Vol. 39 (2026): Proceedings of 2026 International Conference on Life Sciences and Health Engineering
—— Session 3: Health for All Populations: Forum on Public Health and Clinical Health Sciences
(PHCHS-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/PHCHS-LSHE-2026.2026.39.8389

Conflict-Aware Exact-Byte Auditing and Grouped Splitting for Diabetic Retinopathy Grading

Ho Yeung Zhan*

Minxin Hong Kong School (Guangzhou Nansha), Guangzhou, China

**Corresponding author: Ho Yeung Zhan.*

Abstract

Byte-identical fundus files that cross data partitions or carry conflicting labels can distort evaluations of diabetic retinopathy grading. In the four-grade subset of Bangladesh Multi-Disease Retinal Dataset version 1, we audited file identity and compared conflict exclusion, grouped splitting, and duplicate canonicalization. Among 896 records, SHA-256 grouping followed by direct bytewise confirmation found 117 duplicate pairs, including 45 label-conflict pairs; exclusion retained 806 records in 734 groups. A decoded-RGB sensitivity audit yielded no further mergers. Across six controlled protocols, conflict-aware grouped splitting and its unique-record counterpart eliminated the defined overlap and conflict violations, retaining 806 and 734 records, respectively. A paired training-only multiplicity comparison then used shared unique validation and test cohorts across five grouped splits. Mean quadratic weighted kappa reached 0.645 with duplicate-retaining training and 0.601 with unique-record training; the paired mean difference was 0.044 and remained positive in all five splits. In a controlled leakage injection, training size, class counts, row order, validation, and test data were fixed. Mean kappa changed from 0.557 without exact train–test overlap to 0.625 at 74.5% overlap, although the intermediate means were non-monotonic. Claims are confined to exact-byte integrity within BMRD v1 and do not imply patient-level independence or clinical generalization.

Keywords

diabetic retinopathy grading, exact-byte audit, data leakage, label conflict, group-atomic splitting, data integrity

1. Introduction

Diabetic retinopathy (DR) is a common microvascular complication of diabetes and a major target for image-based screening. Deep-learning systems can detect or grade retinal disease from fundus photographs, but credible performance estimates depend on how the evaluation cohort is assembled. Large clinical studies have demonstrated the potential of automated retinal assessment, while work on limited-data grading has shown that evaluation design and loss construction remain consequential even when the model architecture is deliberately compact [1–4].

Data leakage is a recurring threat because copies or related images placed in different partitions expose a model to test-associated information during training and can make generalization appear better than it is. The problem spans modalities: reviews and audits have traced leakage to patient overlap, preprocessing, feature selection, and repeated images across medical imaging and machine learning [5–10]. Keeping known identities

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

within a single partition blocks one route of leakage, yet grouped splitting does not resolve inconsistent labels inside an identity group. Retaining only one representative per group has a different effect, changing both the number of training rows and the empirical weight of duplicated examples. The two interventions therefore call for separate comparisons.

As a narrow and reproducible starting point, exact-byte identity can be represented by cryptographic hashes of the source-file bytes, provided that candidate matches are also confirmed directly [11, 12]. When duplicates carry conflicting grades, they also bear on mislabeled examples, label noise, hidden stratification, and documentation of dataset provenance [13–16]. We study these issues in the public four-grade subset of Bangladesh Multi-Disease Retinal Dataset (BMRD) version 1, a multi-disease retinal-image resource containing repeated byte-identical files and source-label disagreements [17].

The study combines an exact-byte audit, confirmed by direct file comparison and accompanied by a decoded-RGB sensitivity check, with six controlled protocol variants that separate conflict exclusion, grouped partitioning, and duplicate canonicalization. Repeated grouped splits and a controlled leakage-injection experiment complete the evaluation, with the latter preserving cohort composition while changing exact train–test overlap.

