information

MDPI

Article

Evaluating the Generalisability of Convolutional Neural Networks for Diabetic Retinopathy Detection in Latin America and Sub-Saharan Africa

Rogers Mwavu [1,*], Fred Kaggwa [2], Simon Arunga [3] and William Wasswa [4]

[1] Information Technology Department, Mbarara University of Science and Technology, Mbarara P.O. Box 1410, Uganda
[2] Computer Science Department, Mbarara University of Science and Technology, Mbarara P.O. Box 1410, Uganda; kaggwa_fred@must.ac.ug
[3] Ophthalmology Department, Mbarara University of Science and Technology, Mbarara P.O. Box 1410, Uganda; sarunga@must.ac.ug
[4] Department of Biomedical Sciences and Engineering, Mbarara University of Science and Technology, Mbarara P.O. Box 1410, Uganda; wwasswa@must.ac.ug
* Correspondence: rmwavu@must.ac.ug; Tel.: +256-773426328

Abstract

Diabetic retinopathy is a leading cause of vision loss worldwide, particularly impacting individuals in low- and middle-income countries with limited healthcare access. Early detection through automated screening systems is essential for improving outcomes, as timely intervention can prevent severe vision impairment. However, most of the available AI models have not been evaluated in low-resource settings. Hence, this study presents an evaluation of the efficacy of advanced deep learning architectures for detecting rDR across diverse population datasets. A dual-phase validation approach was employed to assess model performance. Internal validation utilised the BrSET dataset to establish baseline performance metrics, while external validation was conducted on the MoDRIA dataset, which encompasses various conditions and demographics, to evaluate model robustness. Key performance metrics, including accuracy, specificity, sensitivity, F1-score, and calibration scores, were systematically recorded and analysed. Internal validation revealed high accuracy across all models, EfficientNetB0 achieved the highest classification accuracy (0.9561; 95% CI 0.9490–0.9630), EfficientNetB3 demonstrated superior overall discriminative performance, achieving the highest AUROC (0.9892; 95% CI 0.9841–0.9934) highest sensitivity (0.9573), and lowest Brier score (0.0168). Meanwhile, DenseNet exhibited the most balanced clinical screening performance, achieving the highest F1-score (0.7259; 95% CI 0.6797–0.7669) and Youden Index (0.2381), indicating improved balance between sensitivity and specificity. In contrast, external validation revealed substantial deterioration in model performance across all architectures, highlighting major limitations in cross-population generalisability. Although EfficientNetB0 achieved the highest external accuracy (0.8821; 95% CI 0.8746–0.8898), AUROC values declined markedly across models (0.5140–0.6104), accompanied by poor sensitivity, reduced F1-scores, and substantial calibration instability. EfficientNetB3 achieved the highest external sensitivity (0.5939), whereas calibration analyses demonstrated unreliable probability estimation under domain-shift conditions. These findings suggest that AI models trained on geographically homogeneous retinal imaging datasets may not generalise reliably across underrepresented populations. Population differences and imaging variability substantially affected external model performance, highlighting the need for diverse datasets, rigorous external validation, and adaptive recalibration before clinical deployment of AI-driven DR screening systems.

Check for updates

Academic Editor: Kohei Arai

Received: 2 February 2026
Revised: 20 May 2026
Accepted: 22 May 2026
Published: 3 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: artificial intelligence; medical image analysis; diabetic retinopathy; deep learning; machine learning; external validation; clinical decision support; generalisability; healthcare AI

1. Introduction

Diabetic retinopathy (DR) is a leading cause of preventable blindness, affecting approximately 283 million of the 463 million individuals living with diabetes mellitus (DM) [1–3]. This burden is expected to increase, with the 11th edition of the International Diabetes Federation (IDF) Atlas projecting that 853 million people will be living with DM by 2050 [4]. Notably, over 40% of affected individuals may remain undiagnosed, particularly in low- and middle-income countries (LMICs), especially in Sub-Saharan Africa (SSA) [2,5]. This growing challenge is driven by demographic and epidemiological transitions, including population ageing, rapid urbanisation, rising obesity rates, reduced physical activity, and socioeconomic barriers that limit access to timely and effective eye care [6].

Early detection and treatment of DR can prevent up to 95% of vision loss associated with the condition [7]. However, systematic screening programmes are often lacking in resource-constrained settings [8]. In SSA, existing screening practices primarily depend on the manual grading of retinal fundus photographs, a labour-intensive process characterised by significant intergrader variability and limited scalability [9]. These challenges exacerbate inequities in eye health, emphasising the urgent need for innovative screening solutions tailored to resource-limited environments [10,11]. Recent advances in artificial intelligence (AI) technologies have transformed ophthalmic care, yielding substantial progress in automated DR screening [12,13]. Deep learning technologies, particularly Convolutional Neural Network (CNN)-based approaches, now achieve diagnostic performance comparable to, and sometimes surpassing, that of specialist ophthalmologists [14–16].

Several autonomous AI systems, including IDx-DR (Digital Diagnostics, Coralville, IA, USA) [16], EyeArt (Eyenuk, Woodland Hills, CA, USA) [17,18], Optomed Aurora with AEYE AI, and AEYE-DS [19], have received US Food and Drug Administration (FDA) clearance to detect more-than-mild and referable DR from fundus photographs without specialist oversight [19,20]. These systems provide point-of-care screening in primary and community health settings, deliver immediate diagnostic results, and reduce dependence on trained graders, but their application remains limited in Africa [21]. Such innovations demonstrate the potential of AI to deliver scalable, cost-effective DR screening, especially in settings with acute ophthalmologist shortages and limited infrastructure [22,23].

Despite advances in AI for DR detection, most models primarily rely on datasets from Europe, China, and the USA, as shown in Figure 1 [24]. This limited representation fails to capture the diversity of the global population, restricting the models' generalisability [24,25]. Consequently, these models can introduce biases, lead to inconsistent outcomes, produce ineffective solutions, and exacerbate social injustices in AI decision-making. When biased decisions affect minority populations, they raise complex ethical and legal questions. For instance, if a model misclassifies cases of DR, it becomes unclear who is responsible: the model developers, healthcare providers, or the institutions that use the model. Even models that achieve 95% accuracy, often deemed acceptable, can still misclassify 5% of patients' disease, potentially resulting in vision-threatening conditions. To address these challenges, rigorous external validation in localised contexts is urgently needed, alongside regulatory oversight and ongoing monitoring across diverse populations [25,26]. Testing models on new, unseen population data not included in the training process is essential for assessing their generalisability and reliability [27]. This approach helps to identify

any biases and confirms that the models perform well across varied demographics in multiple settings. Moreover, external validation enhances stakeholder confidence in model predictions, making it a critical step in their deployment, particularly in healthcare, where accurate predictions are vital [27,28]. Therefore, without comprehensive external validation, biases can persist and further entrench health inequities.

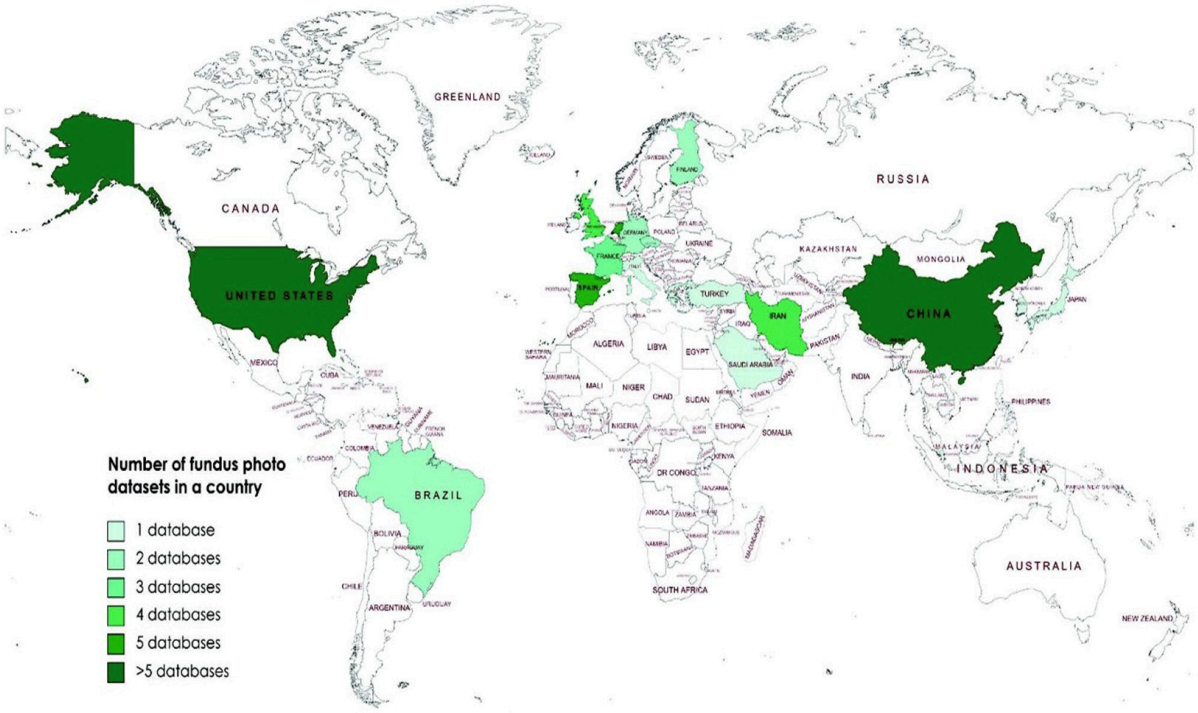


Figure 1. Distribution of Fundus Image Datasets by Country. Adapted from [24].

This study systematically validates multiple state-of-the-art CNN architectures—namely, ConvNeXt, Inception-v3, DenseNet-121, ResNet-50, EfficientNet-B0, and EfficientNet-B3—for the detection of DR using retinal datasets from Latin America and SSA. We specifically focused on detecting referable diabetic retinopathy (rDR) by framing the prediction task as a binary classification problem, categorising ICDR grades 0–1 as non-referable DR and grades 2–4 as referable DR. To enhance traditional performance evaluations, we incorporated calibration analysis, subgroup assessments, and studies on imaging device variability, thereby ascertaining the clinical reliability and transferability of DR detection models in resource-constrained settings. Our objectives include (1) assessing diagnostic performance through comprehensive analyses of accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC), alongside F1 score and predictive value metrics; (2) evaluating clinical utility via decision curve analysis and model calibration to illuminate the practical implications of integrating artificial intelligence into DR screening; (3) investigating performance consistency across diverse demographic subgroups, imaging devices, and varying image quality levels, using calibration metrics to evaluate alignment between predicted probabilities and actual outcomes; and (4) identifying the strengths and limitations of each model to facilitate the selection of AI solutions tailored for resource-limited environments, with a focus on prediction calibration and real-world healthcare applicability.

