iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.84044

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Reliability Reporting in Kaggle-Based Diabetic Retinopathy AI Studies

Mudit Sharma
LK Singhania Education Centre, Gotan, India

Abstract: ***Public retinal fundus datasets have enabled many machine-learning studies on diabetic retinopathy detection, but high reported performance can be difficult to interpret when validation and reporting practices vary across studies. This structured secondary-data audit examined 30 source-traced study rows involving Kaggle-accessible diabetic-retinopathy datasets or closely related external-validation datasets. For each row, dataset use, model family, task framing, reported performance metrics, evidence-use tier, and nine reliability-reporting criteria were extracted into a reproducible workbook. The audit identified incomplete reporting of several reliability markers: external validation was clearly reported in 9 of 30 rows, confidence intervals in 2 of 30 rows, public code availability in 1 of 30 rows, and explicit leakage-control reporting in 12 of 30 rows. No row reached the high-reliability band under the predefined 0-9 rubric. Reported AUROC/AUC and accuracy values were summarized descriptively only because included rows used heterogeneous datasets, tasks, and validation procedures. The study identifies reporting gaps and supports cautious interpretation of high internal Kaggle performance claims. It does not train a new model, clinically validate a diagnostic system, or conduct a formal diagnostic-performance meta-analysis.***

Keywords: Diabetic retinopathy; fundus imaging; Kaggle; EyePACS; APTOS 2019; external validation; reproducibility; medical AI; secondary-data audit

I. INTRODUCTION

Diabetic retinopathy is a diabetes-related retinal disease that can lead to vision loss when it is not detected and managed in time [1]. Because early disease may be asymptomatic, screening programs and retinal imaging have become important tools for identifying patients who need follow-up care. Machine-learning systems for fundus-image analysis are therefore attractive because they may support screening workflows when specialist availability is limited.

Public datasets have strongly shaped student and academic work in this area. Kaggle-hosted datasets, especially APTOS 2019 Blindness Detection and Kaggle Diabetic Retinopathy Detection/EyePACS, are widely used because they provide labeled fundus images and allow reproducible experimentation without collecting private medical data [2,3]. However, easy access to public data also creates a methodological risk: high internal validation scores can be reported without enough information about external validation, confidence intervals, data leakage, class imbalance, or patient-level independence.

This issue matters because medical-image performance can change when a model is tested outside the development distribution. The JAMA diabetic-retinopathy benchmark by Gulshan et al. demonstrated that strong performance can be reported with external validation and confidence intervals [4]. In contrast, the Voets et al. replication showed that performance obtained on Kaggle EyePACS did not fully transfer to Messidor-2, reinforcing the need to distinguish internal dataset performance from evidence of generalization [5].

The present study does not propose another retinal CNN architecture and does not claim clinical diagnostic performance. Instead, it audits the reporting reliability of previously reported diabetic-retinopathy AI studies that use Kaggle-accessible fundus datasets or closely related external-validation datasets. The intended contribution is a transparent reliability-reporting framework that distinguishes headline performance claims from the strength of supporting validation evidence.

A. Research Questions

1) Among diabetic-retinopathy machine-learning studies using Kaggle-accessible fundus datasets, how frequently are external validation, confidence intervals, statistical testing, code availability, leakage-control reporting, class-imbalance handling, patient/image split description, calibration reporting, and clinical metrics reported?
2) How are reported AUROC/AUC and accuracy values distributed across reliability-score bands?
3) Can a reproducible reliability rubric help distinguish headline performance claims from better-supported validation evidence?

©IJRASET: All Rights are Reserved

B. Hypothesis

The working hypothesis was that studies relying primarily on internal Kaggle validation would often report strong headline performance, but would receive lower reliability scores than studies that included external validation, uncertainty intervals, code release, leakage-control reporting, and clinically meaningful metrics.

II. METHODS

A. Study design

A structured secondary-data audit was conducted. The unit of analysis was a source-traced study row in the extraction workbook. A row represented either a Kaggle-anchored diabetic-retinopathy machine-learning study, an adjacent study using Kaggle-accessible datasets with external validation, or a context benchmark retained for interpretation. This study was not a formal systematic review, not a PRISMA meta-analysis, and not a pooled diagnostic-performance meta-analysis. Its purpose was to audit reporting and validation practices rather than estimate a single pooled AUROC, accuracy, sensitivity, or specificity value.

B. Dataset scope

The primary dataset scope included APTOS 2019 Blindness Detection and Kaggle Diabetic Retinopathy Detection/EyePACS. Studies using Messidor, Messidor-2, IDRiD, DDR, or related retinal datasets were included when they were linked to Kaggle-style model development, external validation, or cross-dataset evaluation. Private-only studies were excluded from primary analysis unless they served as clinical benchmark context.

C. Search and screening procedure

Candidate studies were identified using web search, arXiv, JAMA, and Kaggle-related queries documented in the workbook screening log. Search strings included combinations of "APTOS 2019 diabetic retinopathy", "EyePACS Kaggle", "Messidor-2 external validation", "AUROC", "sensitivity", "specificity", "class imbalance", and "code availability". The search procedure was documented for reproducibility, but it should not be described as exhaustive PRISMA screening.

D. Source verification and metric handling

Metric values were treated as source-traced only when they could be located on an official article page, arXiv abstract page, full article, or other stable source page recorded in the workbook. Rows verified only from abstracts or source pages were labelled as such and were not described as full-PDF table audits. Full-article verification was strongest when the complete paper page or article text provided metric values, confidence intervals, or validation details.

The manuscript therefore avoids language such as "proven," "clinically validated," or "model superiority." Reported metrics are interpreted as previously reported values from heterogeneous studies, not as directly comparable measurements under a shared experimental protocol.

TABLE I
ELIGIBILITY LOGIC FOR THE SECONDARY-DATA AUDIT.