The contribution is therefore methodological rather than architectural. First, the audit provides a complete record-to-byte account of the admitted BMRD v1 subset, including duplicate multiplicity and the grade pairs involved in source-label conflicts. Second, the six protocols treat conflict exclusion, group-atomic splitting, and canonicalization as separate decisions instead of collapsing them into a single “deduplication” label. Third, the predictive analysis distinguishes training-set multiplicity from evaluation-cohort duplication and then varies exact train–test overlap while holding training size and class counts fixed. This structure connects data retention, integrity violations, and apparent grading performance without presenting the compact classifier as a competitive clinical model.

2. Method

2.1 Dataset and Exact-byte Audit

The study used the four-grade DR subset of BMRD version 1 from the public archive and retained the archive’s labels 0–3 [17]. Before the audit, thirty-five images without an admitted grade in this range were excluded, leaving 896 records. The released records provided no verified patient identifier for this subset, so identity was defined at the source-file rather than patient level.

For every admitted record, we computed SHA-256 over the unmodified source-file bytes and grouped equal digests. We then compared the bytes of every candidate pair directly; only bytewise-equal files were treated as exact duplicates. A group was labelled source-conflicting when its members carried more than one DR grade. Otherwise, additional copies beyond the first were counted as duplicate excess. Canonical selection, where required, used the lexicographically first normalized relative path and was fixed before model training.

A separate sensitivity audit decoded each file with orientation correction and converted the result to RGB. Grouping by dimensions and decoded pixel bytes then tested whether distinct file containers represented the same decoded raster, without resizing or perceptual similarity thresholds.

Raw-byte identity was the primary unit because it is deterministic and can be reconstructed from archive files without image transformation. The decoded-RGB pass served only as a sensitivity check for distinct containers that decode to identical oriented pixels. Neither layer detects resized, cropped, recompressed, or perceptually similar images, so those cases remained outside the protocol definitions rather than being assigned uncertain matches.

2.2 Protocol Constructtion and Integrity Checks

Six protocols were derived from the same audit. The naive record split assigned all 896 rows independently, whereas hash grouping retained every row but placed each exact-byte group in one partition. After all members of the 45 conflicting groups were removed, conflict exclusion followed by record splitting assigned the remaining rows independently. Canonicalization with conflicts instead retained one selected record per hash

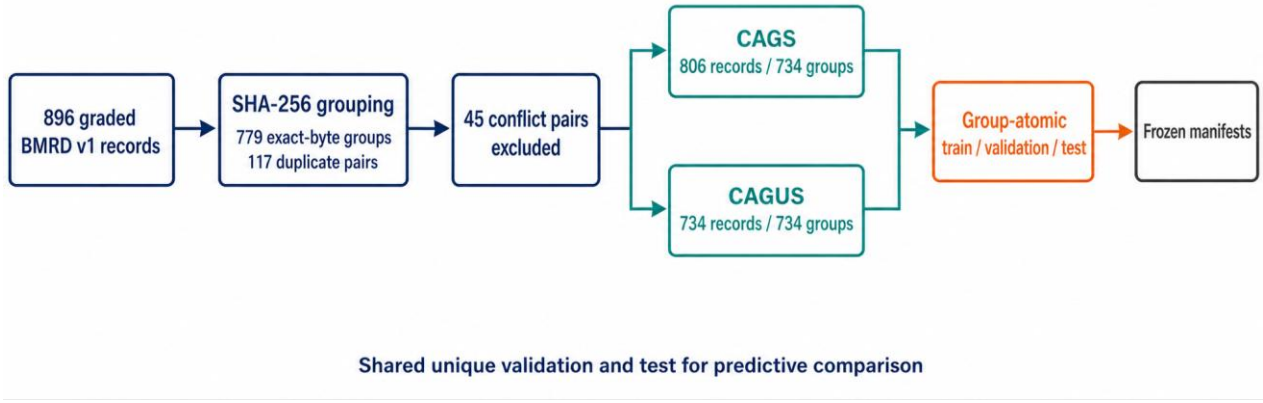
Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

while preserving conflicting group labels in the source audit. Conflict-Aware Grouped Split (CAGS) excluded every conflicting group, kept the remaining same-label copies, and assigned each hash group atomically. Using the same conflict exclusion and grouped assignment, Conflict-Aware Grouped Unique Split (CAGUS) retained one canonical record per non-conflicting hash.