2. Materials and Methods

2.1. Study Design and Setting

We conducted a retrospective, cross-sectional diagnostic accuracy study to develop and externally validate CNN models for automated DR classification using two independent

retinal image datasets: the Brazilian Multilabel Ophthalmological Dataset (BrSET) [29] from Latin America, and the Multimodal Database of Retina Images for Africa (MoDRIA) from Uganda and Kenya [30].

2.2. Datasets

The Brazilian Multilabel Ophthalmological Dataset (BrSET) comprises 16,266 macula-centred fundus images from 8524 patients at three tertiary ophthalmology centres in São Paulo, Brazil, collected between 2010 and 2020. Images were acquired using standardised fundus cameras (Nikon NF-505 fundus camera (Nikon Corporation, Tokyo, Japan) and the Canon CR-2 retinal camera (Canon Inc., Tokyo, Japan)) with pharmacological mydriasis to enhance image quality. All procedures for image collection, de-identification, and classification adhered to institutional standards, with trained non-medical personnel capturing images under the supervision of certified retinal specialists. Each image was annotated for anatomical features and categorised into 5 classifications (0: No retinopathy, 1: Mild non-proliferative diabetic retinopathy, 2: Moderate non-proliferative diabetic retinopathy, 3: Severe non-proliferative diabetic retinopathy, or 4: Proliferative diabetic retinopathy and post-laser status), with participants providing one macula-centred image from each eye to ensure comprehensive representation of retinal pathologies. Although retinal images in the BrSET dataset were originally graded using the five-level ICDR classification system, the prediction task in this study was formulated as a binary classification problem, where ICDR grades 0–1 were categorised as non-referable DR and ICDR grades 2–4 as referable DR.

Quality control assessments evaluated focus, illumination, completeness of the field, and the presence of artefacts, categorising images as satisfactory or unsatisfactory. Pathological classifications included diabetic retinopathy, diabetic macular oedema, and other retinal conditions, with severity graded according to the International Clinical Diabetic Retinopathy (ICDR) classification and the Scottish Diabetic Retinopathy Grading Scheme (SDRG), standardising the determination of DR severity. Ethical approval for the study was obtained from the São Paulo Federal University Institutional Review Board (CAAE: 33842220.7.0000.5505), and informed consent was acquired from all participants.

The MoDRIA served as the external validation cohort. Developed by the Mbarara University Data Science Research Hub with funding from the NIH (1U54TW012043-01), a total of 5867 out of 37,725 well-labelled colour fundus images from the MoDRIA dataset were utilised for external validation. These images were collected from 2933 participants across 14 clinical sites in Uganda and two in Kenya. The data collection process spanned from 2021 to 2025, resulting in a comprehensive dataset to support research and analysis.

This dataset utilised non-mydriatic fundus cameras to reflect real-world clinical practices, employing the Forus 3nethra Classic retinal camera (Forus Health Pvt. Ltd., Bengaluru, Karnataka, India) in Uganda and the Canon CR-2 in Kenya. Each participant provided paired macula-centred and disc-centred images, capturing a diverse range of retinal pathologies.

Demographic and clinical metadata were systematically recorded, including age, sex, blood pressure, glucose levels, HIV status, diabetes history, insulin use, and visual acuity. All images were de-identified and reviewed by retinal specialists using the ICDR for standardised DR grading. The MoDRIA dataset intentionally included both diabetic and non-diabetic participants, with an estimated diabetes prevalence of 28%, enabling simultaneous AI model development and validation for detecting rDR and evaluating non-DR pathologies.

Missing demographic and clinical variables were not used as predictive inputs for CNN training, and therefore did not require imputation during model development. Analyses involving demographic subgroup comparisons were performed using complete-case

data only. Images with incomplete metadata but adequate retinal quality remained eligible for image-based model evaluation.

Ethical approval for MoDRIA was granted by the Mbarara University of Science and Technology Research Ethics Committee (MUST-2021-239) and the Uganda National Council for Science and Technology (HS2094ES). Quality control metrics were employed to assess image quality using QuickQual MEga Minified Estimator (MEME) version 1 software and fractal dimension analysis, excluding images deemed inadequate for clinical interpretation. Images were excluded when severe blur, poor illumination, incomplete retinal fields, motion artefacts, or obscured anatomical structures prevented reliable clinical grading. Quality control procedures were applied before model training and evaluation to minimise label uncertainty and reduce preprocessing-related bias. Both the BrSET and MoDRIA datasets exhibit remarkable diversity, comprising extensive image collections from varied demographics with robust age and gender representation. They provide in-depth visual acuity data, health contexts, and systemic health factors, ensuring high-quality imaging and encompassing multiple classification parameters for thorough diabetic retinopathy detection, as shown in Table 1.

2.3. Rationale for Dataset Selection

Selecting the BrSET and MoDRIA datasets is critical for developing and validating AI-based models to detect DR. The BrSET dataset serves as a substantial training resource due to its large size and diverse population from Latin America, which enables the model to learn complex patterns and relationships inherent to retinal pathologies. It reflects a mature population with an average age of 57.66 years (SD 18.26) and offers a diverse representation of retinal conditions collected under standardised conditions. This comprehensive dataset is particularly advantageous for training and internal validation, facilitating a robust training phase and effective performance assessment. In contrast, the MoDRIA dataset presents participants with an average age of 42.82 years (SD 16.70). This dataset highlights demographic diversity and addresses the unique clinical challenges of DR in SSA. Notably, there are significant differences in DR prevalence: 93.51% of BrSET patients display no retinopathy compared to 91.97% in MoDRIA, with higher rates of mild (1.00% vs. 0.36%) and moderate (1.91% vs. 1.01%) non-proliferative DR in BrSET, while MoDRIA exhibits a higher prevalence of proliferative DR (6.55% vs. 2.42%).

MoDRIA's use of non-mydriatic cameras, aligned with clinical practices in resource-limited settings, enhances the practical applicability of AI technologies. Its large sample size allows for reliable estimation of key performance metrics such as accuracy and sensitivity, ensuring statistical significance in the findings. The methodological and clinical variances represented in MoDRIA introduce essential variability for external validation. Overall, the strategic selection of the BrSET and MoDRIA datasets is scientifically justified in Table 2, as their complementary features enhance the development and validation of AI models for DR detection.

2.4. Model Architectures

We employed six (6) CNN architectures, summarised in Table 3, to enhance the diagnostic performance of DR detection. The selected models, ConvNeXt, Inception-v3, DenseNet-121, ResNet-50, EfficientNet-B0, and EfficientNet-B3, each represent distinctive methodologies for feature extraction and model scaling. These models were chosen for their demonstrated success in medical imaging tasks and their architectural diversity, thereby ensuring a comprehensive evaluation of feature representation and network efficiency strategies.

Table 1. Summary of training and external validation dataset.

Characteristic	Category	Training and Internal Validation Dataset (BrSET)	External Validation Dataset (MoDRIA)
General	Number of images	16,266	37,725
	Number of participants	8524	9438
	Race/ethnicity	Brazilian	Black (purely African population)
	Camera	Canon CR; NIKON NF5050	3nethra Classic; Canon
Age (years)	≤25	417 (4.9%)	1797 (19.0%)
	26–50	1281 (15.0%)	4617 (48.9%)
	51–75	3029 (35.5%)	2506 (26.6%)
	≥76	926 (10.8%)	373 (4.0%)
	Missing	2882 (33.8%)	145 (1.5%)
	Total	8535 (100%)	9438 (100%)
Gender	Female	3300 (38.6%)	5768 (61.1%)
	Male	5243 (61.4%)	3629 (38.5%)
	Missing	0 (0%)	41 (0.4%)
	Total	8543 (100%)	9438 (100%)
Visual acuity (VA)	Normal (≥6/12)	–	8066 (85.5%)
	Mild impairment (<6/12)	–	635 (6.7%)
	Moderate impairment (<6/18)	–	479 (5.1%)
	Severe impairment (<6/60)	–	8 (0.1%)
	Blindness (<3/60)	–	123 (1.3%)
	Missing	–	128 (1.3%)
	Total	–	9439
Medication use	No medication	–	7310 (77.4%)
	HIV only	–	933 (9.9%)
	Hypertension only	–	560 (5.9%)
	Diabetes only	–	302 (3.2%)
	Diabetes + Hypertension	–	189 (2.0%)
	HIV + Hypertension	–	57 (0.6%)
	Diabetes + HIV	–	53 (0.6%)
	CVD only	–	15 (0.2%)
	Diabetes + HIV + Hypertension	–	14 (0.1%)
	Missing	–	5 (0.05%)
	Total	–	9438 (100%)
Systemic risk factors (missing %)	Weight	–	7.3
	Height	–	25.1
	Body mass index	–	25.4
	Systolic BP	–	3.3
	Diastolic BP	–	7.3
	Random blood sugar	–	71.5
Dataset characteristics	DR classification	ICDR	DR_SDRG
	Quality parameters	Focus; illumination	Focus; illumination
	Classification parameters	DR, macular oedema, scar, nevus, AMD, vascular occlusion, hypertensive retinopathy, drusens, haemorrhage, retinal detachment, myopic fundus, increased cup–disc ratio	DR, macular oedema, optic disc ratio
	Anatomical parameters	Optic disc, vessels, macula	Optic disc, vessels, macula

Table 2. Comparison of Training and Validation Datasets with Statistical Significance.

Variable	Train (BrSET)	Validation (MoDRIA)	*p*-Value
Patient age (mean ± SD)	57.66 ± 18.26	42.82 ± 16.70	0.0000
DR_ICDR 0 (No retinopathy)	15,210 (93.51%)	5108 (91.97%)	0.0001
DR_ICDR 1 (Mild non-proliferative DR)	162 (1.00%)	20 (0.36%)	0.0000
DR_ICDR 2 (Moderate non-proliferative DR)	310 (1.91%)	56 (1.01%)	0.0000
DR_ICDR 3 (Severe non-proliferative DR)	190 (1.17%)	6 (0.11%)	0.0000
DR_ICDR 4 (Proliferative DR)	394 (2.42%)	364 (6.55%)	0.0000

Table 3. Architectural traits and preprocessing requirements.

Model	Key Architectural Traits	Input and Preprocessing
ConvNeXt	Modern CNN architecture with ConvNeXt blocks, large 7 × 7 depthwise convolutions, Layer Normalisation	224 × 224 pixels, ImageNet mean/std normalisation
Inception-v3	Inception modules with parallel convolution branches, Batch Normalisation, multi-scale feature extraction	299 × 299 pixels, rescaled by $1/127.5 - 1$
DenseNet-121	Dense connectivity where each layer receives features from all previous layers (feature concatenation)	224 × 224 pixels, rescaled by 1/255
ResNet-50	50-layer residual network with bottleneck residual blocks and skip connections	224 × 224 pixels, rescaled by 1/255
EfficientNet-B0	Mobile inverted bottleneck convolution (MBConv) blocks, squeeze-and-excitation modules, compound scaling	224 × 224 pixels, ImageNet mean/std normalisation
EfficientNet-B3	utilises a novel compound scaling method that uniformly scales the network's depth, width, and image resolution, leading to optimised efficiency and accuracy	300 × 300 resolution

ConvNeXt leverages depthwise convolutions and layer normalisation to achieve superior feature representation, making it particularly effective for capturing intricate patterns in the data. This architecture is designed to improve the training efficiency and accuracy of deep networks while ensuring that the model generalises well to unseen data. During the Input and Preprocessing stage, images were resized to 224 × 224 pixels and normalised using ImageNet mean and standard deviation to ensure consistency across the dataset, which aids in faster convergence during training.

