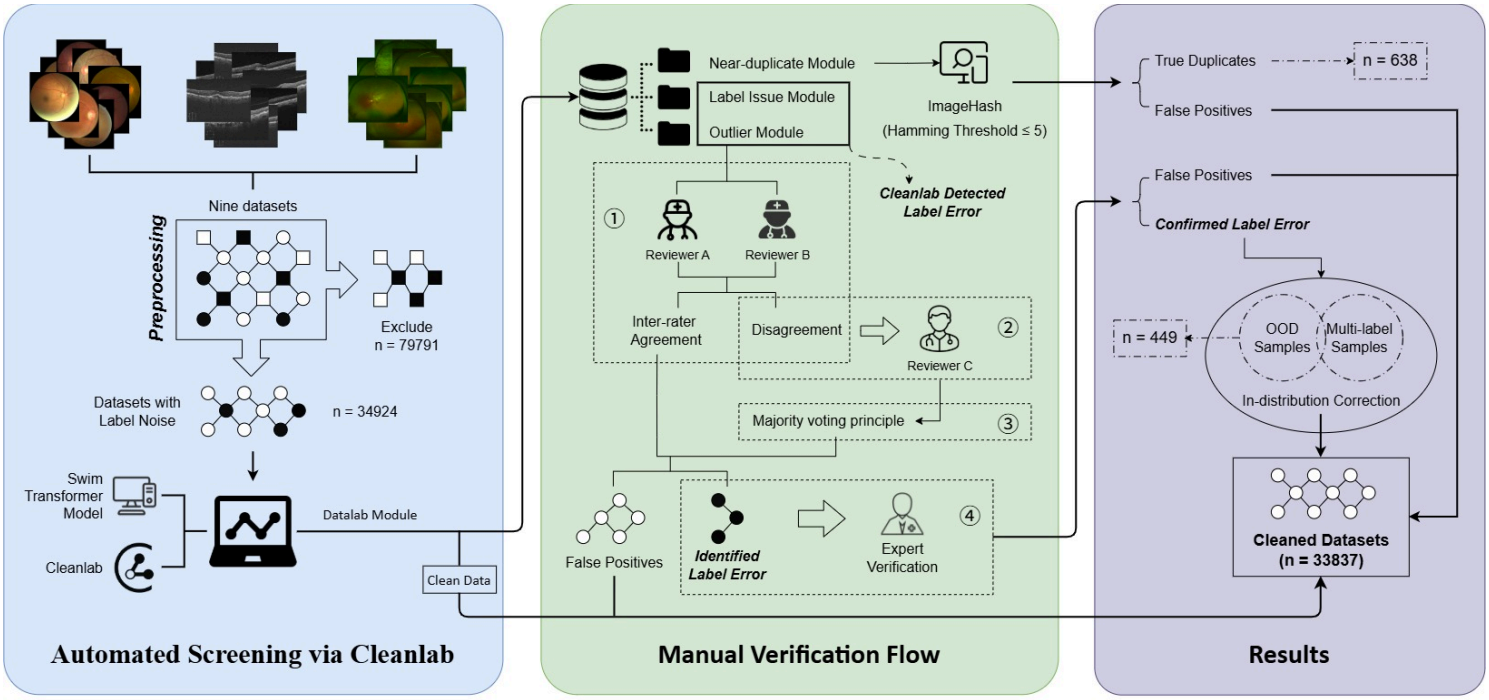


Figure 1

The semi-automated flow including Cleanlab framework and manual review. Note: Prior to analysis, images with undefined or inconsistently categorised labels were excluded from ODIR, MuReD, and TsukazakiUWF, and OCT2017 was deduplicated to one macular-region image per patient, yielding 34,814 images for input to the Cleanlab framework. The white round dots represent clean labels, while black round dots represent label errors. Abbreviations: OOD, out-of-distribution.