Included	Excluded
APTOS 2019, Kaggle EyePACS/Diabetic Retinopathy Detection, or Kaggle-linked studies using Messidor/Messidor-2 for validation	Non-retinal studies, non-DR studies, private-only studies without relevance as benchmarks
Papers reporting at least one performance metric or validation/reporting criterion	Papers with no extractable metric and no reliability-criteria information
Peer-reviewed papers, preprints, replication studies, and benchmark studies when source-level evidence was available	Duplicate reports, nontechnical blog posts, and papers where the source could not be verified

E. Extraction variables

For each eligible row, the extraction table recorded study title, year, authors, source type, dataset or datasets, Kaggle anchor, external-validation dataset, task, model family, reported accuracy, AUROC/AUC, AUPRC, sensitivity, specificity, F1-score, confidence intervals, code availability, leakage-control reporting, class-imbalance handling, calibration or uncertainty reporting, evidence-use tier, and source-verification notes.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

F. Reliability scoring rubric

A 0-9 reliability-reporting rubric was defined before final manuscript revision. One point was awarded for each clearly reported criterion. Partial reporting was coded as 0.5 when a criterion was mentioned but incompletely documented. The score should be interpreted as a reporting-reliability indicator, not as a definitive judgment of model quality, author competence, or clinical usefulness.

TABLE II
RELIABILITY SCORING RUBRIC.

Criterion	Full point definition	Why it matters
External validation	Independent external dataset used	Tests performance beyond one dataset distribution
Confidence intervals	95% CI or comparable uncertainty interval reported	Prevents overinterpretation of single-point metrics
Statistical testing	Formal statistical comparison or p-value reported	Assesses whether performance differences may reflect noise
Code availability	Public code or reproducible scripts confirmed	Supports independent verification
Leakage-control description	Clear split, preprocessing, augmentation, and tuning separation	Reduces inflated performance from test contamination
Class-imbalance handling	Weighting, sampling, focal loss, SMOTE, class-wise reporting, or comparable method	DR labels are often skewed
Patient/image split description	Patient-level, eye-level, or image-level split logic described	Reduces same-patient or same-image leakage risk
Calibration/uncertainty	Calibration, uncertainty, or reliability analysis reported	Medical probabilities should communicate uncertainty
Clinical metrics	Sensitivity, specificity, or comparable screening metrics reported	Accuracy alone is insufficient for screening

G. Evidence-use tiers

Rows were assigned to evidence-use tiers before writing the Results section. Primary quantitative rows were used in descriptive performance plots. Secondary quantitative rows were retained for descriptive support but were not treated as the central evidence base. Criteria-only rows contributed to reporting-practice analysis when exact performance values could not be source-traced. Context-only benchmark rows were used to frame interpretation but were not treated as Kaggle-anchored performance evidence.

TABLE III
EVIDENCE-USE TIERS IN THE COMPLETED WORKBOOK.

Evidence-use tier	Rows	Use in manuscript
Primary quantitative	23	Main performance and reporting analyses
Secondary quantitative	3	Descriptive support only
Context-only benchmark	2	Background comparison, not pooled with Kaggle rows
Criteria-only / qualitative	2	Reporting-criteria analysis only

III. RESULTS

A. Evidence base

The completed workbook contained 30 source-traced rows. All 30 rows were retained for reporting-criteria analysis. For quantitative performance analysis, 23 rows were primary quantitative evidence, 3 rows were secondary quantitative evidence, 2 rows were context-only benchmarks, 1 row was criteria-only, and 1 row was excluded from absolute-performance plots because the extracted metric was not appropriate for direct plotting. Figure 1 summarizes these evidence-use categories.

Figure 1. Evidence-use flow for Stage 3 analysis

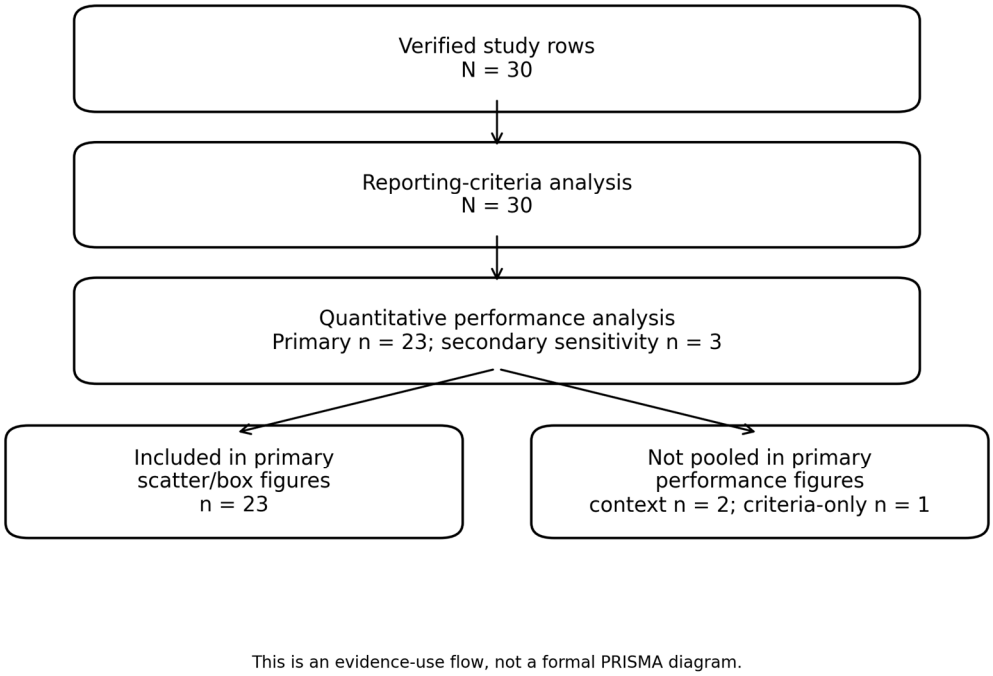


Fig. 1 Evidence-use flow for the 30 source-traced rows. The figure is an evidence-use diagram rather than a formal PRISMA diagram.