Inception-v3 utilises parallel convolutional branches within its inception modules, enabling robust multi-scale feature extraction that enhances its ability to detect variations in retinal images. This model's architecture enables it to capture features at multiple scales and resolutions, providing a rich representation of the input data. During the Input and Preprocessing stage, input images are resized to 299 × 299 pixels and rescaled using the formula (1/127.5)—1 to normalise the pixel values, which helps to stabilise the learning process.

DenseNet-121 emphasises dense connectivity, allowing each layer to receive inputs from all previous layers, facilitating efficient gradient flow and maximising feature reuse. This connectivity pattern helps alleviate the vanishing gradient problem, making it easier to train deeper networks. During the Input and Preprocessing stage, images were resized to 224 × 224 pixels and rescaled by 1/255 to normalise pixel values, ensuring that the model receives data in a consistent format.

ResNet-50 employs residual connections, which are instrumental in training deeper networks by mitigating the vanishing gradient problem. This architecture enables gradients to flow through the network more effectively, thereby improving learning. The input images are resized to 224 × 224 pixels and also rescaled by 1/255 for normalisation, preparing the data for optimal performance.

The EfficientNet models, specifically EfficientNet-B0 and EfficientNet-B3, adopt a compound scaling strategy that optimally balances network depth, width, and image resolution, thereby improving accuracy and computational efficiency. EfficientNet-B0 serves as the baseline model, while EfficientNet-B3 builds on it with deeper and wider architectures for improved performance. Input and Preprocessing: For EfficientNet-B0, images are resized to 224 × 224 pixels and normalised using ImageNet mean and standard deviation. EfficientNet-B3 requires images to be resized to 300 × 300 pixels, allowing for finer detail capture due to the increased resolution.

To further enhance model generalisation and reduce the risk of overfitting, data augmentation techniques such as random rotation, horizontal and vertical flipping, and colour jittering were applied during training for all model architectures. Additionally, regularisation methods, including dropout and weight decay, were used across all models to further improve robustness.

2.5. Model Development and Setup

We developed a comprehensive pipeline for training the six CNN models (ConvNeXt, Inception-v3, DenseNet-121, ResNet-50, EfficientNet-B0, and EfficientNet-B3), ensuring that each model was optimised according to its unique architecture. These architectures were selected for their strong track records in medical imaging tasks and their complementary design philosophies, allowing us to compare their performance across both classic and more recent CNN innovations. Hyperparameters were selected through pilot experiments and iterative optimisation based on validation AUROC and calibration performance. All models were trained using the AdamW optimiser with an initial learning rate of 1×10^{-4}, cosine annealing learning rate scheduling (T_max = 100; η_min = 1×10^{-6}), batch size of 64, dropout rate of 0.3, and L2 regularisation of 1×10^{-4}. Early stopping with a patience threshold of 10 epochs was implemented to minimise overfitting and preserve model generalisability.

The prediction task was formulated as a binary classification problem for rDR, where ICDR grades 0–1 were classified as non-referable DR and ICDR grades 2–4 as referable DR. To address substantial class imbalance, model optimisation employed a weighted binary cross-entropy (BCE) loss function combined with focal loss regularisation. Weighted BCE was used to compensate for unequal class frequencies, while focal loss enhanced learning from difficult minority-class examples by down-weighting easily classified samples. In addition, Class-Specific Discrepancy Reduction (CSDR; λ = 2.0) was incorporated as an auxiliary regularisation mechanism to reduce class-specific prediction bias and improve calibration stability across subgroups.

Model selection was based on mean validation AUROC, calibration performance (Brier score and Expected Calibration Error), and stability across cross-validation folds. Preprocessing and augmentation were implemented using the Albumentations library and included random rotations, horizontal and vertical flips, brightness and contrast adjustments, Gaussian noise, and blur transformations to improve robustness to imaging variability.

A stratified 5-fold cross-validation strategy was employed to preserve the proportion of referable and non-referable DR cases within each training and test splits. Stratification was implemented to minimise sampling bias arising from class imbalance and to ensure robust estimation of model generalisability across heterogeneous data partitions. Final model selec-

tion was based on aggregate cross-validation performance, prioritising AUROC, calibration quality, sensitivity, and inter-fold consistency rather than single-fold accuracy alone.

Importantly, the external validation cohort (MoDRIA) remained entirely independent throughout model development, hyperparameter optimisation, threshold selection, and calibration procedures. No images or metadata from the external dataset were used during training, internal validation, or model selection, thereby minimising information leakage and ensuring an unbiased assessment of model generalisability.

2.6. Model Evaluation

2.6.1. Model Training and Internal Validation

Following model training and internal validation utilising the BrSET dataset, the final model was derived through a comprehensive and rigorous development process incorporating bootstrapping and k-fold cross-validation presented in Figure 2. The dataset was systematically partitioned into K non-overlapping folds, with iterative training conducted on $K - 1$ folds while reserving one fold for validation. This structured approach facilitated a robust assessment of model performance across diverse data splits, effectively mitigating the risk of data leakage.

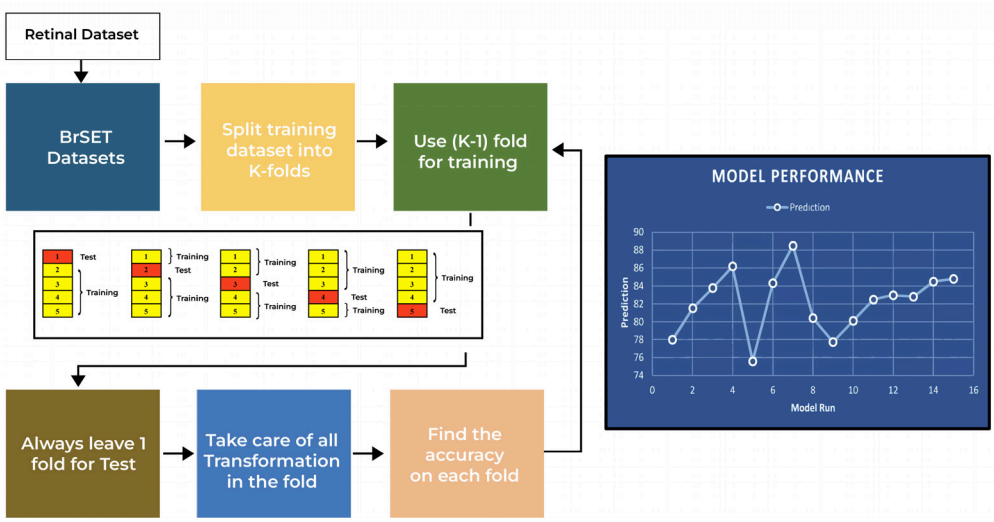


Figure 2. K-fold cross-validation.

To enhance the variability of the training data, bootstrap sampling was employed within each fold, thereby improving the model's stability and generalisation capabilities. Performance metrics, including AUC-ROC, accuracy (with 95% confidence intervals), precision, recall, F1-score, and specificity, were calculated for each fold and subsequently aggregated to identify optimal model configurations and hyperparameters. Classification thresholds were determined using Youden's J statistic on internal validation folds and subsequently fixed during external validation to avoid threshold optimisation on the external dataset.

In addition to these primary metrics, we utilised supplementary evaluation criteria such as Youden's J index, Brier Score, Expected Calibration Error (ECE), and confusion matrices to comprehensively assess classification performance, calibration, and error distribution. The consistency of performance, convergence behaviour, and minimal inter-fold variability observed throughout the cross-validation process provided strong evidence for the robustness of the model.

The final model was subsequently obtained by retraining the chosen architecture on the complete BrSET dataset, utilising the optimal hyperparameters identified during the validation phase. This method ensured that the final model effectively leveraged

all available data while maintaining the generalisation properties validated by the cross-validation framework.

2.6.2. Model External Validation

The model underwent external validation using the MoDRIA dataset, which represents a distinct population from the training data, facilitating a comprehensive assessment of its generalisability and performance in novel contexts. The deliberate selection of the MoDRIA dataset ensured the inclusion of diverse demographics and scenarios not represented in the BrSET dataset, thereby rigorously testing the model's robustness across a broader range of conditions. During the validation process, the trained model was applied to the Modria data to generate predictions, which were subsequently compared against the actual outcomes to evaluate performance.

This study was designed and reported with consideration of key recommendations from the STARD-AI extension [31] and CONSORT-AI reporting guidelines [32], particularly regarding external validation, transparent reporting of diagnostic performance, calibration assessment, and model generalisability. These standards ensured the transparency and reproducibility of our findings, thereby enhancing the reliability and credibility of the research. This systematic approach not only reinforces the model's validity but also enhances its applicability in real-world settings, emphasising the importance of robust external validation in the development of AI-driven diagnostic tools.

3. Results

3.1. Internal Validation

We internally validated the selected architectural performances, detailing significant metrics for each of the models, as presented in Table 4.

Table 4. Summarised internal validation performance metrics for each model.

Metric	ConvNeXt	Dense Net	Efficient NetB0	ResNet50	Inception-V3	Efficient NetB3
Youden	0.1549	0.2381	0.2160	0.1890	0.1773	0.1456
Accuracy	0.9457 (95% CI 0.9374–0.9530)	0.9555 (95% CI 0.9481–0.9621)	0.9561 (95% CI 0.9490–0.9630)	0.9524 (95% CI 0.9451–0.9607)	0.9389 (95% CI 0.9303–0.9478)	0.9481 (95% CI 0.9398–0.9564)
F1-score	0.6954 (95% CI 0.6510–0.7343)	0.7259 (95% CI 0.6797–0.7669)	0.7202 (95% CI 0.6751–0.7632)	0.7114 (95% CI 0.6690–0.7548)	0.6678 (95% CI 0.6234–0.7130)	0.7051 (95% CI 0.6582–0.7479)
AUROC	0.9881 (95% CI 0.9804–0.9936)	0.9816 (95% CI 0.9717–0.9897)	0.9758 (95% CI 0.9667–0.9843)	0.9807 (95% CI 0.9726–0.9883)	0.9840 (95% CI 0.9761–0.9912)	0.9892 (95% CI 0.9841–0.9934)
Sensitivity	0.9573	0.9100	0.8720	0.9052	0.9479	0.9573
Specificity	0.9449	0.9586	0.9619	0.9557	0.9383	0.9475
Brier Score	0.0173	0.0195	0.0209	0.0196	0.0191	0.0168
ECE	0.8870	0.8849	0.8908	0.8867	0.8817	0.9002
Internal calibration best fit slope:	1.0832	1.1783	1.1324	1.1280	1.1513	1.1700
Internal calibration best fit intercept:	−0.1204	−0.1274	−0.1143	−0.1090	−0.1355	−0.1095

To determine thresholds, we used the Youden's index. Notably, these values varied across the models: ConvNeXt 0.1549, DenseNet 0.2381, EfficientNetB0 0.2160, ResNet50 0.1890, Inception-V3 0.1773, and EfficientNetB3 0.1456.