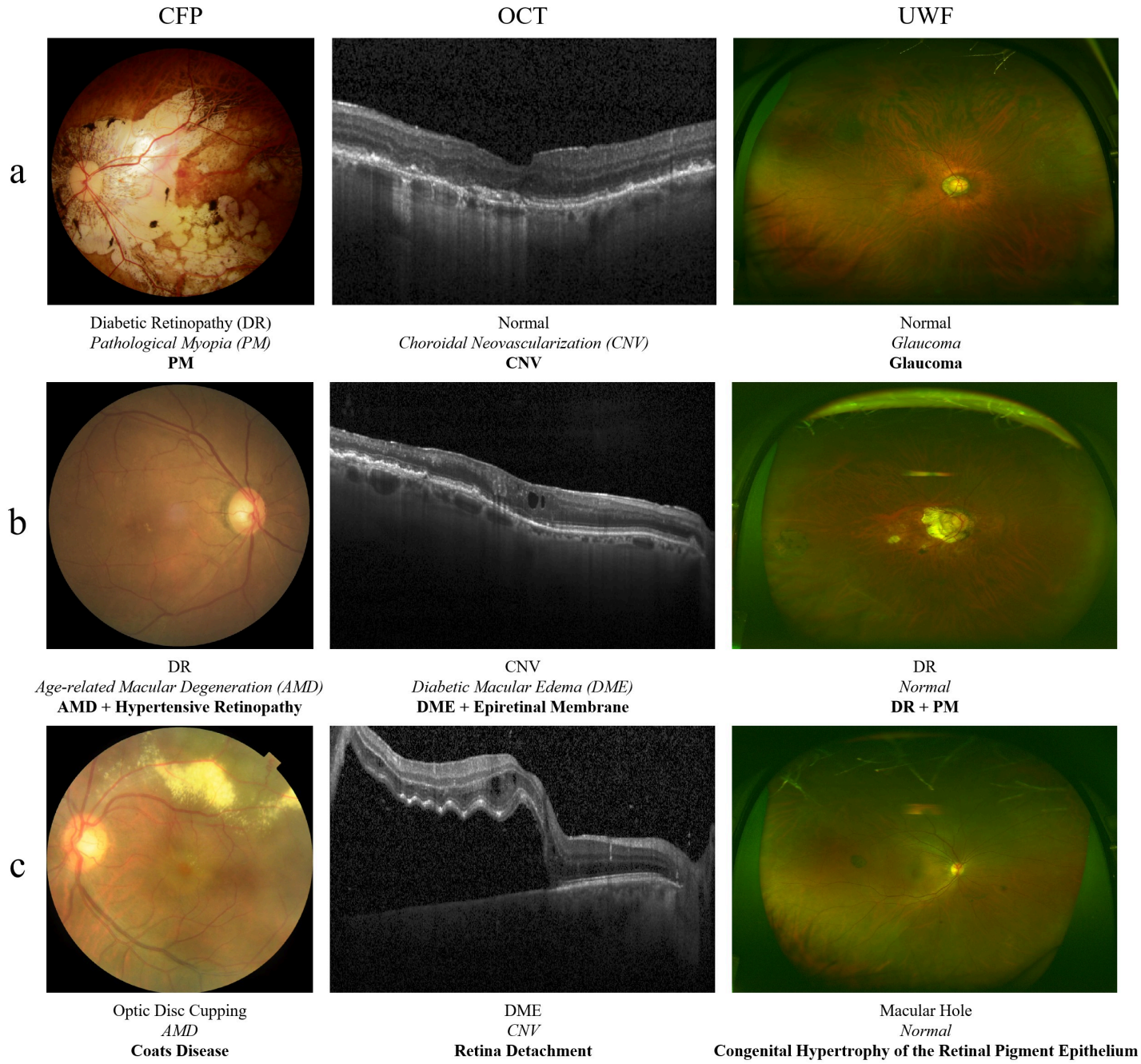


Figure 2

Representative images of label errors for three modalities.(a) in-distribution label errors, (b) multi-label errors, (c) out-of-distribution (OOD) label errors. Below each image are its disease labels. The first to third rows contain the original label, *the Cleanlab's suggested label*, and **the final label after manual verification**. Abbreviations: CFP, color fundus photography; OCT, optical coherence tomography; UWF, ultra-widefield fundus imaging. Note: Sample images from the ODIR dataset (Li et al., 2021), MuReD dataset (Rodriguez et al., 2023), OCT2017 dataset (Kermany et al., 2018), and TsukazakiUWF dataset (GitHub repository: https://github.com/DateCazuki/Fundus_Diagnosis), used under CC BY 4.0.