B. Reporting and validation coverage

External validation was clearly reported in 9 of 30 rows and partially reported in 4 rows. Confidence intervals were reported in 2 rows. Statistical testing was not clearly reported in any row and was partially described in 3 rows. Code availability was confirmed in 1 row. Leakage-control reporting was clearly described in 12 rows and partially described in 9 rows. These counts describe reporting completeness; they should not be interpreted as proof that unreported safeguards were absent.

TABLE IV
REPORTING-CRITERIA COVERAGE ACROSS ALL SOURCE-TRACED ROWS.

Criterion	Yes	Partial	No/Unknown
External validation	9	4	17
Confidence intervals	2	0	28
Statistical testing	0	3	27
Code availability	1	2	27
Leakage control described	12	17	1
Class imbalance handled	6	9	15
Patient/image split described	2	3	25
Calibration/uncertainty	1	0	29
Clinical metrics reported	13	13	4

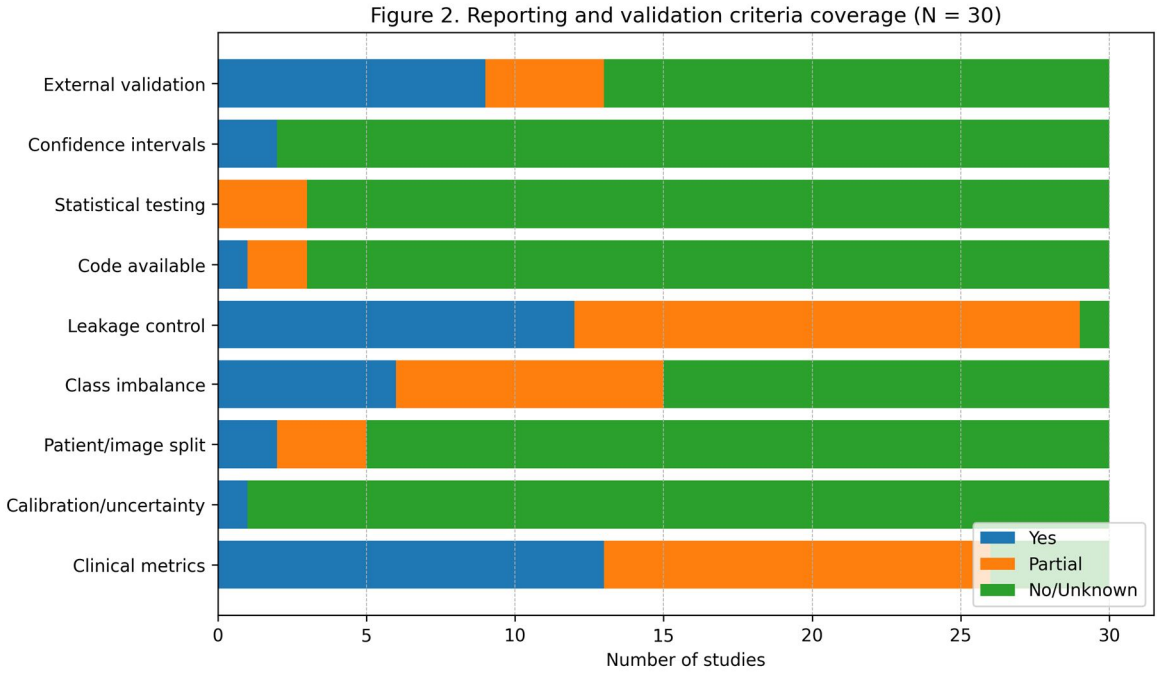


Fig. 2 Reporting-criteria coverage across the 30 source-traced rows. Counts are descriptive and do not imply causal effects.

C. Reliability-score distribution

The mean reliability-reporting score was 2.85/9 and the median was 2.50/9. Reliability scores were concentrated in low and moderate bands: 20 rows were low-reliability, 9 rows were moderate-reliability, 0 rows were high-reliability, and 1 row was uncoded. Under this rubric, a low score indicates limited reporting of reliability safeguards, not necessarily poor model design.

