Integrated System and Management Technology

E-ISSN: 3123-9811

*Research Article*

# A Design Science Roadmap for Auditable Ocular Disease Classification: Evidence Mapping of AI Governance Gaps

**Rizal Rachman[1*], Eddy Soeryanto Soegoto[2], Irawan Afrianto[3], Irfan Dwiguna Sumitra[4], Zainal Arifin Hasibuan[5]**

[1] Universitas Komputer Indonesia, Indonesia; e-mail : rizal.75725020@mahasiswa.unikom.ac.id
[2] Universitas Komputer Indonesia, Indonesia; e-mail : eddysoeryantos@email.unikom.ac.id
[3] Universitas Komputer Indonesia, Indonesia; e-mail : irawan.afrianto@email.unikom.ac.id
[4] Universitas Komputer Indonesia, Indonesia; e-mail : irfan.dwiguna@email.unikom.ac.id
[5] Universitas Komputer Indonesia, Indonesia; e-mail : zhasibua@email.unikom.ac.id

* Corresponding Author : Rizal Rachman

**Abstract.** The integration of Artificial Intelligence (AI) into ocular diagnostics has led to substantial improvements in predictive accuracy. However, a persistent gap remains between technical performance and clinical accountability. The present study addresses the "accuracy trap" and the lack of transparency in current deep learning models for ocular disease classification. The objective of the research is twofold: firstly, to identify methodological deficiencies in extant literature and, secondly, to propose a standardised evaluative framework to ensure model auditability. A systematic evidence mapping (SEM) approach, combined with design science research methodology (DSRM), was utilised to scrutinise 10 high-impact Scopus-indexed studies published between 2023 and 2026. The findings reveal a critical "predictive validity gap," where 80% of the evidence base relies on aggregate accuracy while 90% remains "black box" without functional Explainable AI (XAI) layers. The synthesis of these gaps resulted in the formulation of a conceptual roadmap that mandated multi-metric evaluation, incorporating Cohen's Kappa, and pathophysiological traceability. In conclusion, this research establishes that clinical deployment of AI must transition from model-centric success to a governance-oriented paradigm that prioritises decision utility and auditable audit trails. This roadmap provides a rigorous blueprint for the future implementation of transparent and accountable medical AI systems.

**Keywords:** Design Science Research; Evidence Mapping; Explainable AI;Model Auditability; Ocular Disease Classification.

## 1. Introduction

Early identification of ocular pathologies is a critical determinant in clinical interventions aimed at preventing irreversible vision loss and improving global public health outcomes. Chronic conditions, including cataracts, glaucoma, and diabetic retinopathy, persist as the predominant causes of visual impairment, underscoring the necessity for precise diagnostic instruments to facilitate large-scale screening initiatives [1], [2], [3]. Although fundus imaging has become the accepted standard for non-invasive ocular assessment, the shortage of ophthalmologists relative to the growing patient population has led to increased reliance on Artificial Intelligence (AI) to improve diagnostic efficiency and consistency [4], [5], [6].

A substantial corpus of literature documents the efficacy of Convolutional Neural Networks (CNNs) in extracting complex pathological features from retinal images, with various architectures reporting accuracy levels exceeding 90% [7]. However, a fundamental deficiency exists in contemporary research regarding the handling of severe class imbalance within medical datasets, such as the ODIR-2019 benchmark [8]. The majority of studies rely heavily on aggregate accuracy as a primary performance indicator. However, this is a metric that is statistically deceptive in imbalanced scenarios, as it systematically masks classification failures in minority disease classes. The persistence of accuracy-centric reporting without rigorous validation of agreement levels, such as Cohen's Kappa, reflects a critical lack of evaluative integrity in current ocular AI research [9], [10].