The protocol set decomposes three decisions that affect different parts of an evaluation. Conflict treatment determines whether contradictory supervision remains available to a model. The partition unit determines whether the same known byte identity can occur in more than one data split. Canonicalization determines the empirical weight assigned to an identity after conflicts have been handled. These decisions are not interchangeable. Group-atomic splitting can remove cross-partition overlap while leaving both source-label conflicts and repeated training rows, whereas canonicalization can remove repeated rows while concealing which label was selected from a conflicting group. CAGS and CAGUS share conflict exclusion and group assignment, so their difference is specifically the retention of same-label copies. Keeping the three decisions explicit also prevents a lower violation count from being interpreted as a free improvement, because every protocol must be read together with its retained record and unique-group counts.

For each protocol, we counted exact-byte groups spanning train, validation, and test partitions, retained source-conflicting groups, and duplicate excess. The integrity rate was the proportion of retained records implicated either in a cross-partition exact-byte group or in a retained source-conflicting group; the zero values for CAGS and CAGUS follow from their construction. Figure 1 traces the audit into the CAGS and CAGUS branches and their shared group-atomic partitioning step. Every manifest and split seed was frozen before training.

Figure 1: Exact-byte audit and protocol construction. SHA-256 grouping identifies duplicate pairs, conflicting pairs are excluded, and the retained groups follow either the CAGS or CAGUS branch before group-atomic partitioning. The paired predictive comparison uses shared unique validation and test cohorts. Counts report records or exact-byte groups at the indicated stage



2.3 Predictive Evaluation

The paired predictive comparison was restricted to a training-only multiplicity intervention. For each of five grouped split seeds (20260804, 20260811, 20260818, 20260825, and 20260901), both branches used the same unique validation cohort and the same unique test cohort. The duplicate-retaining training branch kept all non-conflicting training rows, including same-label byte-identical copies; the unique-record training branch kept one canonical training row per exact-byte group. Thus, only training multiplicity and training-row count differed. Full CAGS, by contrast, retains non-conflicting rows in every partition and is reported only in the six-protocol integrity comparison.

For controlled measurement rather than a proposed clinical architecture, we used a compact convolutional network with inputs resized to 128 × 128 pixels. Class-weighted cross-entropy and AdamW were used for eight epochs, with checkpoint selection based on validation performance. Every grouped split was trained with model seeds 13, 37, and 71. The reported metrics were quadratic weighted kappa (QWK), macro-F1, balanced accuracy, and severe-error rate, where a severe error differed from the reference grade by at least two levels. Within each split, metrics were first averaged across the three model seeds; the five split-level means were then summarized as mean and standard deviation.

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

The controlled leakage experiment started from fixed unique-record training, validation, and test cohorts containing 514, 110, and 110 records. At four intended overlap levels, selected training rows were replaced by test records of the same grade while training size, class counts, row order, validation data, and test data remained unchanged. The realized exact train–test overlaps were 0.0%, 25.5%, 50.9%, and 74.5%. Five model seeds (13, 37, 71, 97, and 113) were evaluated at each level, with seed-level means and standard deviations reported descriptively.

3. Results

3.1 Exact-byte Audit and Protocol Integrity

Direct byte comparison confirmed that the 896 admitted records formed 779 raw-byte groups, comprising 662 singletons and 117 two-record duplicate groups. Of these duplicate pairs, seventy-two shared a grade, with 0, 69, 1, and 2 pairs corresponding to grades 0–3. The other 45 pairs were source-conflicting: 11 linked grades 0 and 1, 22 linked grades 1 and 2, five linked grades 1 and 3, and seven linked grades 2 and 3. Removing conflicts excluded 90 records and left 806 records in 734 exact-byte groups. The decoded-RGB audit recovered those same 779 groups and 45 conflicts, without any further mergers.