In terms of accuracy, EfficientNetB0 exhibited the highest performance at 0.9561 (95% CI 0.9490–0.9630), closely followed by DenseNet at 0.9555 (95% CI 0.9481–0.9621) and ResNet50 at 0.9524 (95% CI 0.9451–0.9607). EfficientNetB3 also demonstrated a commendable accuracy of 0.9481 (95% CI 0.9398–0.9564), ConvNeXt achieved 0.9457 (95% CI 0.9374–0.9530), while Inception-V3 recorded the lowest accuracy at 0.9389 (95% CI 0.9303–0.9478).

The F1-scores indicated that DenseNet led with a score of 0.7259 (95% CI 0.6797–0.7669), followed by EfficientNetB0 at 0.7202 (95% CI 0.6751–0.7632), ResNet50 at 0.7114 (95% CI 0.6690–0.7548), EfficientNetB3 at 0.7051 (95% CI 0.6582–0.7479), ConvNeXt at 0.6954 (95% CI 0.6510–0.7343), and Inception-V3 at 0.6678 (95% CI 0.6234–0.7130). The AUROC for each model demonstrated excellent discriminative performance, with ConvNeXt achieving an AUROC of 0.9881 (95% CI 0.9804–0.9936) and EfficientNetB3 the highest at 0.9892 (95% CI 0.9841–0.9934). DenseNet and ResNet50 followed with AUROCs of 0.9816 (95% CI 0.9717–0.9897) and 0.9807 (95% CI 0.9726–0.9883), respectively. EfficientNetB0 and Inception-V3 had AUROCs of 0.9758 (95% CI 0.9667–0.9843) and 0.9840 (95% CI 0.9761–0.9912) as presented in Table 5.

Table 5. Internal Validation of ROC and Calibration Curves.

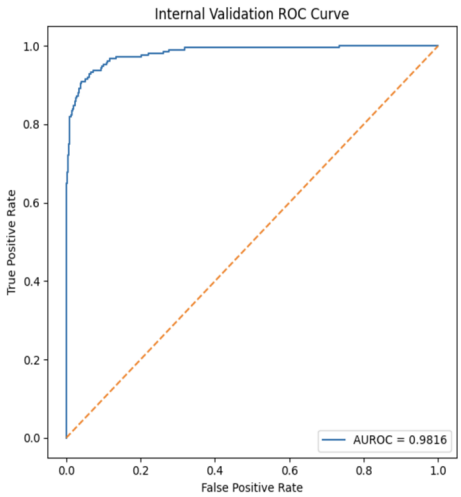
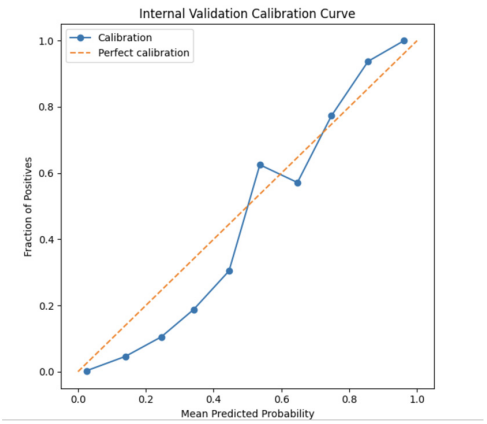
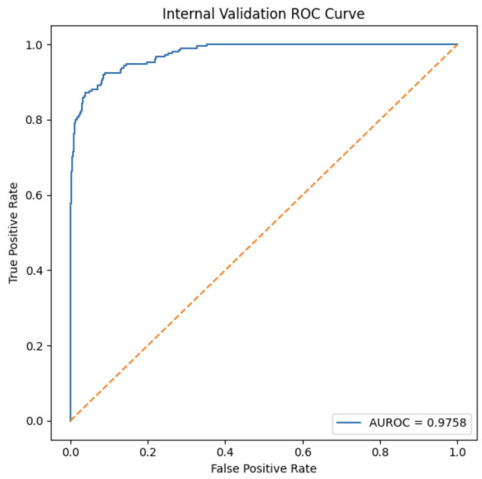
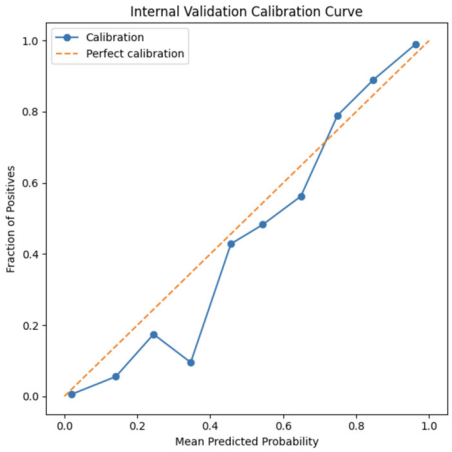
DenseNet	**DenseNet**
	
EfficientB0	**EfficientB0**
	

Table 5. *Cont.*

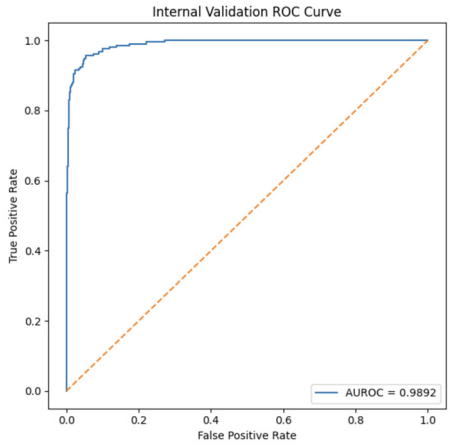
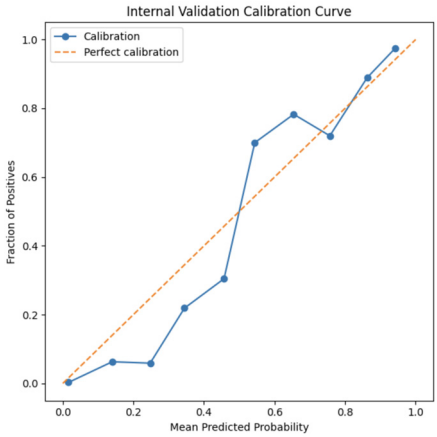
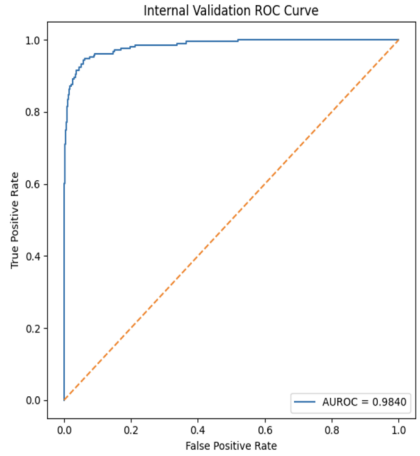
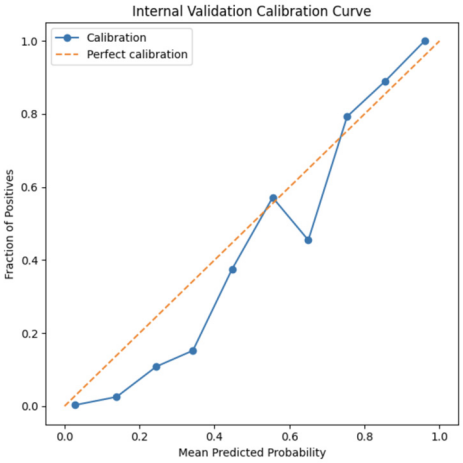
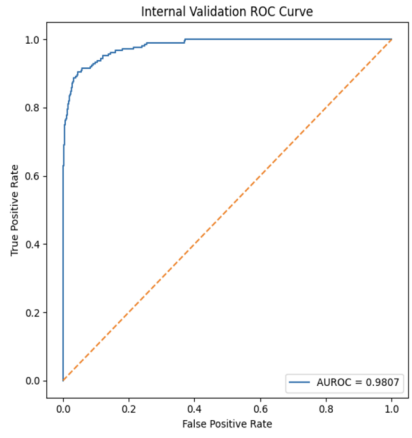
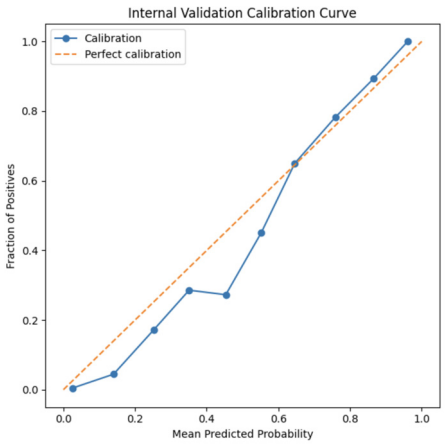
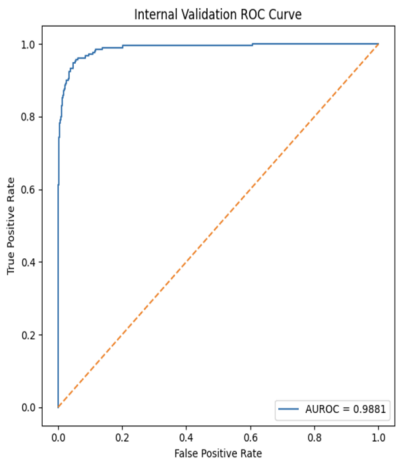
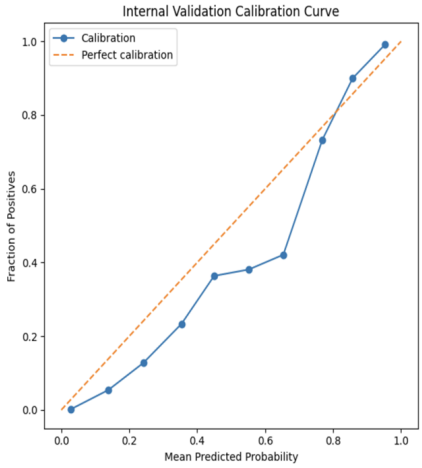
EfficientB3	**EfficientB3**
	
InceptionV3	**InceptionV3**
	
ResNet-50	**ResNet-50**
	

Table 5. *Cont.*

ConvNext	**ConvNext**
	

Sensitivity and specificity metrics varied across models, with ConvNeXt and EfficientNetB3 achieving the highest sensitivity at 0.9573. In contrast, the models demonstrated varying specificity, with DenseNet showing the highest value at 0.9586, closely followed by EfficientNetB0 at 0.9619.