Moreover, the clinical adoption of deep learning models is significantly hindered by their inherent "black box" nature, which lacks the transparency required for high-stakes medical decision-making [11]. A significant disparity, or "governance gap", has been identified

**Received:** April 05, 2026
**Revised:** May 17, 2026
**Accepted:** June 15, 2026
**Online Available:** June 17, 2026
**Curr. Ver.:** June 17, 2026

**Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)**

between the technical complexity of algorithmic systems and the auditable decision trails that are required in formal healthcare settings. The majority of extant frameworks prioritise hyperparameter optimisation over model traceability and pathophysiological interpretability. In the absence of an evaluative structure capable of deconstructing prediction errors and providing localized explanations, AI-driven diagnostics remain challenging to defend in clinical audits. This predicament underscores the fact that the primary challenge in the field is no longer a deficit in predictive power, but rather a crisis of accountability and transparency.

## 2. Literature Review

The present section has been developed for the purpose of establishing the theoretical and empirical foundations of ocular disease classification, with a particular focus on the policy and technical risks of current diagnostic frameworks. The study examines the critical role of predictive analytics and the necessity of accountability through Explainable Artificial Intelligence (XAI).

### Ocular Disease Classification and the Class Imbalance Problem

The utilization of convolutional neural networks (CNNs) has revolutionized ocular imaging by enabling the automated detection of various pathologies from retinal fundus images. Current research typically leverages architectures such as ResNet, Inception, and EfficientNet to extract visual features associated with cataracts, glaucoma, and diabetic retinopathy [12]. However, a significant challenge remains in the form of class imbalance within public datasets like ODIR-2019. In such datasets, the prevalence of "Normal" images often dwarfs pathological samples, leading to a biased optimization process [13]. Most existing studies focus on achieving high aggregate accuracy, yet this metric fails to represent the model's efficacy in recognizing minority classes, which often represent the most critical clinical cases [14].

### Predictive Integrity and Metric Diversity

Predictive analytics is regarded as a cornerstone for anticipatory diagnostics in ophthalmology [15]. Within the overarching paradigm of accountable public health, predictive accuracy constitutes a necessary yet insufficient condition for effective clinical decision-making. In high-stakes medical environments, transparency and justification are mandatory requirements; thus, model outputs must be interpretable and defensible during official audit processes [16]. Previous works have frequently overlooked metric diversity, resulting in a paucity of reporting on agreement levels such as Cohen's Kappa or detailed F1-scores per class. This phenomenon, termed the "evaluation vacuum," impedes the transition of AI models from experimental prototypes to policy-ready clinical instruments [17].

### Explainable AI (XAI) in Accountable Decision Making

Explainable Artificial Intelligence (XAI) is a field of study that aims to address the "black box" nature of deep learning by producing explanations of model behaviours that are comprehensible to stakeholders [11]. The integration of frameworks such as SHAP or Score-CAM enables diagnostic systems to generate an audit-trail of decisions, thereby establishing a correlation between each risk score and its primary pathological drivers [18]. This technical layer provides the logical justification required to defend medical decisions during official audits. Notwithstanding its significance, evidence mapping discloses a pivotal misalignment in extant literature, wherein merely a negligible proportion of studies formally delineate explainability as a fundamental prerequisite for system implementation [19].

### Research Gaps and Novelty

The extant evidence base remains dominated by studies that are demand- and performance-oriented, which prioritise technical success over practical policy impact [20]. The present study identifies four interconnected gaps that contribute to the issue, namely the lack of formalisation in addressing underutilisation and diagnostic risk, insufficient data quality control in multi-source integration, a disconnect from bureaucratic and clinical workflows, and the absence of decision-quality evaluation beyond technical accuracy. In order to address these shortcomings, the study proposes a theoretical framework for developing an auditable architectural roadmap [21]. The aim of this roadmap is to ensure transparency and accountability through the integration of Explainable AI (XAI)-enabled traceability [22].