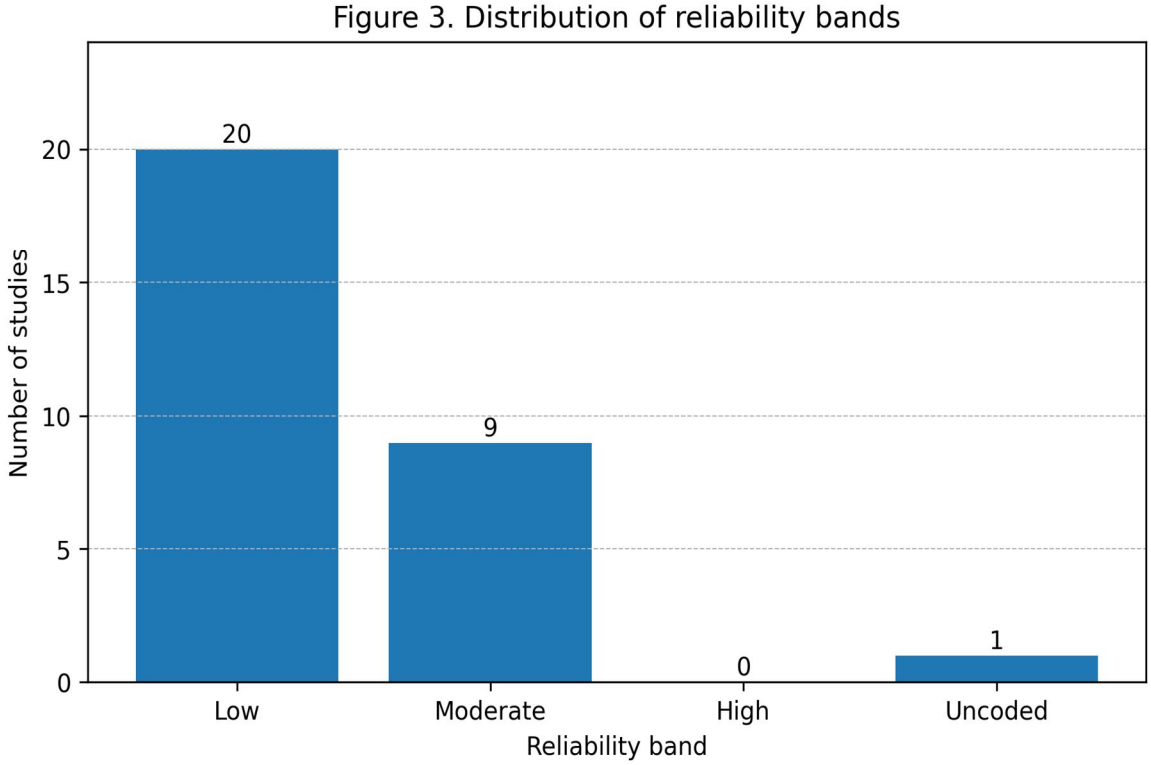


Fig. 3 Distribution of reliability bands based on the 0-9 rubric.

D. Reported performance and reliability score

AUROC/AUC values were extractable for 8 primary quantitative rows, and accuracy values were extractable for 14 primary quantitative rows. Because these metrics were drawn from heterogeneous tasks, datasets, model families, and validation procedures, the scatter plots are descriptive. They should not be interpreted as pooled estimates or as direct head-to-head comparisons between models.

Figure 4. Reported AUROC/AUC vs reliability score (n = 8)

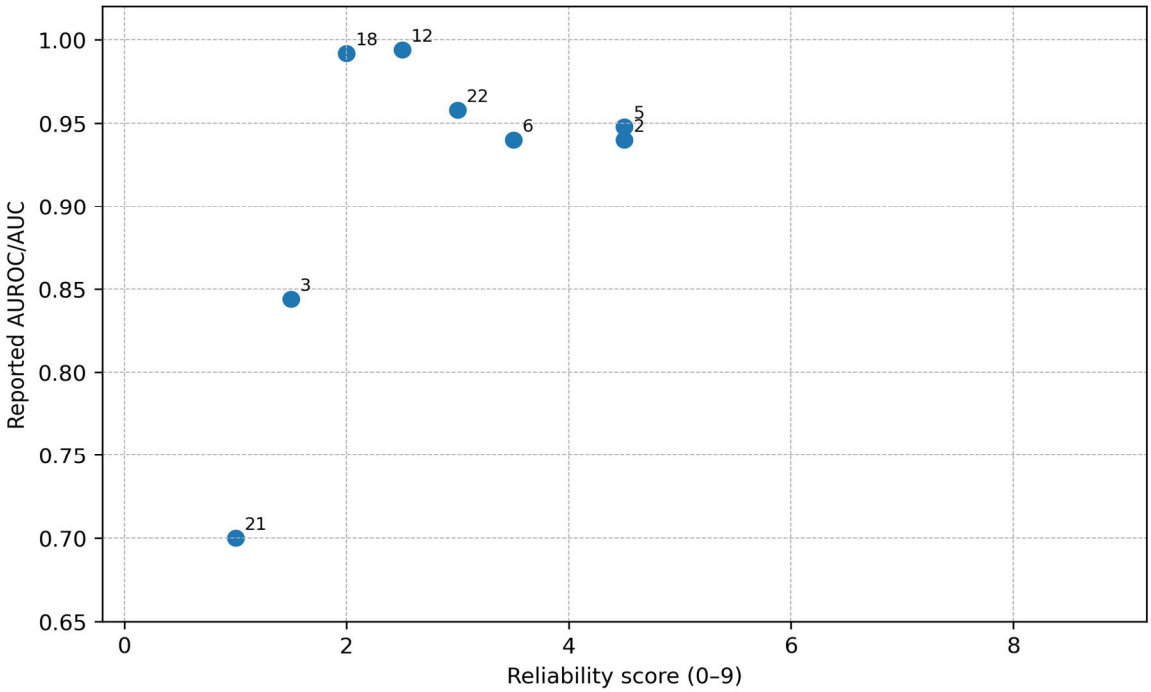


Fig. 4 Reported AUROC/AUC versus reliability score among primary quantitative rows with extractable AUROC/AUC values.

Figure 5. Reported accuracy vs reliability score (n = 14)