Calibration performance was assessed using the Brier score, calibration slope, calibration intercept, and calibration curves. Internal Brier scores were low across architectures, ranging from 0.0168 to 0.0209, suggesting relatively low average probabilistic error within the internal validation dataset. The Calibration slopes ranged from 1.0832 to 1.1783, and the intercepts ranged from −0.1355 to −0.1090, indicating modest miscalibration rather than perfect calibration.

Additionally, the internal calibration metrics revealed the best-fit slope and intercept for each model. The best fit slope values ranged from 1.0832 for ConvNeXt to 1.1783 for DenseNet, indicating the degree of calibration. The best-fit intercept values ranged from −0.1204 for ConvNeXt to −0.1355 for Inception-V3, reflecting each model's calibration bias. Although calibration slopes and intercepts suggested only modest systematic calibration deviation, the relatively high ECE values indicate that probability estimates remained imperfectly calibrated despite strong discriminative performance. These findings suggest that the models may still exhibit overconfident probability estimates even under internal validation conditions. The confusion matrices for each model further elucidated their performance, showing true positives, false positives, true negatives, and false negatives across the datasets, as detailed in Table 4. In summary, the internal validation findings demonstrate strong overall discriminative capability across all evaluated CNN architectures, with DenseNet and EfficientNetB3 consistently achieving superior performance across multiple evaluation metrics. Nevertheless, the observed calibration imperfections reinforce the importance of incorporating calibration assessment and potential post hoc recalibration strategies before clinical deployment of AI-assisted diabetic retinopathy screening systems.

3.2. External Validation

Using the Modria dataset, we conducted external validation of all our model architectures, including ConvNeXt, DenseNet, EfficientNetB0, ResNet50, Inception-V3, and EfficientNetB3. The performance metrics for each model are summarised in Table 6.

Table 6. Summarised external performance metrics for each model.

Metric	ConvNeXt	Dense Net	Efficient NetB0	ResNet50	Inception-V3	Efficient NetB3
Youden	0.1549	0.2381	0.2160	0.1890	0.1773	0.1456
Accuracy	0.5142 (95% CI 0.5011–0.5245)	0.7593 (95% CI 0.7480–0.7699)	0.8821 (95% CI 0.8746–0.8898)	0.8772 (95% CI 0.8692–0.8858)	0.7547 (95% CI 0.7429–0.7664)	0.5013 (95% CI 0.4884–0.5138)
F1-score	0.1350 (95% CI 0.1171–0.1497)	0.1461 (95% CI 0.1228–0.1704)	0.1571 (95% CI 0.1246–0.1848)	0.1160 (95% CI 0.0846–0.1452)	0.1568 (95% CI 0.1312–0.1794)	0.1483 (95% CI 0.1318–0.1638)
AUROC	0.5425 (95% CI 0.5136–0.5721)	0.5602 (95% CI 0.5302–0.5891)	0.6104 (95% CI 0.5806–0.6380)	0.5140 (95% CI 0.4803–0.5443)	0.5815 (95% CI 0.5496–0.6113)	0.5734 (95% CI 0.5445–0.6041)
Sensitivity	0.5188	0.2817	0.1502	0.1103	0.3122	0.5939
Specificity	0.5138	0.7970	0.9398	0.9376	0.7896	0.4940
Brier Score	0.0882	0.0868	0.0689	0.0712	0.0778	0.1048
ECE	0.7546	0.7685	0.8509	0.8569	0.8013	0.7485
Internal calibration best fit slope:	0.7456	0.6926	0.8968	0.6479	0.8214	0.4979
Internal calibration best fit intercept:	−0.1278	−0.1160	−0.0987	−0.0505	−0.1336	−0.0726

The models' accuracy scores during external validation were generally lower than during internal validation. EfficientNetB0 achieved the highest accuracy at 0.8821 (95% CI 0.8746–0.8898), followed closely by ResNet50 at 0.8772 (95% CI 0.8692–0.8858) and DenseNet at 0.7593 (95% CI 0.7480–0.7699). In contrast, ConvNeXt declined to 0.5142 (95% CI 0.5011–0.5245), while EfficientNetB3 recorded the lowest accuracy at 0.5013 (95% CI 0.4884–0.5138).

The discrepancy between accuracy and other performance metrics during external validation highlights the substantial class imbalance in the MoDRIA dataset, where non-referable DR cases predominated. In this context, models may achieve seemingly high accuracy by predominantly predicting the majority class, which compromises the identification of referable DR cases. Consequently, metrics such as sensitivity, F1-score, AUROC, and calibration measures provide more clinically relevant assessments of model utility than accuracy alone. For instance, despite external accuracies exceeding 0.87 for EfficientNetB0 and ResNet50, their poor sensitivity (0.1502 and 0.1103, respectively) and low F1-scores indicate limited effectiveness in detecting clinically significant referable DR cases.

Throughout model training, the F1-scores showed an overall decline, with EfficientNetB3 achieving the highest value at 0.1571 (95% CI 0.1246–0.1848). The AUROC scores followed a similar pattern, with EfficientNetB3 attaining the highest score of 0.6104 (95% CI 0.5806–0.6380) and ConvNeXt attaining the lowest at 0.5425 (95% CI 0.5136–0.5721). Sensitivity and specificity metrics revealed variable performance across models. ConvNeXt exhibited a sensitivity of 0.5188, while DenseNet and EfficientNetB0 excelled in specificity, achieving scores of 0.7970 and 0.9398, respectively.

Confusion matrix analysis further demonstrated that several architectures disproportionately favoured true-negative predictions at the expense of true-positive detection, resulting in reduced sensitivity despite relatively preserved specificity. This pattern suggests that threshold optimisation and probability calibration substantially influenced model behaviour during external validation. Although thresholds were selected using Youden's index during internal validation, they may not have remained optimal with the demographic and disease prevalence shifts observed in the external cohort. The marked decline

in AUROC values towards near-random performance for some architectures further indicates reduced discriminative capacity under domain-shift conditions, likely driven by differences in imaging devices, population characteristics, disease severity distributions, and acquisition protocols across datasets.

Calibration metrics further highlighted the challenges faced during external validation Table 7. Brier Scores ranged from 0.0689 to 0.1048, and expected calibration error (ECE) values varied from 0.7485 to 0.8569, indicating inferior calibration compared to internal validation metrics. The decline in model performance when applied to external datasets can be attributed to several interrelated factors, emphasising the importance of population-specific validation. Variability in demographic factors, including age, gender, and clinical backgrounds, significantly affects model performance, particularly when the training dataset does not reflect the intended application context. For instance, the mean age of patients in the training dataset (BrSET) was 57.66 years, compared to 42.82 years in the validation dataset (MoDRIA; p-value = 0.0000). This discrepancy suggests the model may struggle to generalise across age groups, potentially skewing estimates of disease prevalence and comorbid conditions. These findings emphasise the need to interpret performance metrics in the context of class imbalance and clinical screening priorities.

Table 7. External Validation of ROC and Calibration Curves.

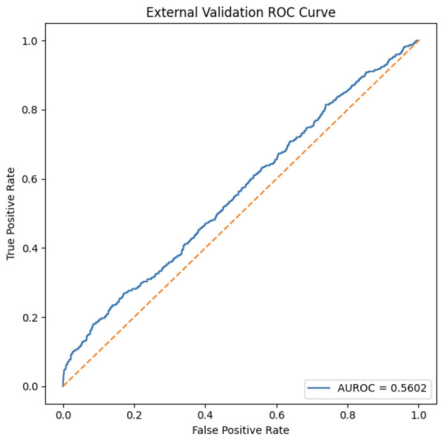
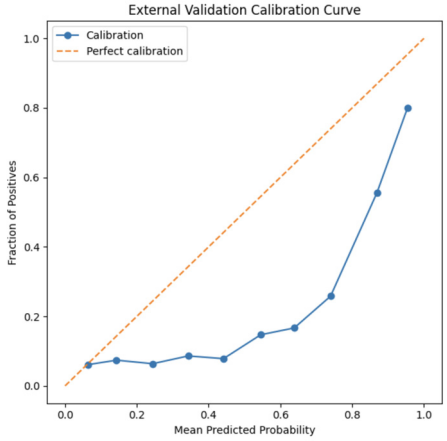
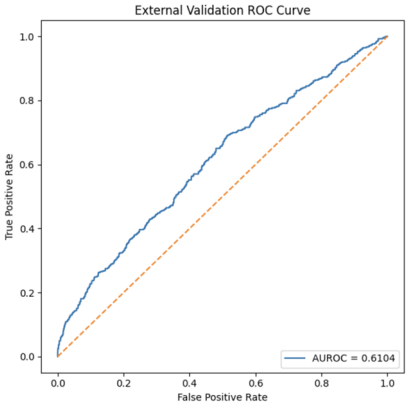
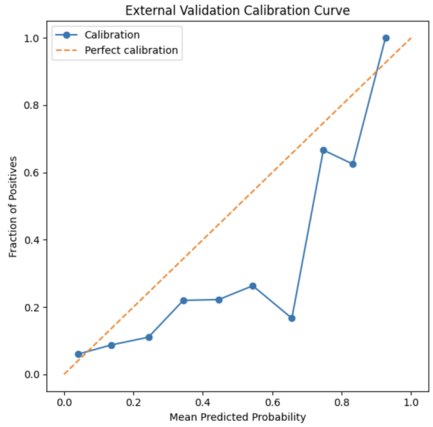
DenseNet	**DenseNet**
	