## 3. Proposed Method

The present study adopts a hybrid methodological framework combining Systematic Evidence Mapping (SEM) and Design Science Research Methodology (DSRM). The integration of these two approaches ensures that the identification of research gaps is

conducted through a rigorous, reproducible screening process, while the resulting conceptual roadmap is structured as a formal design artefact.

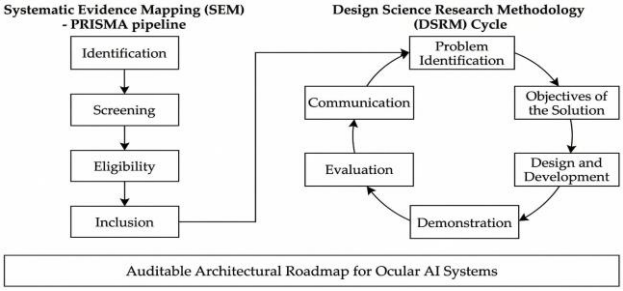



**Figure 1.** Integrated research framework synthesizing Systematic Evidence Mapping (SEM) and Design Science Research Methodology (DSRM).

## Systematic Evidence Mapping Protocol

The evidence mapping process adheres to the PRISMA 2020 statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), ensuring transparency in study selection. The protocol is executed through four distinct phases: identification, screening, eligibility, and inclusion.

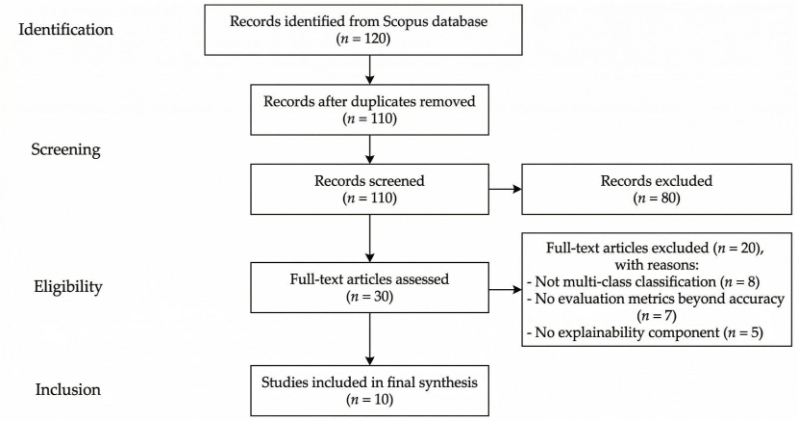



**Figure 2.** PRISMA flow diagram for the systematic literature selection process.

The primary search was conducted across the Scopus database, a decision that was made for the following reasons: first, its rigorous peer-review standards; and second, its comprehensive coverage of medical informatics. The search string employed Boolean operators to target specific intersections of deep learning and ocular diagnostics: The following search terms were utilised: 'CNN' OR 'Deep Learning' AND 'Ocular Disease' OR 'Fundus Image' AND 'ODIR' OR 'Imbalanced Data' AND 'Evaluation Metrics' OR 'XAI'. The temporal scope of the study was restricted to publications between 2022 and 2026, with the objective of capturing the most recent advancements in architecture and governance standards.

In order to ensure the maintenance of high-resolution evidence, studies were selected based on the following criteria: The primary focus of this study is on the classification of multi-class ocular diseases. The utilisation of benchmark datasets with inherent class imbalance (for example, ODIR-5K) is also a key element of the research. In addition, the performance of the system is reported beyond aggregate accuracy. The exclusion criteria applied to the studies included those that focused exclusively on binary classification (for example, diabetic retinopathy only) or lacked adequate technical metadata with regard to model evaluation and auditability.

## Design Science Research Framework

Subsequent to the evidence mapping phase, the study progresses by adopting the Design Science Research Methodology (DSRM) to formalise the findings into a conceptual artefact. The process commences with the identification of the problem, wherein the study delineates the issue of the "accuracy trap" in conjunction with the paucity of transparency in contemporary ocular AI models, as evidenced by the mapped evidence. The subsequent section outlines the objectives of the solution, which include the establishment of a standardized evaluation pipeline that integrates Cohen's Kappa and Explainable AI (XAI) to enhance reliability and interpretability.