Table 1 compares the six frozen protocols. Independent record splitting allowed 55 exact-byte groups to cross partitions; removing conflicts before independent splitting reduced this count to 34 but did not eliminate it. Grouping prevented cross-partition hashes, whereas canonicalization removed duplicate excess. Only CAGS and CAGUS had neither cross-partition exact-byte groups nor retained source-label conflict groups, with integrity rates of 0%; they differed in whether the 72 same-label duplicate rows were retained. A single rate nevertheless masks the source of failure: the intermediate protocols have nonzero values for different reasons.

The integrity rate was therefore interpreted as an implementation check rather than a predictive endpoint. Its numerator records whether either defined violation remains, while the accompanying record, group, and duplicate-excess counts show what was retained to obtain that value. In particular, the shared zero rate of CAGS and CAGUS confirms that both construction rules were satisfied; it does not rank their training distributions. The paired predictive analysis below addresses that remaining distinction with common unique validation and test cohorts, so evaluation images receive equal weight in both branches.

Table 1: Integrity outcomes for the six controlled protocol variants. Duplicate excess counts retained records beyond the first member of each same-label exact-byte group

Protocol	Records /hashes	Any cross-split groups	Source-conflict groups retained	Duplicate excess	Integrity rate
Naive record split	896/779	55	45	117	18.5%
Hash grouping	896/779	0	45	117	10.0%
Conflict exclusion + record split	806/734	34	0	72	8.4%
Canonicalization, conflicts retained	779/779	0	45	0	5.8%
CAGS	806/734	0	0	72	0%
CAGUS	734/734	0	0	0	0%

3.2 Training Multiplicity and Controlled Leakage

Duplicate-retaining training added 44, 43, 53, 52, and 47 rows across the five grouped splits, respectively, relative to unique-record training. Ordered by grades 0–3, the grade-wise excess counts were (0, 42, 1, 1), (0, 42, 0, 1), (0, 52, 0, 1), (0, 50, 1, 1), and (0, 45, 0, 2). Almost every added row therefore carried grade 1, meaning that the intervention primarily up-weighted grade 1 rather than systematically oversampling the minority grades.

Block A of Table 2 summarizes the five split-level means after averaging the three model seeds within each split. Duplicate-retaining training reached QWK 0.645 ± 0.047, compared with 0.601 ± 0.036 for unique-record training. The paired mean difference, ΔQWK = 0.044, was positive in all five grouped splits. Macro-F1 and balanced accuracy were also higher for duplicate-retaining training, while severe-error rate was lower. These rows isolate training multiplicity under shared unique validation and test cohorts; they are not a direct predictive comparison of the full CAGS and CAGUS manifests.

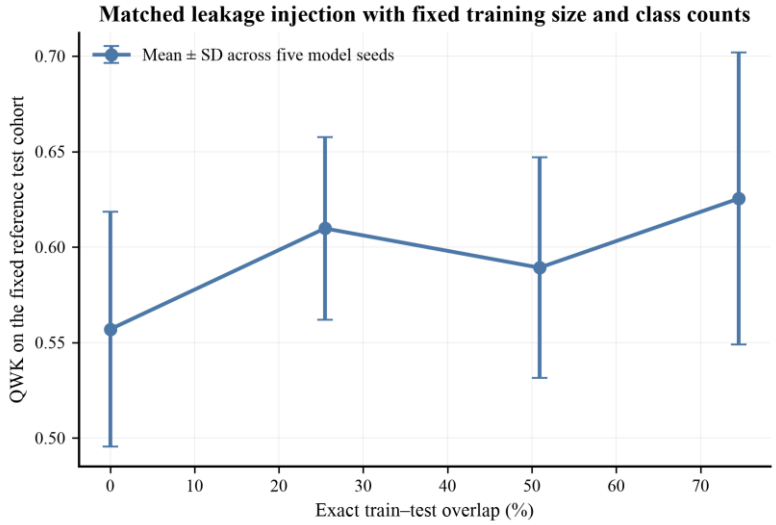
Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

Block B reports the matched leakage injection. Mean QWK was 0.557 at zero overlap and 0.625 at 74.5% overlap, with non-monotonic intermediate means. Macro-F1, balanced accuracy, and severe-error rate changed alongside QWK, but none varied monotonically across all four overlap levels. Figure 2 shows the mean QWK and standard deviation across the five model seeds while keeping the fixed validation and test cohorts explicit.