EfficientB0	**EfficientB0**
	

Table 7. *Cont.*

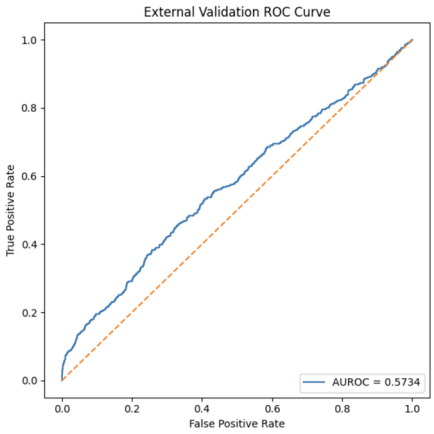
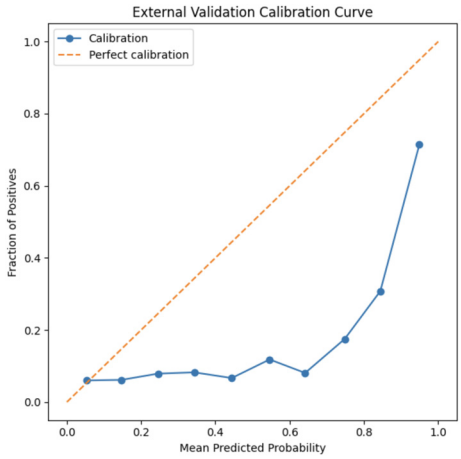
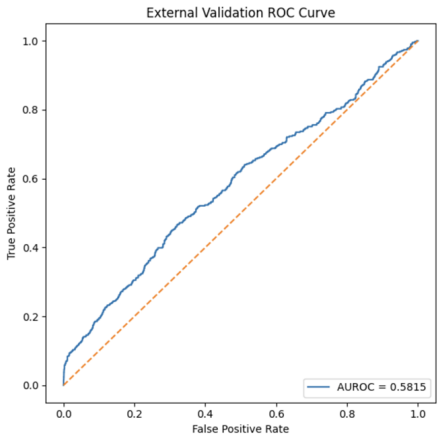
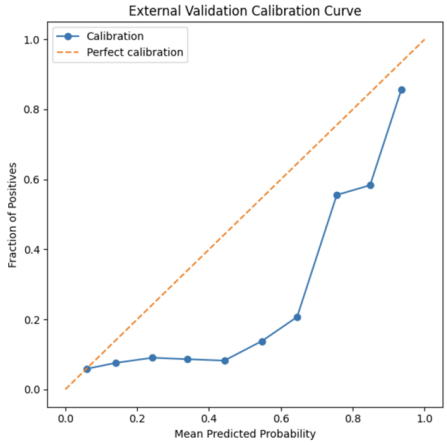
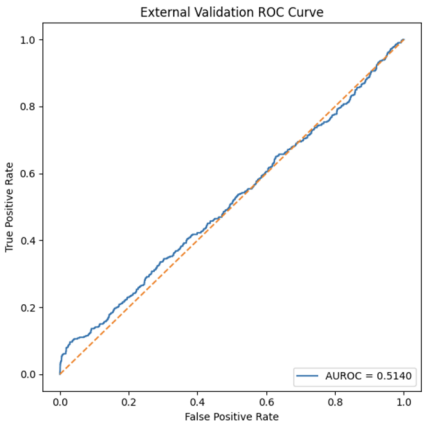
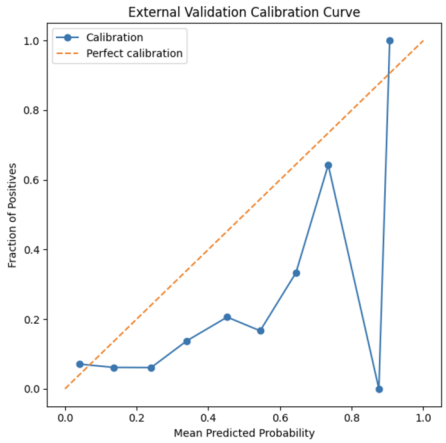
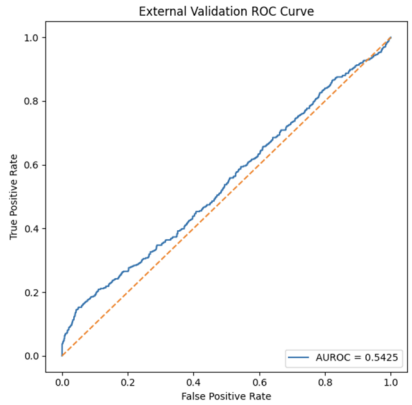
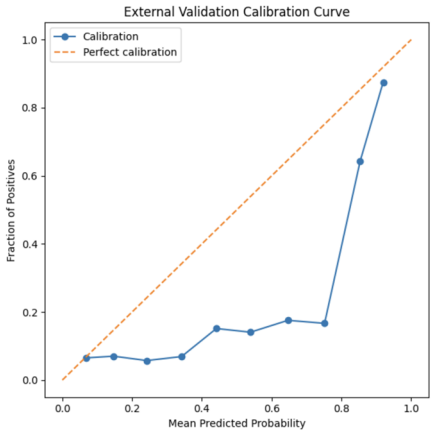
EfficientB3	**EfficientB3**
	
InceptionV3	**InceptionV3**
	
ResNet-50	**ResNet-50**
	

Table 7. *Cont.*

ConvNext	ConvNext
	

Additionally, covariate shifts, arising from differences in feature distributions between the training and validation datasets, can further compromise model accuracy. Table 2 also illustrates critical variances in the distribution of DR severity, where the proportion of patients without retinopathy (DR_ICDR 0) is notably higher in the training set (93.51%) compared to the validation set (91.97%; p-value = 0.0001), while the rate of severe cases (DR_ICDR 4) escalates from 2.42% to 6.55% (p-value = 0.0000). Such discrepancies highlight the potential for the model, primarily trained on distinct demographic and clinical profiles, to inadequately capture the full spectrum of presentations encountered in clinical practice. Therefore, rigorous testing within the specific populations for which models are intended is essential for ensuring their accuracy, relevance, and ethical application, ultimately leading to enhanced patient safety and improved clinical outcomes.

3.3. Subgroup Analysis

Quantitative comparisons between the BrSET and MoDRIA datasets (Tables 1 and 2) demonstrated substantial domain shift during external validation. Significant differences were observed in demographic composition, disease severity distribution, and imaging acquisition protocols. For example, the mean participant age differed markedly between BrSET and MoDRIA (57.66 $\pm$ 18.26 years vs. 42.82 $\pm$ 16.70 years; $p < 0.001$), while the prevalence of proliferative diabetic retinopathy increased from 2.42% in BrSET to 6.55% in MoDRIA ($p < 0.001$). In addition, BrSET predominantly utilised mydriatic Canon and Nikon imaging systems, whereas MoDRIA employed non-mydriatic 3nethra and Canon CR-2 cameras under real-world screening conditions. These inter-dataset differences likely altered retinal feature distributions, image quality characteristics, and disease presentation patterns encountered during inference, thereby contributing to reduced discriminative performance, calibration deterioration, and subgroup instability during external validation.

Table 8 presents the AUROC values for various models stratified by age and imaging technology, showcasing both internal and external validation performance metrics.

3.3.1. Internal Subgroup Analysis

In the internal subgroup analysis, models performed exceptionally well for patients aged <60 years. Notably, ConvNeXt achieved the highest AUROC of 0.9960, followed closely by DenseNet at 0.9930 and EfficientNetB3 at 0.9952. All models in this age group exhibited AUROC values exceeding 0.9800, indicating their robust ability to effectively identify DR in younger patients. In contrast, for the subgroup aged $\geq$60 years, there was a noticeable decline in performance across all models, yet ConvNeXt maintained the

highest AUROC of 0.9837. The performance for this age group was still strong but lower in the younger cohort, highlighting the models' decreased effectiveness as the patient age increases, with EfficientNetB0 achieving the lowest AUROC of 0.9690 in this subgroup.

Table 8. Summarised internal and external subgroup AUCs for each model.

Subgroup	ConvNeXt (auc)	DenseNet	Efficient NetB0	ResNet50	Inception-V3	Efficient NetB3
Internal subgroup AUROC						
Age <60	0.9960	0.9930	0.9896	0.9935	0.9938	0.9952
Age ≥ 60	0.9837	0.9751	0.9690	0.9739	0.9789	0.9859
Canon CR	0.9867	0.9777	0.9747	0.9769	0.9796	0.9874
External subgroup AUROC						
Age < 60	0.5298	0.5491	0.6054	0.5027	0.5755	0.5687
Age ≥ 60	0.6602	0.6338	0.6616	0.5977	0.6160	0.5921
3Nethra	0.5602	0.5750	0.5925	0.5402	0.5925	0.5827

When examining performance based on imaging technology, AUROC values indicate relatively consistent results across models for Canon CR imaging, with ConvNeXt at 0.9867 and EfficientNetB3 marginally higher at 0.9874. This suggests that the models can capitalise on the high-quality images provided by Canon CR systems, thereby yielding reliable detectability for DR.

3.3.2. External Subgroup Analysis

The external subgroup analysis highlighted stark discrepancies in model performance compared to internal validation. For patients aged <60 years, ConvNeXt dropped sharply to an AUROC of 0.5298, highlighting significant generalisation challenges beyond the training dataset. Similarly, other models, such as DenseNet and ResNet50, also exhibited declining performance, with AUROCs of 0.5491 and 0.5027, respectively. This indicates that while the models were effective on internal datasets, they struggled to maintain performance when applied to external validation datasets that may differ in demographic characteristics or imaging conditions.

In the ≥60 years age group, while the models showed some improvement compared to the younger cohort, AUROC values remained moderate, with ConvNeXt at 0.6602 and EfficientNetB0 slightly higher at 0.6616. These results further emphasise the need for improved model adaptability across diverse populations and imaging contexts.

In terms of imaging technology, external validation of the 3nethra classic imaging system yielded an AUROC of 0.5602, suggesting variability in model reliability across validations. The variance in results between Canon and 3nethra classic cameras may stem from discrepancies in image quality, resolution, and colour accuracy. If a model predominantly trained on one type of camera encounters images from another during validation, it may misclassify or fail to recognise certain features indicative of DR, thus affecting its predictive accuracy. Using two different high-quality cameras, even with skilled ophthalmologists performing screening, may not yield the same results in DR screening for several reasons. First, different cameras can vary in image quality and resolution, affecting the clarity and detail of captured images. These differences can affect the detection of subtle signs of DR. Additionally, colour accuracy may vary between camera systems, and accurate colour representation is crucial for identifying certain features of DR; poor fidelity in one camera could lead to misinterpretation of the images. The field of view can vary between cameras, with some offering a wider perspective that provides a more comprehensive assessment, while others may miss important regions.

Furthermore, calibration and settings unique to each camera can influence image acquisition, so different cameras may produce inconsistent image quality depending on exposure and focus settings. Finally, while skilled ophthalmologists are adept at compensating for some technical differences through their expertise, variability in interpretation may still occur due to nuances in the images. Consequently, although high-quality cameras and skilled professionals enhance the accuracy of DR screening, using different camera systems introduces variability in results, suggesting that standardising imaging protocols and consistently using a single camera system may be more effective for routine assessments.

4. Discussion

The results obtained from both internal and external validation of the models present a compelling landscape of strengths and weaknesses, emphasising the importance of rigorous model evaluation in AI solution development prior to deployment in real world settings. The disparity observed between internal and external performance metrics aligns with findings in the literature emphasising the challenges of generalisation in predictive modelling, particularly in computer vision tasks.

4.1. Model Performance Insights

The internal validation results, with accuracy metrics exceeding 0.95 across most models, indicate effective learning and adaptation to the training dataset. EfficientNetB3 demonstrated superior overall discriminative performance, achieving the highest AUROC (0.9892), highest sensitivity (0.9573), and lowest Brier score (0.0168). Meanwhile, DenseNet exhibited the most balanced clinical screening performance, achieving the highest F1-score (0.7259) and Youden Index (0.2381), indicating improved balance between sensitivity and specificity. Collectively, these findings suggest that EfficientNetB3 and DenseNet were the strongest-performing architectures during internal validation.

However, the significant drop in external validation performance serves as a reminder of the common pitfalls in machine learning, where models may exhibit exceptional performance in controlled settings yet struggle when applied to real-world scenarios. This resonates with the work of Liang et al. [33] and Song et al. [34], which highlights the overfitting issues commonly encountered when models are not sufficiently exposed to variability during the training phase.