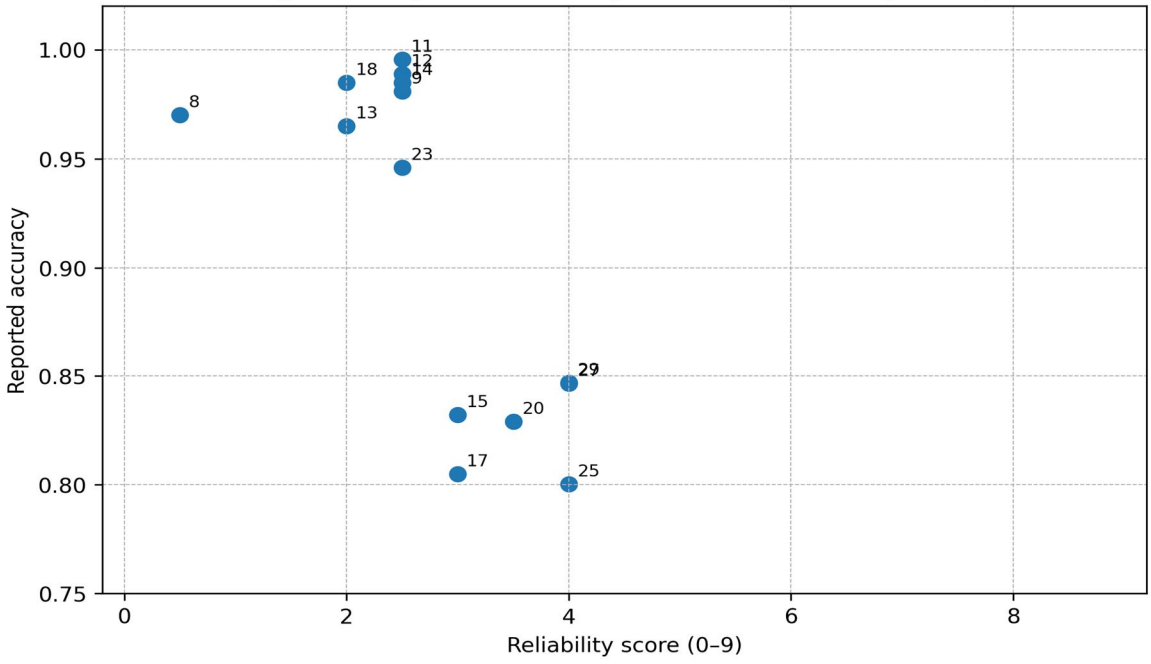


Fig. 5 Reported accuracy versus reliability score among primary quantitative rows with extractable accuracy values.

Figure 6. Reported accuracy distribution by reliability band

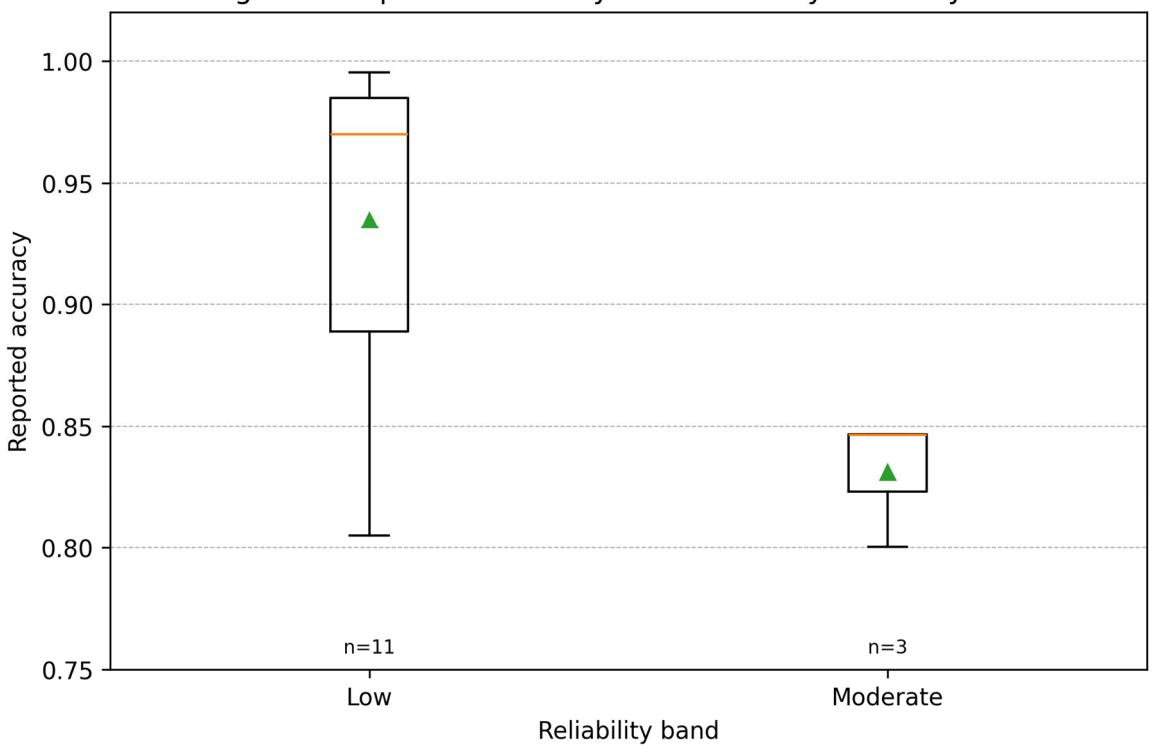


Fig. 6 Reported accuracy by reliability band. Box plots summarize heterogeneous tasks and should not be interpreted as pooled model performance.

E. Selected evidence examples

TABLE V
SELECTED EVIDENCE EXAMPLES ILLUSTRATING INTERNAL PERFORMANCE, EXTERNAL VALIDATION, AND CONTEXT BENCHMARKS.