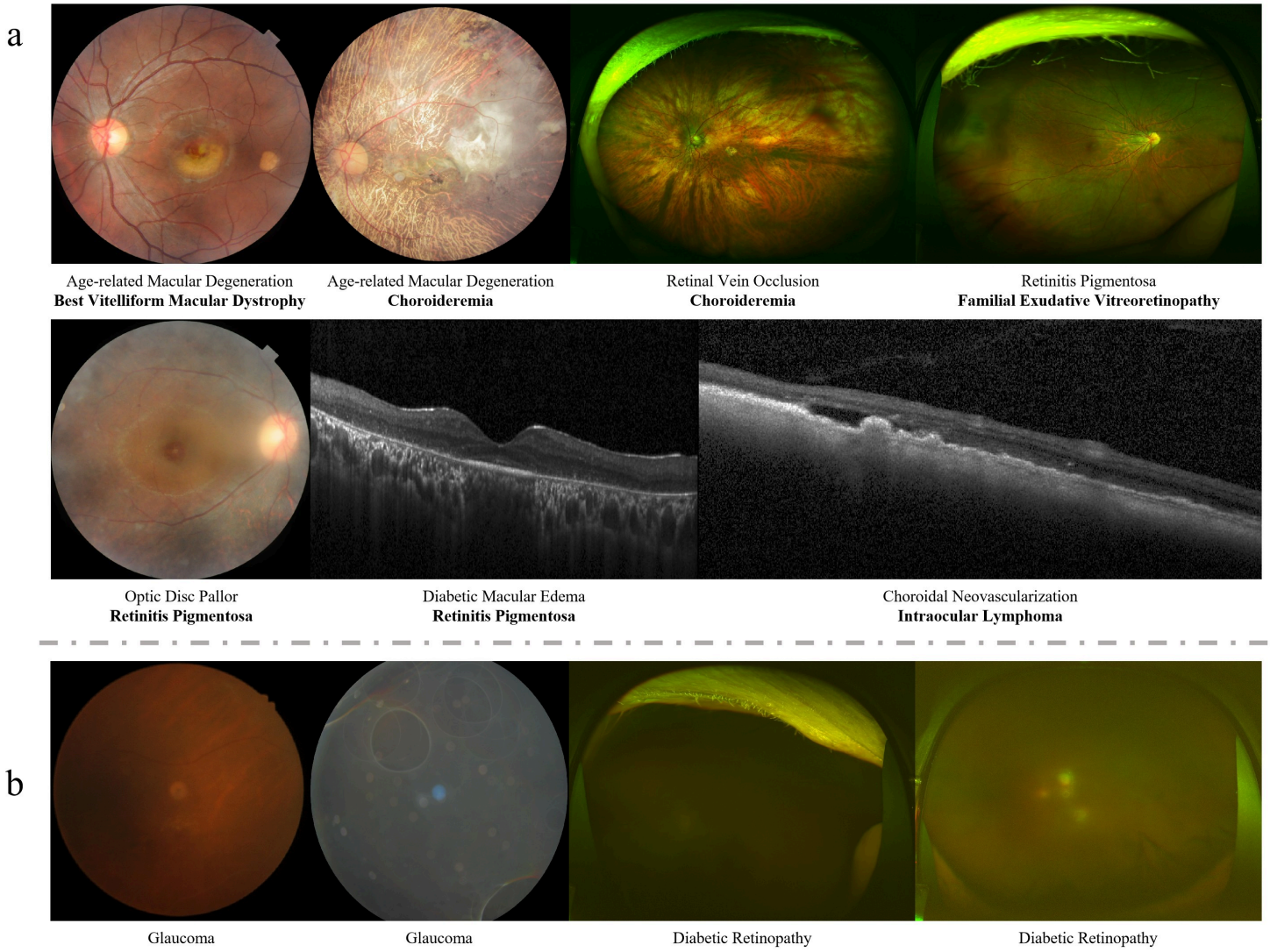


Figure 3

Representative images of rare ocular diseases and poor quality pictures. (a) rare ocular diseases, (b) poor quality pictures. Below each image are its disease labels (the original label and **the final label after manual verification**). Note: Sample images from the ODIR dataset (Li et al., 2021), MuReD dataset (Rodriguez et al., 2023), Eye Diseases Classification dataset (Kaggle dataset: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification), OCT2017 dataset (Kermany et al., 2018), and TsukazakiUWF dataset (GitHub repository: https://github.com/DateCazuki/Fundus_Diagnosis), used under CC BY 4.0.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.Summaryofincludeddatasets.pdf
- SupplementaryTable3.Cleanlabscoreassociationwithconfirmedlabelerrorsacrossninedatasets.pdf
- SupplementaryTable4.Summaryofduplicatesandlabelerrors.pdf
- Table1.SummaryofCleanlabidentifieddataqualityissuesacrossnineretinaldatasets.docx
- SupplementaryTable2.Informationofthemanualverificationworkflow.pdf
- SupplementaryFigure1.jpg