The performance degradation observed in external validation metrics, especially the low sensitivity values, correlates with challenges reported in recent studies, such as those by Aroquaux et al. [35] and Erwin et al. [36], which emphasise the importance of dataset diversity in developing models that are resilient to variations in input data. The need for models to adapt to various conditions (such as lighting, camera quality, and scene complexity) is crucial for achieving high performance in practical applications.

Significantly, the observed declines in performance were accompanied by measurable differences in demographic and disease characteristics between the training (BrSET) and external validation (MoDRIA) cohorts. The MoDRIA dataset comprised younger participants, exhibited a higher prevalence of proliferative DR, and included images captured using differing imaging protocols and camera systems. These variations likely introduced substantial covariate and domain shifts, altering the distribution of retinal features presented to the models during inference.

While the present study did not conduct formal feature space similarity analyses or ablation experiments, the statistically significant inter-dataset differences presented in Tables 1 and 2 provide quantitative evidence supporting the impact of dataset heterogeneity on reduced external generalisability.

Notably, the divergence between overall accuracy and clinically meaningful metrics, such as sensitivity and F1-score, highlights the limitations of relying solely on accuracy in highly imbalanced diagnostic datasets. Several models achieved relatively high external accuracy primarily by correctly classifying the dominant non-referable DR class, while simultaneously demonstrating poor detection of referable disease. These findings indicate that accuracy may substantially overestimate real-world clinical utility when disease prevalence is uneven. Consequently, metrics such as AUROC, calibration performance, sensitivity, and F1-score provide more robust indicators of model reliability and screening effectiveness in AI-assisted diagnostic systems.

4.2. Calibration and Generalisation

Calibration issues remain highly prevalent in modern deep architectures. Notably, models such as ConvNeXt and Inception-v3 frequently exhibit severe overconfidence in their predictive outputs, which can compromise downstream clinical decision-making. Furthermore, our evaluation reveals a critical gap in generalization: models demonstrating high confidence during internal validation suffer from severe miscalibration and degraded reliability when evaluated on external test sets. This generalization failure is consistent with findings reported by Roy et al. [37] and Collins et al. [38]. This phenomenon further supports the observations of Wang et al. [39] and Hekler et al. [40], who argue that calibration is routinely overlooked during optimization, leading to overly optimistic predictions that fail to capture true predictive uncertainty. Consequently, addressing these calibration deficits is essential prior to clinical deployment, where reliable uncertainty estimation is paramount to patient safety.

Calibration analysis indicated that probabilistic predictions were not uniformly reliable across architectures. Although internal Brier scores were relatively low, calibration slopes greater than 1 and non-zero intercepts suggested modest miscalibration. Furthermore, calibration deteriorated substantially during external validation, indicating that probability estimates learned from the BrSET dataset did not transfer reliably to the MoDRIA cohort. Several architectures exhibited overconfident probability estimates despite poor discriminative performance, highlighting inadequate uncertainty calibration under domain-shift conditions. These findings emphasise the need for recalibration, threshold adaptation, and continuous post-deployment monitoring before clinical implementation in geographically and demographically distinct populations.

The present study contributes important evidence regarding the external validity and clinical robustness of AI models for diabetic retinopathy screening in underrepresented populations. By systematically evaluating multiple contemporary CNN architectures across retinal datasets from Latin America and SSA, we demonstrate that strong internal performance does not necessarily translate into reliable real-world generalisability. Although several models achieved excellent discrimination during internal validation, substantial reductions in AUROC, sensitivity, and calibration performance were observed during external testing, emphasising the vulnerability of DL systems to demographic, epidemiological, and imaging domain shifts. Importantly, our findings extend beyond conventional accuracy metrics by incorporating calibration assessment, subgroup analysis, and imaging device variability, thereby providing a more clinically meaningful evaluation of model reliability. These results reinforce growing concerns that AI systems developed using geographically homogeneous datasets may inadequately serve diverse populations, particularly in low-resource settings where disease presentation, imaging infrastructure, and healthcare pathways differ substantially from those represented in high-income training cohorts. Collectively, our findings emphasise the necessity of geographically diverse train-

ing datasets, rigorous external validation, and continuous post-deployment monitoring to ensure equitable, safe, and clinically transferable AI-assisted screening systems.

These findings also have important implications for fairness and equity in medical AI. Recent studies [24–26] have raised concerns that AI systems trained predominantly using data from high-income countries may systematically underperform in underrepresented populations, thereby exacerbating existing healthcare disparities. In ophthalmology, differences in retinal pigmentation, disease prevalence, imaging infrastructure, and healthcare access may substantially influence model behaviour across populations. The present study provides empirical evidence supporting these concerns, demonstrating that models achieving excellent internal performance may nevertheless exhibit clinically inadequate sensitivity and calibration when applied to African populations and real-world screening environments. Consequently, equitable deployment of AI-assisted screening systems will require geographically representative training datasets, continuous external validation, and adaptive recalibration strategies tailored to local clinical contexts. From a deployment perspective, these findings indicate that the evaluated models should not yet be considered suitable for fully autonomous clinical implementation across heterogeneous healthcare environments. The substantial deterioration observed during external validation emphasises the importance of prospective real-world validation, threshold adaptation, post-deployment recalibration, and continuous performance monitoring before integration into routine DR screening programmes. These considerations are consistent with key STARD-AI and CONSORT-AI recommendations for trustworthy clinical AI deployment.

4.3. Subgroup Analyses

The findings from our subgroup analyses provide significant insights into the performance of deep learning models for DR screening across different patient groups and imaging modalities. In the internal analysis, models such as ConvNeXt, DenseNet, and EfficientNetB3 excelled at DR detection among younger patients (under 60 years), with AUROCs (a statistical measure of model accuracy) over 0.9800. This reflects their robustness on high-quality training datasets and their ability to recognise early DR signs.

However, a notable decline in performance for older patients ($\geq$60 years) was observed. ConvNeXt achieved an AUROC of 0.9837 in this group, which is significantly lower than its performance in younger cohorts. This decline suggests challenges in generalising the model due to age-related ocular changes and increased comorbidities. Models must adapt to address the complexities of older populations. The results from the external validation reinforce these findings. Performance dropped sharply, with the AUROC of ConvNeXt falling to 0.5298 for patients aged <60 years in external datasets. This highlights the need for diverse training strategies that cover a range of imaging conditions and demographics. Variability between Canon CR (Canon Computed Radiography) and 3Nethra imaging (a specific retinal imaging device) further indicates that models must be trained on multiple modalities to improve generalisation.

Imaging technology directly affects performance. Image quality, resolution, and colour calibration greatly impact model accuracy in detecting rDR. Researchers should develop standardised imaging protocols to reduce disparities. This will enhance model performance and support the widespread adoption of AI-based DR screening. While skilled ophthalmologists improve image interpretation, differences in camera systems can still lead to inconsistent diagnostic results. Future studies should ensure the capabilities of imaging devices match the needs of automated screening. This alignment can help to translate AI advancements into better clinical outcomes.

5. Conclusions

This study aimed to evaluate the performance of various DL models for DR screening across different demographics and imaging modalities. Our internal validation demonstrates that advanced architectures such as ConvNeXt, EfficientNetB0, DenseNet, and EfficientNetB3 excel at detecting rDR in younger patients, with AUROCs exceeding 0.9800. While these results indicate strong potential for these models under conditions where the training and target populations are closely matched, the subsequent external validation highlights that this potential is currently limited by significant domain shifts.

The significant drop in performance observed during external validation highlights critical challenges in model generalisation. The models struggled to maintain accuracy, particularly in older patients and across varied imaging systems, as evidenced by marked declines in AUROC values. These findings emphasise the necessity for comprehensive training strategies that include diverse datasets reflective of various demographic characteristics and imaging conditions.

The implications of this research are profound, indicating the need for enhanced model adaptability to ensure high performance in real-world applications. Furthermore, standardising imaging protocols and calibrating models to account for differences in camera technology will be vital in improving diagnostic reliability.

While this study identified important limitations related to population variability and imaging heterogeneity, these findings provide valuable directions for future research aimed at improving AI generalisability in real-world clinical settings. Future work should investigate population-specific adaptation strategies and incorporate rigorous domain-shift analyses, including feature distribution assessment, latent representation similarity evaluation, and controlled ablation studies, to better understand the mechanisms underlying performance degradation across diverse populations and imaging environments. Additionally, advanced domain adaptation and data diversification approaches, such as Domain-Adversarial Neural Networks (DANN), CycleGAN-based synthetic image generation, and ensemble learning frameworks, should be explored to improve model robustness. Prospective multi-country validation, adaptive recalibration methods, and longitudinal post-deployment monitoring will also be essential to ensure fairness, reliability, and clinical applicability across heterogeneous healthcare settings. Ultimately, these findings emphasise that successful AI deployment in ophthalmology requires not only strong technical performance but also robust validation strategies that prioritise diversity, equity, and real-world clinical reliability.

Author Contributions: R.M. conceived the study, developed the methodology, implemented the deep learning models, performed data analysis, and wrote the original manuscript. F.K. contributed to the study's conceptualisation, provided supervision, and critically reviewed the manuscript. S.A. provided clinical expertise, contributed to study design, supervised the research, and reviewed the manuscript for intellectual content. W.W. contributed to model development, data curation, statistical analysis, and manuscript revision. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Mbarara University Data Science Research Hub (MUDSReH) through funding from the Fogarty International Center, U.S. National Institutes of Health (NIH; Award Number U54TW012043-01).

Institutional Review Board Statement: The study was conducted in accordance with Mbarara University of Science and Technology Research Ethics Committee (MUST-2021-239) on 14 October 2022 and the Uganda National Council for Science and Technology (HS2094ES).

Informed Consent Statement: This study was approved by the Mbarara University of Science and Technology Research Ethics Committee (MUST-2022-632) on 14 October 2022 and the Uganda

National Council for Science and Technology (HS2094ES) on 27 September 2023. Informed consent from the study populations is not required for registry linkage studies.

Data Availability Statement: The data presented in this study are openly available in physionet at https://www.physionet.org/, reference number [29].