ID	Authors/Year	Dataset(s)	Metric highlighted	Use tier	Reliability
1	Gulshan et al. (2016)	None	Acc NR; AUC 0.991	Context-only benchmark	6 (Moderate)
2	Voets, Møllersen, Bongo (2018)	None	Acc NR; AUC 0.94	Primary quantitative	4.5 (Moderate)
5	delaPava, Ríos, Rodríguez, Perdomo, González (2021)	None	Acc NR; AUC 0.948	Primary quantitative	4.5 (Moderate)
11	Mardianta, Affandy, Supriyanto, Wijaya (2025)	None	Acc 0.996; AUC NR	Primary quantitative	2.5 (Low)
12	Ahmed (2025)	None	Acc 0.989; AUC 0.994	Primary quantitative	2.5 (Low)
19	Hashmi (2026)	None	Acc NR; AUC NR	Primary quantitative	4 (Moderate)
25	Joshi, Sharma, Rathod, Sawant, Musale, Kalamkar (2026)	None	Acc 0.8; AUC NR	Primary quantitative	4 (Moderate)
27	Urooj, Banerjee, Shaikh, Thakur, Gupta (2025)	None	Acc 0.847; AUC NR	Primary quantitative	4 (Moderate)
29	Urooj, Banerjee, Gupta (2025)	None	Acc 0.847; AUC NR	Primary quantitative	4 (Moderate)
30	Sánchez-Gutiérrez et al. (2022)	None	Acc NR; AUC 0.988	Context-only benchmark	6 (Moderate)

IV. DISCUSSION

A. Principal findings

This secondary-data audit found that Kaggle-based diabetic-retinopathy AI studies often report strong performance, but reliability-related reporting is inconsistent. Of the 30 source-traced rows, only 9 clearly reported external validation, only 2 reported confidence intervals, and only 1 confirmed public code availability. These findings identify reporting gaps rather than proving that any specific model is unreliable.

B. Meaning for student medical-AI research

The findings are directly relevant to student researchers because public datasets make it possible to build impressive-looking AI projects quickly. A diabetic-retinopathy classifier can report high accuracy on an internal split while still lacking external validation, uncertainty intervals, patient-level split details, or code release. The audit therefore shifts the evaluation question from "Which model has the highest accuracy?" to "Which performance claim is supported by the strongest validation evidence?" This is a more mature framing for a first research paper because it avoids untested clinical claims and evaluates reporting quality using a transparent rubric.

C. External validation and generalization

External validation emerged as a central reliability marker. Rows using Messidor or Messidor-2 as external validation were more informative than Kaggle-only internal validation rows because they tested model behavior across a different data source. However, this audit does not claim that every externally validated model is clinically ready or that every internally validated model is invalid. The appropriate conclusion is narrower: external validation gives reviewers stronger evidence for generalization.

D. Why accuracy alone is insufficient

Accuracy was frequently reported, but accuracy alone is not enough for diabetic-retinopathy screening. A model can appear accurate when class distributions are skewed, while still missing clinically important referable cases. Sensitivity, specificity, AUROC/AUC, AUPRC, calibration, and confidence intervals provide more information about screening behavior. The low frequency of confidence intervals and calibration reporting in this audit suggests that many studies provide limited information about uncertainty, even when their headline metrics are high.

E. Reproducibility and leakage control

The audit also found gaps in reproducibility and leakage-control reporting. Only one row confirmed public code availability, and many rows described splitting or preprocessing only partially. In image classification, leakage can occur when preprocessing decisions, augmentation, duplicate handling, or threshold selection are influenced by validation or test images. The audit scored only what was clearly reported, so missing details should be treated as reporting uncertainty rather than confirmed methodological failure.

V. LIMITATIONS

This study has several important limitations. First, the audit is a structured secondary-data review rather than a formal systematic review or meta-analysis.

Although search queries and screening decisions were documented, the search may not have captured every relevant diabetic-retinopathy paper using APTOS, EyePACS, Messidor, or related datasets. Second, several metrics were source-page or abstract verified rather than extracted from complete full-PDF tables. Those rows were retained only with explicit verification-status labels. Third, included studies used heterogeneous datasets, tasks, class definitions, validation procedures, and model families, so their reported metrics cannot be pooled into a single diagnostic-performance estimate.

Fourth, the reliability score is an author-defined reporting rubric. Although each criterion was chosen for methodological relevance, the score has not been externally validated and should not be interpreted as a definitive measure of paper quality. Fifth, absence of reporting was scored as absence or uncertainty of evidence, not necessarily absence of the underlying practice. Sixth, this study did not rerun model code, reprocess image datasets, or independently validate any neural network. The conclusions are therefore limited to reporting reliability and evidence transparency.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

TABLE VI
MAIN LIMITATIONS AND MITIGATION STRATEGIES.

Limitation	Mitigation in this manuscript
Not a formal systematic review	The paper uses the more cautious label "structured secondary-data audit."
Heterogeneous metrics and tasks	AUROC and accuracy plots are separated and interpreted descriptively.
Preprint-heavy evidence base	Source type and verification level are retained in the workbook.
Author-defined reliability rubric	Rubric criteria are transparent and reproducible.
Missing reporting may not mean missing practice	The paper frames results as reporting reliability, not actual model invalidity.
No new model training	The manuscript explicitly states that it analyzes previously reported results only.

VI. CONCLUSION

This study presents a reproducible secondary-data audit of Kaggle-based and Kaggle-adjacent diabetic-retinopathy AI studies. Across 30 source-traced rows, reliability-related reporting was inconsistent: external validation, confidence intervals, code availability, calibration analysis, and detailed leakage-control reporting were not uniformly documented. The main conclusion is intentionally cautious: high internal Kaggle performance should be interpreted alongside validation design and reporting completeness. The study does not establish clinical performance of any model, but it provides a practical rubric for reliability-aware evaluation of public-dataset medical-AI claims.

VII. DATA AND MATERIALS AVAILABILITY

The extraction workbook, figure data, and generated figures are submitted as supplementary materials. The workbook includes source URLs, evidence-use tiers, verification-status labels, and reliability-scoring decisions. The analysis did not use private patient data, did not access individual medical records, and did not train a new model.

VIII. ETHICAL CONSIDERATIONS

This study used previously published, publicly available aggregate research results and did not involve human-subject data collection, private medical records, or clinical intervention. It should not be interpreted as a diagnostic medical-device study.

IX. AUTHOR CONTRIBUTIONS

Mudit Sharma conceived the research direction, selected the secondary-data audit design, compiled and organized the evidence workbook, developed the reliability-scoring framework, prepared the figures, and wrote and revised the manuscript.

X. CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

XI. ACKNOWLEDGMENT

The author thanks the researchers, dataset contributors, and open-data communities that made public diabetic-retinopathy research possible.

XII. AI-ASSISTED PREPARATION DISCLOSURE

AI assistance was used for organization, drafting support, editing, and document preparation. The author reviewed the manuscript, verified the source-tracing workbook, and remains responsible for the final content.

REFERENCES

[1] National Eye Institute. Diabetic Retinopathy. National Eye Institute. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy Accessed: 29 June 2026.
[2] Kaggle. APTOS 2019 Blindness Detection. https://www.kaggle.com/c/aptos2019-blindness-detection Accessed: 29 June 2026.

[3] Kaggle. Diabetic Retinopathy Detection. https://www.kaggle.com/c/diabetic-retinopathy-detection Accessed: 29 June 2026.
[4] Gulshan et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. https://jamanetwork.com/journals/jama/fullarticle/2588763.
[5] Voets, Møllersen, Bongo. (2018). Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. https://arxiv.org/abs/1803.04337.
[6] Islam, Hasan, Abdullah. (2018). Deep Learning based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images. https://arxiv.org/abs/1812.10595.
[7] Tymchenko, Marchenko, Spodarets. (2020). Deep Learning Approach to Diabetic Retinopathy Detection. https://arxiv.org/abs/2003.02261.
[8] delaPava, Ríos, Rodríguez, Perdomo, González. (2021). A deep learning model for classification of diabetic retinopathy in eye fundus images based on retinal lesion detection. https://arxiv.org/abs/2110.07745.
[9] Arrieta Ramos, Perdómo, González. (2022). Deep Semi-Supervised and Self-Supervised Learning for Diabetic Retinopathy Detection. https://arxiv.org/abs/2208.02408.
[10] Karkera, Adak, Chattopadhyay, Saqib. (2023). Detecting Severity of Diabetic Retinopathy from Fundus Images: A Transformer Network-based Review. https://arxiv.org/abs/2301.00973.
[11] Jain, Gupta, Singhal. (2024). Diabetic Retinopathy Detection Using Quantum Transfer Learning. https://arxiv.org/abs/2405.01734.
[12] Pakdelmoez et al. (2024). Controllable retinal image synthesis using conditional StyleGAN and latent space manipulation for improved diagnosis and grading of diabetic retinopathy. https://arxiv.org/abs/2409.07422.
[13] Alam, Roy, Kawsar, Rimi. (2024). Enhancing Transfer Learning for Medical Image Classification with SMOTE: A Comparative Study. https://arxiv.org/abs/2412.20235.
[14] Mardianta, Affandy, Supriyanto, Wijaya. (2025). Diabetic Retinopathy Detection Based on Convolutional Neural Networks with SMOTE and CLAHE Techniques Applied to Fundus Images. https://arxiv.org/abs/2504.05696.
[15] Ahmed. (2025). Addressing High Class Imbalance in Multi-Class Diabetic Retinopathy Severity Grading with Augmentation and Transfer Learning. https://arxiv.org/abs/2507.17121.
[16] Chaturvedi, Gupta, Ninawe, Prasad. (2020). Automated Diabetic Retinopathy Grading using Deep Convolutional Neural Network. https://arxiv.org/abs/2004.06334.
[17] Shakibania, Raoufi, Pourafkham, Khotanlou, Mansoorizadeh. (2023). Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy. https://arxiv.org/abs/2308.09945.
[18] Hannan, Mahmood, Qureshi, Ali. (2025). Enhancing Diabetic Retinopathy Classification Accuracy through Dual Attention Mechanism in Deep Learning. https://arxiv.org/abs/2507.19199.
[19] Kumar, Aditya, Kumar, Bikku, Thota, Kumar. (2025). Stage Aware Diagnosis of Diabetic Retinopathy via Ordinal Regression. https://arxiv.org/abs/2511.14398.
[20] Doshi, Shah. (2026). Bridging the Rural Healthcare Gap: A Cascaded Edge-Cloud Architecture for Automated Retinal Screening. https://arxiv.org/abs/2605.14108.
[21] Ahmed. (2025). HOG-CNN: Integrating Histogram of Oriented Gradients with Convolutional Neural Networks for Retinal Image Classification. https://arxiv.org/abs/2507.22274.
[22] Hashmi. (2026). Robust Diabetic Retinopathy Grading Using Dual-Resolution Attention-Based Deep Learning with Ordinal Regression. https://arxiv.org/abs/2604.17341.
[23] Islam, Jany, Ahmed, Khan. (2025). Balancing Accuracy and Efficiency: CNN Fusion Models for Diabetic Retinopathy Screening. https://arxiv.org/abs/2512.21861.
[24] Al-Kamachy, Hassanpour, Choupani. (2024). Classification of Diabetic Retinopathy using Pre-Trained Deep Learning Models. https://arxiv.org/abs/2403.19905.
[25] Zhu, Qiu, Lepore, Dumitrascu, Wang. (2022). Self-Supervised Equivariant Regularization Reconciles Multiple Instance Learning: Joint Referable Diabetic Retinopathy Classification and Lesion Segmentation. https://arxiv.org/abs/2210.05946.
[26] [26] Mok, Bum, Tai, Choo. (2024). Cross Feature Fusion of Fundus Image and Generated Lesion Map for Referable Diabetic Retinopathy Classification. https://arxiv.org/abs/2411.03618.
[27] [27] Kar, Akib, Hasib, Yaser, Azim. (2026). Temporal-Enhanced Interpretable Multi-Modal Prognosis and Risk Stratification Framework for Diabetic Retinopathy (TIMM-ProRS). https://arxiv.org/abs/2601.08240.
[28] [28] Joshi, Sharma, Rathod, Sawant, Musale, Kalamkar. (2026). Mobile-Ready Automated Triage of Diabetic Retinopathy Using Digital Fundus Images. https://arxiv.org/abs/2602.21943.
[29] [29] Florindo. (2026). Chaos-SSL: An Attention-Based Self-Supervised Learning Framework with Chaotic Transformation for Medical Image Classification. https://arxiv.org/abs/2605.27146.
[30] [30] Urooj, Banerjee, Shaikh, Thakur, Gupta. (2025). Single Domain Generalization in Diabetic Retinopathy: A Neuro-Symbolic Learning Approach. https://arxiv.org/abs/2509.02918.
[31] [31] Florindo, Ornelas. (2026). Attention-Based Chaotic Self-Supervision for Medical Image Classification. https://arxiv.org/abs/2605.04985.
[32] [32] Urooj, Banerjee, Gupta. (2025). NEURO-GUARD: Neuro-Symbolic Generalization and Unbiased Adaptive Routing for Diagnostics: Explainable Medical AI. https://arxiv.org/abs/2512.18177.
[33] [33] Sánchez-Gutiérrez et al. (2022). Performance of a deep learning system for detection of referable diabetic retinopathy in real clinical settings. https://arxiv.org/abs/2205.05554.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