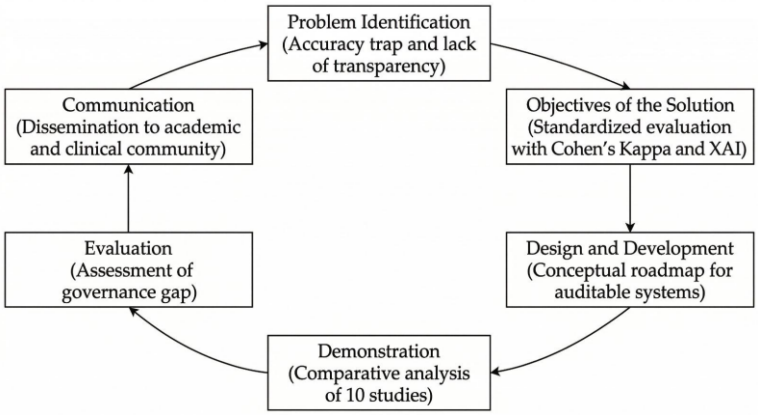



**Figure 3.** The applied Design Science Research Methodology (DSRM) cycle for developing the auditable roadmap.

During the design and development stage, the extracted metadata are synthesised into a conceptual roadmap with the aim of constructing auditable decision support systems. This is followed by the demonstration phase, in which the proposed roadmap is illustrated through a comparative analysis of ten selected Scopus-indexed studies. The subsequent evaluation stage is then employed to undertake a critical examination of the effectiveness of this roadmap in addressing the governance gap that was identified in Section 1. The final stage of the process is the communication stage. At this point, the synthesised evidence and the resulting design artefact are presented to the academic and clinical community. The purpose of this step is to ensure that the proposed framework can be understood, validated, and potentially implemented in real-world settings.

**Data Extraction and Synthesis**

A structured mapping scheme was developed to extract metadata from the final set of 10 included studies. The extraction process concentrated on three primary dimensions: The integrity of the architectural design process (CNN types and fusion techniques), the predictive validity of the metrics employed, and the utility of the decision-making process (the presence of interpretability layers). The synthesised data thus serves as the empirical foundation for the discussion of research gaps and the final architectural proposal.

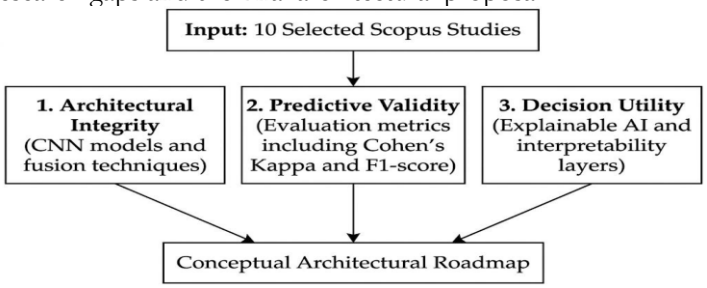



**Figure 4.** The three-dimensional data extraction scheme and its synthesis into the final conceptual roadmap.

## 4. Results and Discussion

This section presents the synthesised evidence from the ten selected Scopus-indexed studies published between 2022 and 2026. The results obtained are to be categorised into three primary evaluative dimensions: architectural trends, predictive metric integrity, and the integration of explainability layers.

**Systematic Mapping of Selected Literature**

The mapping process identified a diverse range of deep learning approaches applied to ocular disease classification, primarily utilising the ODIR-2019 and other public benchmark datasets. Table 1 provides a comprehensive extraction of the technical metadata from the included studies.