Table 2: Predictive results for the paired training-multiplicity comparison (Block A) and matched exact-overlap injection (Block B). Values are mean ± standard deviation over five split-level means in Block A and five model seeds in Block B

Condition	QWK	Macro-F1	Balanced accuracy	Severe error
Block A: paired training-only multiplicity				
Duplicate-retaining training	0.645 ± 0.047	0.582 ± 0.027	0.589 ± 0.029	0.039 ± 0.012
Unique-record training	0.601 ± 0.036	0.555 ± 0.020	0.569 ± 0.027	0.057 ± 0.010
Block B: matched exact train–test overlap				
0.0% overlap	0.557 ± 0.061	0.516 ± 0.052	0.541 ± 0.058	0.062 ± 0.012
25.5% overlap	0.610 ± 0.048	0.569 ± 0.049	0.590 ± 0.051	0.051 ± 0.024
50.9% overlap	0.589 ± 0.058	0.566 ± 0.079	0.571 ± 0.072	0.053 ± 0.008
74.5% overlap	0.625 ± 0.077	0.592 ± 0.076	0.600 ± 0.062	0.044 ± 0.016

Figure 2: Quadratic weighted kappa under matched exact-overlap injection. Markers show the mean and vertical bars show the standard deviation across five model seeds. Training size, class counts, row order, validation data, and test data are fixed; only same-grade replacement with exact test records changes the realized train–test overlap



4. Discussion

The audit shows why "deduplication" is an incomplete description of evaluation design. Conflict exclusion, group-atomic splitting, and canonicalization act on supervision, partition independence, and sample weighting, respectively. Reporting both retained records and exact-byte groups makes these choices more interpretable than a binary statement that a dataset was deduplicated.

The paired training experiment gives a narrower view of same-label multiplicity. Because the additional rows were concentrated in grade 1, the contrast primarily changes grade-1 weighting and total training-row count, not the amount of independent image information. Because the epoch count was fixed, the duplicate-retaining branch also underwent more minibatch updates. The observed difference therefore combines duplicate weighting, training-set size, and optimization-step count. Its positive QWK difference across all five splits is a descriptive property of this BMRD setup, not evidence that duplication is generally beneficial. The matched experiment addresses a different question. It shows that exact overlap can alter apparent performance, but the non-monotonic intermediate means do not support a simple dose–response interpretation.

For benchmark practice, the audit suggests a compact reporting bundle that can be applied before model fitting. Authors should retain a record-to-hash manifest, state whether equality refers to file bytes or decoded pixels, and report the number of identities crossing each partition boundary. Source-label conflict groups

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

should be counted before any canonical record is chosen, because single-record selection can hide the original disagreement. Retained records, unique identities, and duplicate excess should accompany every integrity indicator so that removal and reweighting remain visible. Predictive comparisons should use a common unique evaluation cohort when the question concerns training multiplicity, and split seeds should be reported separately from model seeds. These items do not require a new model, but they make the evaluation target and the source of any performance change easier to audit.

The evidence is confined to one four-grade subset. Without verified patient identifiers, exact-byte grouping cannot establish patient-level independence, while the decoded-RGB check misses resized, cropped, recompressed, or perceptually similar images. Precision is further limited by five grouped splits and small fixed test cohorts; the compact network was selected for controlled measurement, not state-of-the-art comparison. Accordingly, these results do not establish clinical validity, external generalization, deployment performance, model superiority, or universal causal effects of duplication.

