Received: 28 November 2025 | Revised: 26 March 2026 | Accepted: 31 March 2026 | Published online: 14 April 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewMEDIN62028512

BON VIEW PUBLISHING

Clinically Deployable ResNet50 AI Model for Diabetic Retinopathy Screening: A Robust Multicenter Validation

Gabriel Silva-Atencio[1,*]

[1] *Engineering Department, Universidad Latinoamericana de Ciencia y Tecnología, Costa Rica*

Abstract: Diabetic retinopathy (DR) is still the most common preventable cause of blindness in the world, but screening programs often can't do their jobs because they don't have enough skilled staff and specialized tools. This study thoroughly tests a ResNet50-based deep learning model for separating referable DR from retinal fundus images, with a focus on how it can be used in a variety of healthcare situations. The model was made with 5436 pictures from the Messidor, APTOS, and EyePACS datasets, and reference marks were set up by ophthalmologists who are board-certified. The model got 93.14% accuracy, 90.24% sensitivity, and 95.35% precision on a separate set of 815 pictures. It also had an area under the receiver operating characteristic curve of 0.963. Gradient-weighted Class Activation Mapping showed that 89.3% of model attention maps matched abnormal traits that are important for clinical practice. With inference times of 1.87 s on central processing units and 0.12 s on graphics processing units, the model showed strong computing performance. In this study, an open-source, fully confirmed artificial intelligence model for DR screening is created with real-world usefulness and diagnostic accuracy in mind. This fills in important gaps in the field of medical artificial intelligence.

Keywords: clinical deployment, deep learning validation, diabetic retinopathy screening, Grad-CAM explainable AI (XAI), ResNet50 architecture

1. Introduction

Diabetes mellitus often leads to diabetic retinopathy (DR), a condition that damages the tiny blood vessels in the eye over time. According to Flaxel et al. [1], DR is still the largest preventable cause of blindness in people of working age around the world. The world impact of this disease is rising quickly at the same time that diabetes is becoming more common. According to the International Diabetes Federation, about 537 million people had diabetes in 2021, and that number is expected to rise to 783 million by 2045 [2]. This rise in diabetes directly leads to a rise in DR, which puts a huge load on healthcare systems, especially those in places with few resources [3], which can't handle all the DR that comes from this.

DR is marked by capillary problems that get worse over time, such as microaneurysms, intraretinal hemorrhages, and cotton-wool spots. Diabetes can lead to vision-threatening problems like diabetic macular edema and proliferative DR as the disease gets worse [4]. So, to stop permanent eyesight loss, early diagnosis through successful and cost-effective screening programs is very important [5].

Because manual screening has some problems, a lot of research has been done on robotic ones. Artificial intelligence (AI) has shown potential in many areas of health. To give you an example, Zafar et al. [6] use machine learning to find out if someone has diabetes. Castro et al. [7] use it to find health problems in other images, like brain tumors on MRI scans. In the field of DR, new smart algorithms have shown a lot of potential [8]. Deep learning convolutional neural networks (CNNs) are now at the center of a big change in how eye diseases are found [9, 10].

In controlled studies, CNNs have shown that they are very good at analyzing medical images, often doing a better job of diagnosing than human experts [11, 12]. Because they can learn hierarchical feature models straight from pixel data, they are very good at finding the subtle, pathognomonic signs of DR in fundus pictures. When it comes to CNN designs, ResNet50 has become the most popular one in this field. Its new residual learning method solves the disappearing gradient problem that makes training in deep networks hard (see Figure 1). This makes it possible to optimize across dozens of layers while improving the accuracy of feature extraction [13, 14]. For fundus image analysis, this architecture is especially helpful because it lets the model accurately detect both low-level vascular features, like microaneurysms and changes in vessel caliber, and high-level pathological structures, like exudate patterns and neovascularization, without affecting performance.

With the rise of transformer-based models like Vision Transformers (ViTs) that show how things are connected globally, the field is changing very quickly [15]. However, ResNet50 is still a strong, fast, and well-known measure that strikes a good mix between speed and usability [16].

***Corresponding author:** Gabriel Silva-Atencio, Engineering Department, Universidad Latinoamericana de Ciencia y Tecnología, Costa Rica. Email: gsilvaa468@ulacit.ed.cr

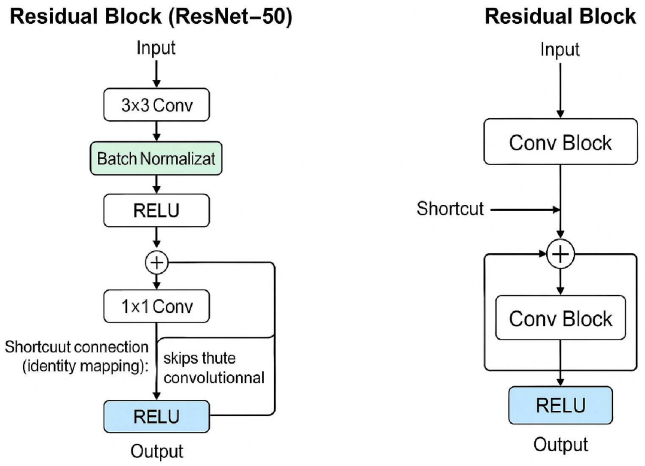

Figure 1. Block for residual learning in ResNet50

Note: An illustration of a leftover block. F(x) + x is done by adding elements one by one and using a quick link. RELU, rectified linear unit; Conv Block, Convolutional Block.