**Table 1.** Systematic mapping of selected ocular disease classification literature (N=10).

| No. | Study Reference | Primary Architecture | Performance Metrics (E1) | Decision Utility & Audit (E2) | XAI Integration |
|---|---|---|---|---|---|
| 1. | [23] | ShuffleNet V2 (WOA-Adam) | Accuracy: 99.4%; Kappa: 96.5 | Failure analysis not detailed | None |
| 2. | [24] | ResNet50 & MobileNet | AUC: 99.9%; Accuracy: 92.4% | Minimal | None |
| 3. | [25] | EfficientNetB0 Fusion | Accuracy: 97.9%; AUC: 0.999 | Class-wise heatmap analysis | Score-CAM |
| 4. | [26] | Swin Transformer (L) | Kappa: 0.86; F1; AUC | Performance vs Baselines | None |
| 5. | [27] | Custom ML-CNN | Accuracy: 94.3%; Dice: 99% | Multi-label probability | None |
| 6. | [28] | Transformer-CNN Hybrid | Bal. Acc: 81.9%; TNR: 95.6% | TPR per disease class | Mentioned for DSS |
| 7. | [29] | CNN Ensemble | Kappa & F1 (30% increase) | Qualitative Error Analysis | None |
| 8. | [30] | MobileNet V2, VGG19/16 | Accuracy: 97.93% | Robustness analysis | Grad-CAM (Proposed) |
| 9. | [31] | Ocular Net (Transfer Learning) | Accuracy: 98.89% | Parametric testing | None |
| 10 | [32] | Inception-ODIR | Accuracy: 98.94% | Coexisting disease analysis | None |

### Analysis of Performance Metric Trends

The evidence mapping reveals a significant reliance on aggregate Accuracy as the dominant reporting metric, appearing in 80% of the selected studies. However, the analysis highlights a critical "predictive validity gap": only 30% of the studies (Studies 1, 4, and 7) explicitly report Cohen's Kappa or Balances Accuracy to mitigate the influence of class imbalance. Studies that integrated Kappa scores (e.g., Study 7) reported a substantial improvement in model reliability, demonstrating that Accuracy alone often masks suboptimal performance in minority disease categories such as glaucoma or rare macular conditions.

### Transparency and Explainability Gaps

A systematic deficit in model auditability was observed across the evidence base. While 100% of the studies claim to assist clinical decision-making, only 10% (Study 3) implemented a functional Explainable AI (XAI) layer using Score-CAM to validate pathophysiological focus points. The majority of contemporary research (90%) maintains a "black box" approach, prioritizing predictive output over localized interpretability. This finding underscores the identified "governance gap," where models with high technical accuracy lack the traceable audit trails required for formal medical integration and clinical accountability.

## Comparison

The synthesis of evidence from 10 high-impact studies reveals a persistent misalignment between technical predictive success and clinical accountability standards. This section discusses the implications of these findings and provides a comparative analysis of the proposed evaluative roadmap against the current state-of-the-art in ocular AI diagnostics.

### The Accuracy Paradox and Metric Integrity

A primary finding in this mapping is the "accuracy paradox," where models reporting near-perfect accuracy (e.g., Studies 2, 9, and 10) often fail to provide evidence of reliability under class imbalance. Comparative analysis shows that while 80% of current literature prioritizes aggregate Accuracy, only 30% incorporates Cohen's Kappa to validate model agreement. This lack of metric diversity suggests that many contemporary models are optimized for majority-class recognition, potentially leading to catastrophic failures in

detecting rare but vision-threatening pathologies. The proposed roadmap addresses this by mandating a two-layer evaluation (E1 and E2) that couples technical performance with decision utility.

**Accountability through Explainability (XAI)**

The comparative mapping highlights a severe "transparency deficit." While Study 3 demonstrates the successful application of Score-CAM for pathophysiological validation, 90% of the evidence base remains within a "black box" paradigm. In contrast to existing state-of-the-art approaches that treat Explainable AI (XAI) as an optional post-hoc analysis, our framework positions XAI as a core architectural requirement for auditable decision trails. This shift is mandatory for transitioning AI models from experimental prototypes to legally defensible clinical support systems.