Conflicts of Interest: The funders had no role in study design, data collection, analysis, interpretation, or manuscript preparation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Meaning
AI	Artificial Intelligence
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BrSET	Brazilian Multilabel Ophthalmological Dataset
CI	Confidence Interval
CNN	Convolutional Neural Network
DL	Deep Learning
DR	Diabetic Retinopathy
ICDR	International Clinical Diabetic Retinopathy
LMICs	Low- and Middle-Income Countries
MoDRIA	Multimodal Database of Retina Images for Africa
MUDSReH	Mbarara University Data Science Research Hub
NIH	National Institutes of Health
Nikon NF5050	Nikon NF5050 Non-Mydriatic Retinal Camera
rDR	Referable Diabetic Retinopathy
SSA	sub-Saharan Africa

References

1. Kropp, M.; Golubnitschaja, O.; Mazurakova, A.; Koklesova, L.; Sargheini, N.; Vo, T.-T.K.S.; de Clerck, E.; Polivka, J., Jr.; Potuznik, P.; Polivka, J. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—Risks and mitigation. *EPMA J.* **2023**, *14*, 21–42. [CrossRef]
2. Teo, Z.L.; Tham, Y.-C.; Yu, M.; Chee, M.L.; Rim, T.H.; Cheung, N.; Bikbov, M.M.; Wang, Y.X.; Tang, Y.; Lu, Y. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology* **2021**, *128*, 1580–1591. [CrossRef]
3. Sahiledengle, B.; Assefa, T.; Negash, W.; Tahir, A.; Regasa, T.; Tekalegn, Y.; Mamo, A.; Teferu, Z.; Solomon, D.; Gezahegn, H. Prevalence and factors associated with diabetic retinopathy among adult diabetes patients in Southeast Ethiopia: A hospital-based cross-sectional study. *Clin. Ophthalmol.* **2022**, *16*, 3527. [CrossRef] [PubMed]
4. Duncan, B.B.; Magliano, D.J.; Boyko, E.J. *IDF Diabetes Atlas 11th edition 2025: Global Prevalence and Projections for 2050*; Oxford University Press: Oxford, UK, 2025; p. gfaf177.
5. Cheloni, R.; Gandolfi, S.A.; Signorelli, C.; Odone, A. Global prevalence of diabetic retinopathy: Protocol for a systematic review and meta-analysis. *BMJ Open* **2019**, *9*, e022188. [CrossRef]
6. Olamoyegun, M.A.; Alare, K.; Afolabi, S.A.; Aderinto, N.; Adeyemi, T. A systematic review and meta-analysis of the prevalence and risk factors of type 2 diabetes mellitus in Nigeria. *Clin. Diabetes Endocrinol.* **2024**, *10*, 43. [CrossRef]
7. Hayati, A.; Abdol Homayuni, M.R.; Sadeghi, R.; Asadigandomani, H.; Dashtkoohi, M.; Eslami, S.; Soleimani, M. Advancing diabetic retinopathy screening: A systematic review of artificial intelligence and optical coherence tomography angiography innovations. *Diagnostics* **2025**, *15*, 737. [CrossRef]
8. Yim, D.; Chandra, S.; Sondh, R.; Thottarath, S.; Sivaprasad, S. Barriers in establishing systematic diabetic retinopathy screening through telemedicine in low-and middle-income countries. *Indian J. Ophthalmol.* **2021**, *69*, 2987–2992. [CrossRef]
9. Rogers, M.; Kaggwa, F.; Arunga, S.; Wasswa, W. Feasibility and Equity of AI-Driven Models for Diabetic Retinopathy Screening in Sub-Saharan Africa: A Systematic Review and Meta-Analysis. *Cureus J. Comput. Sci.* **2025**, *2*, es44389-025-07263-5. [CrossRef]
10. Irodi, A.; Zhu, Z.; Grzybowski, A.; Wu, Y.; Cheung, C.Y.; Li, H.; Tan, G.; Wong, T.Y. The evolution of diabetic retinopathy screening. *Eye* **2025**, *39*, 1040–1046. [CrossRef] [PubMed]

11. Balakrishnan, K.; Velusamy, D.; Hinkle, H.E.; Li, Z.; Ramasamy, K.; Khan, H.; Ramaswamy, S.; Shah, P.M. Artificial intelligence in rural healthcare delivery: Bridging gaps and enhancing equity through innovation. *arXiv* **2025**, arXiv:2508.11738. [CrossRef]
12. Rajalakshmi, R.; PramodKumar, T.A.; Naziyagulnaaz, A.S.; Anjana, R.M.; Raman, R.; Manikandan, S.; Mohan, V. *Leveraging Artificial Intelligence for Diabetic Retinopathy Screening and Management: History and Current Advances*; Taylor & Francis: Abingdon, UK, 2025; pp. 719–726.
13. Olawade, D.B.; Weerasinghe, K.; Mathugamage, M.D.D.E.; Odetayo, A.; Aderinto, N.; Teke, J.; Boussios, S. Enhancing ophthalmic diagnosis and treatment with artificial intelligence. *Medicina* **2025**, *61*, 433. [CrossRef]
14. Panahi, O. Deep learning in diagnostics. *J. Med. Discov.* **2025**, *2*, 1–6.
15. Costin, H.-N.; Fira, M.; Goraș, L. Artificial intelligence in ophthalmology: Advantages and limits. *Appl. Sci.* **2025**, *15*, 1913. [CrossRef]
16. Khan, Z.; Gaidhane, A.M.; Singh, M.; Ganesan, S.; Kaur, M.; Sharma, G.C.; Rani, P.; Sharma, R.; Thapliyal, S.; Kushwaha, M. Diagnostic accuracy of IDX-DR for detecting diabetic retinopathy: A systematic review and meta-analysis. *Am. J. Ophthalmol.* **2025**, *273*, 192–204. [CrossRef]
17. Vought, R.; Vought, V.; Shah, M.; Szirth, B.; Bhagat, N. EyeArt artificial intelligence analysis of diabetic retinopathy in retinal screening events. *Int. Ophthalmol.* **2023**, *43*, 4851–4859. [CrossRef] [PubMed]
18. Kim, T.N.; Aaberg, M.T.; Li, P.; Davila, J.R.; Bhaskaranand, M.; Bhat, S.; Ramachandra, C.; Solanki, K.; Myers, F.; Reber, C. Comparison of automated and expert human grading of diabetic retinopathy using smartphone-based retinal photography. *Eye* **2021**, *35*, 334–342. [CrossRef]
19. Kąpa, M.; Koryciarz, I.; Kustosik, N.; Jurowski, P.; Pniakowska, Z. Modern approach to diabetic retinopathy diagnostics. *Diagnostics* **2024**, *14*, 1846. [CrossRef]
20. Vanessa, C.; Au, S. Artificial intelligence systems for diabetic retinopathy screening: Appraisal on the 3rd US FDA approved algorithms-AEYE-DS. *J. Ophthalmol. Adv. Res.* **2023**, *4*, 1–3.
21. Grzybowski, A.; Singhanetr, P.; Nanegrungsunk, O.; Ruamviboonsuk, P. Artificial intelligence for diabetic retinopathy screening using color retinal photographs: From development to deployment. *Ophthalmol. Ther.* **2023**, *12*, 1419–1437. [CrossRef] [PubMed]
22. Abdalla, M.M.I.; Mohanraj, J. Revolutionizing diabetic retinopathy screening and management: The role of artificial intelligence and machine learning. *World J. Clin. Cases* **2025**, *13*, 101306. [CrossRef] [PubMed]
23. Tan, T.F.; Thirunavukarasu, A.J.; Jin, L.; Lim, J.; Poh, S.; Teo, Z.L.; Ang, M.; Chan, R.P.; Ong, J.; Turner, A. Artificial intelligence and digital health in global eye health: Opportunities and challenges. *Lancet Glob. Health* **2023**, *11*, e1432–e1443. [CrossRef]
24. Jacoba, C.M.P.; Celi, L.A.; Lorch, A.C.; Fickweiler, W.; Sobrin, L.; Gichoya, J.W.; Aiello, L.P.; Silva, P.S. *Bias and Non-Diversity of Big Data in Artificial Intelligence: Focus on Retinal Diseases: Massachusetts Eye and Ear Special Issue*; Taylor & Francis: Abingdon, UK, 2023; pp. 433–441.
25. Vrudhula, A.; Kwan, A.C.; Ouyang, D.; Cheng, S. Machine learning and bias in medical imaging: Opportunities and challenges. *Circ. Cardiovasc. Imaging* **2024**, *17*, e015495. [CrossRef] [PubMed]
26. Dortche, K.; McCarthy, G.; Banbury, S.; Yannatos, I. Promoting health equity through improved regulation of artificial intelligence medical devices. *J. Sci. Policy Gov.* **2023**, *21*, 3. [CrossRef]
27. Kinney, M.; Anastasiadou, M.; Naranjo-Zolotov, M.; Santos, V. Expectation management in AI: A framework for understanding stakeholder trust and acceptance of artificial intelligence systems. *Heliyon* **2024**, *10*, e28562. [CrossRef]
28. Preece, A.; Harborne, D.; Braines, D.; Tomsett, R.; Chakraborty, S. Stakeholders in explainable AI. *arXiv* **2018**, arXiv:1810.00184. [CrossRef]
29. Nakayama, L.; Goncalves, M.; Zago Ribeiro, L.; Santos, H.; Ferraz, D.; Malerbi, F. A Brazilian Multilabel Ophthalmological Dataset (BRSET). 2023. Available online: https://physionet.org/content/brazilian-ophthalmological/1.0.0/ (accessed on 14 August 2024).
30. Kwaga, T. Multimodal Database of Retina Images for Africa (MoDRIA): The First Open Access Database for Retinal Fundus Photos in Sub_Saharan Africa. *Investig. Ophthalmol. Vis. Sci.* **2025**, *66*, 3870.
31. Sounderajah, V.; Ashrafian, H.; Golub, R.M.; Shetty, S.; De Fauw, J.; Hooft, L.; Moons, K.; Collins, G.; Moher, D.; Bossuyt, P.M. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. *BMJ Open* **2021**, *11*, e047709. [CrossRef] [PubMed]
32. Liu, X.; Rivera, S.C.; Moher, D.; Calvert, M.J.; Denniston, A.K.; Ashrafian, H.; Beam, A.L.; Chan, A.-W.; Collins, G.S.; Deeks, A.D.J. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Lancet Digit. Health* **2020**, *2*, e537–e548. [CrossRef]
33. Liang, W.; Tadesse, G.A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; Zou, J. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **2022**, *4*, 669–677. [CrossRef]
34. Song, C.; Ristenpart, T.; Shmatikov, V. *Machine Learning Models That Remember Too Much*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 587–601.

35. Varoquaux, G.; Colliot, O. Evaluating machine learning models and their diagnostic value. *Mach. Learn. Brain Disord.* **2023**, *197*, 601–630.
36. Erwin, G.D.; Oksenberg, N.; Truty, R.M.; Kostka, D.; Murphy, K.K.; Ahituv, N.; Pollard, K.S.; Capra, J.A. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.* **2014**, *10*, e1003677. [CrossRef] [PubMed]
37. Roy, P. Enhancing real-world robustness in AI: Challenges and solutions. *J. Recent Trends Comput. Sci. Eng.* **2024**, *12*, 34–49. [CrossRef]
38. Collins, G.S.; Dhiman, P.; Ma, J.; Schlussel, M.M.; Archer, L.; Van Calster, B.; Harrell, F.E.; Martin, G.P.; Moons, K.G.; Van Smeden, M. Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ* **2024**, *384*, e074819. [CrossRef] [PubMed]
39. Wang, C. Calibration in deep learning: A survey of the state-of-the-art. *arXiv* **2023**, arXiv:2308.01222.
40. Hekler, A.; Kuhn, L.; Buettner, F. Beyond overconfidence: Foundation models redefine calibration in deep neural networks. *arXiv* **2025**, arXiv:2506.09593. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.