5. Conclusion

A practical BMRD-based DR grading workflow begins with auditing source bytes before partitioning, removing unresolved label-conflict groups, assigning known identities atomically, and reporting whether same-label copies remain. Frozen manifests and explicit retention counts can then separate integrity control from changes in the training distribution. Testing whether the observed patterns persist beyond this archive will require patient-level identifiers, broader near-duplicate detection, and external cohorts.

References

[1] Gulshan, V., and others, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, 316, (22), 2402–2410, 2016, https://doi.org/10.1001/jama.2016.17216.

[2] Ting, D. S. W., and others, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, 318, (22), 2211–2223, 2017, https://doi.org/10.1001/jama.2017.18152.

[3] Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K., and Kaski, K., "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Scientific Reports*, 9, 10750, 2019, https://doi.org/10.1038/s41598-019-47181-w.

[4] Galdran, A., Dolz, J., Chakor, H., Lombaert, H., and Ayed, I. B., *Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images*, *Medical image computing and computer assisted intervention–MICCAI 2020*, 665–674, 2020, https://doi.org/10.1007/978-3-030-59722-1_64.

[5] Kapoor, S., and Narayanan, A., "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, 4, (9), 100804, 2023, https://doi.org/10.1016/j.patter.2023.100804.

[6] Davis, S. E., Matheny, M. E., Balu, S., and Sendak, M. P., "A framework for understanding label leakage in machine learning for health care," *Journal of the American Medical Informatics Association*, 31, (1), 274–280, 2024, https://doi.org/10.1093/jamia/ocad178.

[7] Varoquaux, G., and Cheplygina, V., "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digital Medicine*, 5, (1), 2022, https://doi.org/10.1038/s41746-022-00592-y.

[8] Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Garcia, D. V. V., Singh, Y., Zhang, K., Conte, G. M., and Erickson, B. J., "Mitigating bias in radiology machine learning: 1. Data handling," *Radiology: Artificial Intelligence*, 4, (5), e210290, 2022, https://doi.org/10.1148/ryai.210290.

[9] Yagis, E., Atnafu, S. W., Herrera, A. G. S. de, Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., and Diciotti, S., "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Scientific Reports*, 11, (1), 2021, https://doi.org/10.1038/s41598-021-01681-w.

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

[10] Tampu, I. E., Eklund, A., and Haj-Hosseini, N., "Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images," *Scientific Data*, 9, (1), 580, 2022, https://doi.org/10.1038/s41597-022-01618-6.

[11] National Institute of Standards and Technology, *Secure hash standard (SHS)*, National Institute of Standards; Technology, 2015, https://doi.org/10.6028/NIST.FIPS.180-4.

[12] Truong, T., Jush, F. K., and Lenga, M., *Benchmarking pretrained vision embeddings for near- and duplicate detection in medical images*, *2024 IEEE international symposium on biomedical imaging*, 1–5, 2024, https://doi.org/10.1109/ISBI56570.2024.10635550.

[13] Northcutt, C. G., Athalye, A., and Mueller, J., *Pervasive label errors in test sets destabilize machine learning benchmarks*, *Advances in neural information processing systems, datasets and benchmarks track*, 2021, https://openreview.net/forum?id=XccDXrDNLek.

[14] Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., and Wu, J., "A survey of label-noise deep learning for medical image analysis," *Medical Image Analysis*, 95, 103166, 2024, https://doi.org/10.1016/j.media.2024.103166.

[15] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C., *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging*, *Proceedings of the ACM conference on health, inference, and learning*, 151–159, 2020, https://doi.org/10.1145/3368555.3384468.

[16] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K., "Datasheets for datasets," *Communications of the ACM*, 64, (12), 86–92, 2021, https://doi.org/10.1145/3458723.

[17] Ema, R. R., *Bangladesh multi-disease retinal dataset (BMRD)*, Mendeley Data, 2025, https://doi.org/10.17632/959whsbc6v.1.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group