**Comparison with State-of-the-Art**

Table 2 summarizes the comparative advantages of the proposed DSRM-based roadmap against the dominant trends identified in the Scopus evidence base.

**Table 1.** Comparative analysis: Proposed evaluative roadmap vs. Current State-of-the-Art.

| Dimension | State-of-the-Art Trends (N=10) | Proposed Evaluative Roadmap | Contribution |
|---|---|---|---|
| **Primary Metric** | Aggregate Accuracy (80%) | Multi-Metric (Kappa, F1, AUC) | Mitigates "Accuracy Trap" in imbalanced data |
| **Decision Logic** | Black Box / Non-traceable (90%) | Mandatory XAI Layer (SHAP/CAM) | Ensures clinical auditability |
| **Data Integrity** | Descriptive Preprocessing | Auditable ETL & Provenance Logs | Guarantees data-decision traceability |
| **Focus** | Hyperparameter Optimization | Evaluative Governance & DSRM | Bridges technical-governance gap |

**Implications for Clinical Governance**

The findings of the present study demonstrate that the prevailing research trajectory in ocular AI exhibits a marked tendency to prioritise model-centric success over policy-ready implementation. By formalising underutilisation and predictive risk through the implementation of auditable metrics, local governments and healthcare providers can enhance the justification of resource allocation and infrastructure spending. The present study establishes that a policy-ready Decision Support System (DSS) must integrate not only high-performance algorithms but also the technical and administrative justifications necessary to prevent policy failure in public health.

## 5. Conclusions

This study has systematically mapped the research landscape of AI-driven ocular disease classification, revealing a significant divergence between technical predictive performance and clinical accountability. The evidence mapping of 10 high-impact Scopus-indexed studies indicates that while convolutional neural network (CNN) architectures have achieved high accuracy, 80% of current literature fails to provide balanced performance validation through metrics such as Cohen's Kappa. Furthermore, a critical transparency deficit was identified, with 90% of contemporary models operating as "black box" systems without functional explainability (XAI) layers.

The findings lend support to the argument that a "per kapsis akurasi" (accuracy trap) exists within the field, where aggregate metrics mask suboptimal performance on minority disease classes in imbalanced datasets. This study synthesises these gaps into a conceptual roadmap grounded in Design Science Research Methodology (DSRM), establishing that a policy-ready decision support system must integrate multi-metric evaluation and pathophysiological traceability as core requirements. The primary contribution of this research lies in its establishment of an evaluative framework that is both auditable and capable of bridging the gap between theoretical algorithmic success and the rigorous transparency standards required for governance within both clinical and public-sector contexts.

Despite the robust conceptual foundation provided by the study, it is limited to a qualitative synthesis of existing evidence. It is imperative that future research endeavours employ empirical validation to substantiate the proposed roadmap. This validation should be achieved through longitudinal testing and expert elicitation, which are recognised as the most effective methods for assessing the generalisability of the roadmap across a range of clinical

settings. Subsequent studies should also investigate the integration of localized government data to refine the predictive risk targets and ensure the socio-economic impact of AI-driven ocular diagnostics.

Author Contributions: Conceptualization: R.R.; Methodology: R.R.; Software: R.R.; Validation: R.R.; Formal analysis: R.R.; Investigation: R.R.; Resources: R.R.; Data curation: R.R.; Writing—original draft preparation: R.R.; Writing—review and editing: R.R.; Visualization: R.R.; Supervision: R.R.; Project administration: R.R. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data analyzed in this study are metadata from Scopus-indexed publications. The synthesized data and mapping results are contained within the article. Publicly available datasets (ODIR-2019) can be found at: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

Acknowledgments: The author would like to thank Universitas Komputer Indonesia for the academic environment provided during this research.