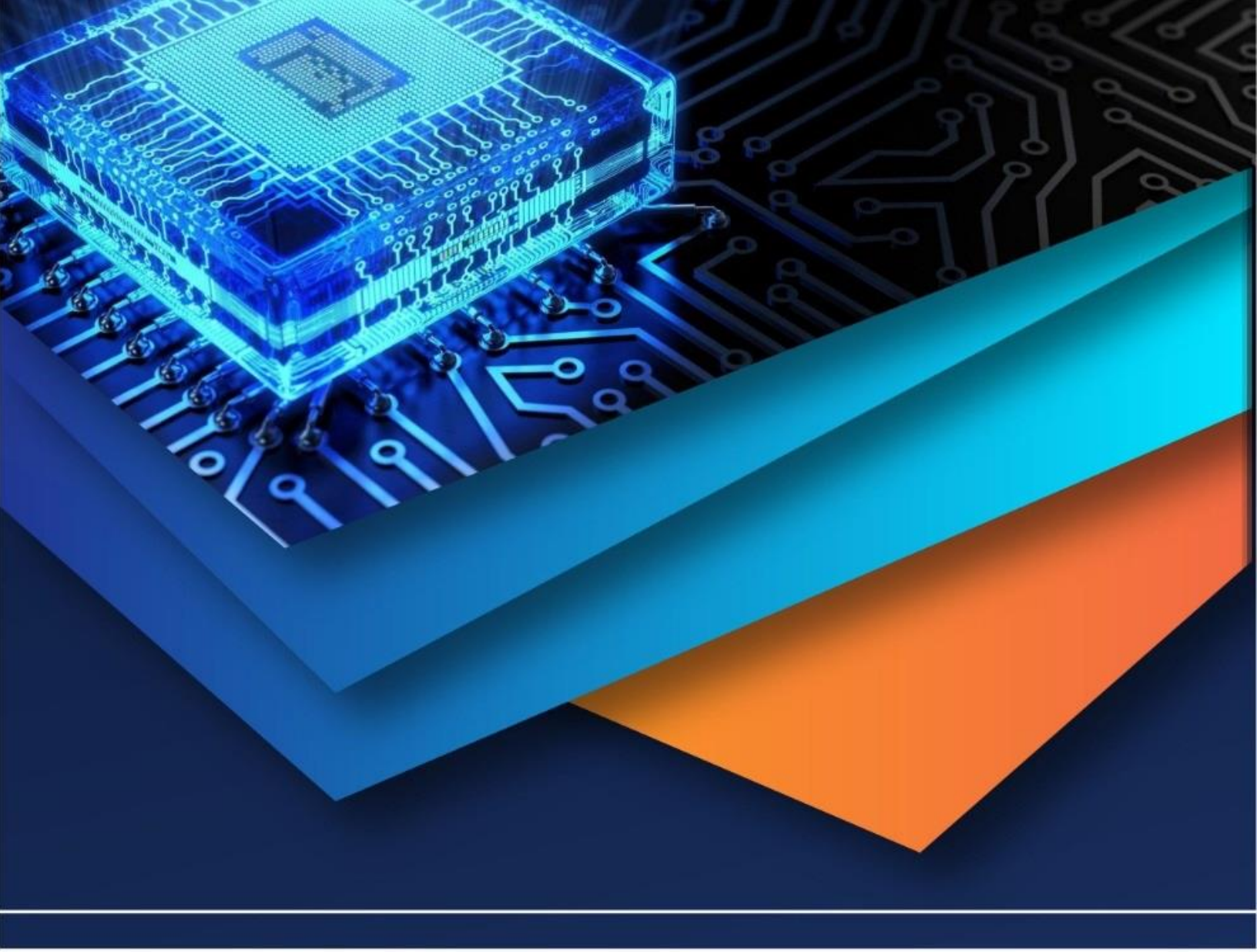
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

SUPPLEMENTARY MATERIALS

Table S1. Suggested supplementary package.

File	Purpose
DR_Kaggle_Reliability_Audit_Workbook_v4_stage3_figures.xlsx	Full extraction table, rubric, dashboard, figure data, verification notes
DR_Kaggle_Reliability_Audit_Stage3_Figures.zip	All generated figure PNGs
DR_Kaggle_Reliability_Audit_Stage5_Full_Manuscript.docx	Full manuscript draft
DR_Kaggle_Reliability_Audit_Stage5_Full_Manuscript.pdf	Rendered PDF preview for review

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)