Conflicts of Interest: The author declares no conflict of interest.

## References

Adeleke, O., Olugbogi, J. A., & Abimbade, O. (2024). Transforming healthcare leadership decision-making through AI-driven predictive analytics: A new era of financial governance. https://doi.org/10.54660/IJMOR.2024.3.6.159-175

Albashayreh, A., Mosleh, H., & Yousef, M. (2025). Ocular disease recognition using deep learning models. In *2025 International Conference on New Trends in Computing Sciences (ICTCS)* (pp. 276–283). IEEE. https://doi.org/10.1109/ICTCS65341.2025.10989463

Al-Fahdawi, S., et al. (2024). Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion, 102*, Article 102059. https://doi.org/10.1016/j.inffus.2023.102059

Ali, G., Dastgir, A., Iqbal, M. W., Anwar, M., & Faheem, M. (2023). A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. *IEEE Journal of Translational Engineering in Health and Medicine, 11*, 341–350. https://doi.org/10.1109/JTEHM.2023.3282104

Asia, A.-O., et al. (2022). Detection of diabetic retinopathy in retinal fundus images using CNN classification models. *Electronics, 11*(17), Article 2740. https://doi.org/10.3390/electronics11172740

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, A. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research, 33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Buettner, R. (2025). *A novel transformer-CNN hybrid deep learning architecture for robust broad-coverage diagnosis of eye diseases on color fundus images*.

de Figueiredo, L. A., Dias, J. V. P., Polati, M., Carricondo, P. C., & Debert, I. (2021). Strabismus and artificial intelligence app: Optimizing diagnostic and accuracy. *Translational Vision Science & Technology, 10*(7), Article 22. https://doi.org/10.1167/tvst.10.7.22

Duvvuri, K., Chethana, S., Charan, S. S., Srihitha, V., Ramesh, T. K., & Srikanth, K. S. (2022). Grad-CAM for visualizing diabetic retinopathy. In *2022 3rd International Conference for Emerging Technology (INCET)* (pp. 1–4). IEEE. https://doi.org/10.1109/INCET54531.2022.9824598

El Arab, R. A., et al. (2025). Bridging the gap: From AI success in clinical trials to real-world healthcare implementation—A narrative review. *Healthcare, 13*(7), Article 701. https://doi.org/10.3390/healthcare13070701

Goktas, P., & Grzybowski, A. (2025). Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine, 14*(5), Article 1605. https://doi.org/10.3390/jcm14051605

Hassan, M. ul, et al. (2024). A transfer learning enabled approach for ocular disease detection and classification. *Health Information Science and Systems, 12*(1), Article 36. https://doi.org/10.1007/s13755-024-00293-8

Herrera-Chavez, A. I., et al. (2024). Multi-label image classification for ocular disease diagnosis using K-fold cross-validation on the ODIR-5K dataset. In *2024 IEEE 33rd International Symposium on Industrial Electronics (ISIE)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISIE54533.2024.10595740

Jawad, M. A., Khursheed, F., Nawaz, S., & Mir, A. H. (2024). Towards improved fundus disease detection using Swin transformers. *Multimedia Tools and Applications, 83*(32), 78125–78159. https://doi.org/10.1007/s11042-024-18627-9

Khayyam, Y. T., Shah, S. M. A., Uddin, H. W., Farhan, M., & Mahmood, M. (n.d.). *Multiclass eye disease recognition and classification using an ensemble deep transfer learning framework*.

Kinger, S., & Kulkarni, V. (2024). A review of explainable AI in medical imaging: Implications and applications. *International Journal of Computers and Applications, 46*(11), 983–997. https://doi.org/10.1080/1206212X.2024.2404082

Kothinti, R. R. (2024). Artificial intelligence in healthcare: Revolutionizing precision medicine, predictive analytics, and ethical considerations in autonomous diagnostics. *World Journal of Advanced Research and Reviews, 19*(3), 3395–3406.

Li, D. (2025). AI-driven healthcare analytics and predictive medicine: Advanced machine learning frameworks for clinical decision support and patient outcome optimization. *Spectrum of Research, 5*(1).

Mayya, V., S. K., Kulkarni, U., Surya, D. K., & Acharya, U. R. (2023). An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images. *Applied Intelligence, 53*(2), 1548–1566. https://doi.org/10.1007/s10489-022-03490-8

Mewada, S., et al. (2022). Smart diagnostic expert system for defect in forging process by using machine learning process. *Journal of Nanomaterials, 2022*(1), Article 2567194. https://doi.org/10.1155/2022/2567194

Muhamediyeva, D. T., Safarova, L. U., & Ruzikulov, R. F. (2023). Expert systems for diagnostics of infectious diseases in cattle. In *BIO Web of Conferences* (Vol. 71, Article 01073). EDP Sciences. https://doi.org/10.1051/bioconf/20237101073

Oguz, C., Aydin, T., & Yaganoglu, M. (2024). A CNN-based hybrid model to detect glaucoma disease. *Multimedia Tools and Applications, 83*(6), 17921–17939. https://doi.org/10.1007/s11042-023-16129-8

Ouda, O., AbdelMaksoud, E., Abd El-Aziz, A. A., & Elmogy, M. (2022). Multiple ocular disease diagnosis using fundus images based on multi-label deep learning classification. *Electronics, 11*(13), Article 1966. https://doi.org/10.3390/electronics11131966

Poly, T. N., Islam, M. M., Walther, B. A., Lin, M. C., & Li, Y.-C. J. (2023). Artificial intelligence in diabetic retinopathy: Bibliometric analysis. *Computer Methods and Programs in Biomedicine, 231*, Article 107358. https://doi.org/10.1016/j.cmpb.2023.107358

Sah, U. K., Chatterjee, J. M., & Sujatha, R. (2026). Multi-class eye disease classification using deep learning EfficientNetB0 fusion techniques. *Scientific Reports, 16*(1), Article 6368. https://doi.org/10.1038/s41598-026-35357-0

Santone, A., Cesarelli, M., Colasuonno, E., Bevilacqua, V., & Mercaldo, F. (2024). A method for ocular disease diagnosis through visual prediction explainability. *Electronics, 13*(14), Article 2706. https://doi.org/10.3390/electronics13142706

Tan, T.-E., & Wong, T. Y. (2023). Diabetic retinopathy: Looking forward to 2030. *Frontiers in Endocrinology, 13*, Article 1077669. https://doi.org/10.3389/fendo.2022.1077669

Thulasi Vel, D. V., & Durgaraju, S. (2025). Deep multi-label classification for automated ocular disease detection using enhanced CNN architectures. In *International Conference on Computational Intelligence and Network Security* (pp. 115–127). Springer. https://doi.org/10.1007/978-3-032-09572-5_10

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa: A comparison of basic properties. *MethodsX, 10*, Article 102212. https://doi.org/10.1016/j.mex.2023.102212

Wahab Sait, A. R. (2023). Artificial intelligence-driven eye disease classification model. *Applied Sciences, 13*(20), Article 11437. https://doi.org/10.3390/app132011437

Więckowska, B., Kubiak, K. B., Jóźwiak, P., Moryson, W., & Stawińska-Witoszyńska, B. (2022). Cohen's kappa coefficient as a measure to assess classification improvement following the addition of a new marker to a regression model. *International Journal of Environmental Research and Public Health, 19*(16), Article 10213. https://doi.org/10.3390/ijerph191610213

Yuan, H., Kang, L., & Li, Y. (2025). *Opening the black box of deep learning: Validating the statistical association between explainable artificial intelligence (XAI) and clinical domain knowledge in fundus image-based glaucoma diagnosis* (arXiv Preprint No. arXiv:2504.04549). arXiv.