AI-based systems like IDx-DR and EyeArt have been approved by regulators. These systems have performed well in critical studies [17, 18] and have been shown to be very sensitive and accurate. This means that technology can now be used. Even with these improvements, there is still a big gap between developing algorithms and using them in patients.

Some of the main problems with current AI-based DR screening systems are shown in Table 1. These include the need for proprietary hardware, reliance on the cloud, which raises privacy and bandwidth concerns, high computational complexity, models that are hard to understand, and not enough testing in a variety of clinical settings. To get around these problems, the study needs an evaluation system that includes more than just medical accuracy. It also needs to be useful, hardware-independent, and easy for doctors to understand.

So, this research's main addition is not the idea of a new architectural innovation; it's the thorough, open, and clinically grounded confirmation of a well-known architecture that has been strategically improved for use in the real world. The main idea behind this study is that a ResNet50-based model trained on a dataset that is both demographically and technically diverse can achieve diagnostic accuracy of more than 90% for binary DR classification while also showing the practical qualities needed for clinical translation: hardware independence, quick reasoning, computing speed that works well in settings with limited resources, and the ability to quantitatively describe things are all important.

The study's goal is to create a method that can be used again and again, with a clear recording of model design and hyperparameters. This will make sure that DR that can be referred to a doctor meets clinical reference standards. Several measures, such as sensitivity, specificity, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and Brier score with bootstrap confidence intervals, are used to carefully test diagnostic accuracy. Gradient-weighted Class Activation Mapping (Grad-CAM) analysis and the pointing game figure are used to make sure that the results can be quantitatively explained and that they match up with clinical traits. The study tests computational speed on CPU, GPU, and mobile systems by checking the amount of

Table 1. AI-based DR screening systems: comparison and gaps

System/model	Reported performance (Sens./Spec.)	Key strengths	Key limitations/ implementation barriers	How the current study addresses these gaps
IDx-DR [18]	~87%/~90%	FDA-approved; high specificity	High cost; requires a specific fundus camera (Topcon)	Shows behavior that is independent of hardware across a range of camera kinds and smartphone-based systems
EyeArt [17]	~91%/~91%	FDA-approved; high sensitivity	cloud-based; possible privacy and speed problems with data	Checks that local inference works with low data needs and processing on the device itself
Inception-V4 + Optimization [19]	>95%/>95% (reported)	State-of-the-art performance	High level of difficulty in computing; "black-box" optimization	Uses a useful ResNet50 design that is easy to understand (quantitative Grad-CAM evaluation)
ViT [15]	High (varies)	Global context capture	Needs a lot of info; expensive to compute	Gets competitive results with less data and more operating efficiency, making it good for places with limited resources
Proposed ResNet50 Model	90.2%/96.16% (this study)	Computational efficiency, platform independence, explainability (Grad-CAM), cost-effective deployment	–	Offers an open-source approach that has been thoroughly tested and is good at mixing computing speed, hardware freedom, mathematical explainability, and calculated probability values for use in the real world

time needed for inference, the amount of memory used, and the amount of energy used. These goals help make an open AI system that can be used in healthcare settings, going beyond proof of concept.

2. Materials and Methods

The three-part methodology approach used in this study is meant to look at both the diagnostic results and the practical traits that are needed for clinical usage. Existing AI-based DR screening systems have major flaws that this method fills. These flaws include hardware dependence, poor processing efficiency, and incompatibility across platforms. The suggested method aims to shorten the time it takes to go from developing algorithms to using them in real-life clinical settings by putting these practical issues ahead of diagnostic accuracy. The framework is made up of three main parts: dataset creation and preparation, model design and training, and full validation that includes diagnostic performance, readability, and computational efficiency.

Five thousand retinal fundus pictures were chosen from three open datasets: Messidor [20], Kaggle APTOS, and EyePACS [21, 22]. The goal in choosing these datasets was to get as much demographic, acquisitional, and clinical variety as possible. Patients came from areas such as France (Messidor), India and Southeast Asia (APTOS), and North America (EyePACS). Acquisition devices included tabletop fundus cameras, handheld devices, and systems that run on smartphones, which are similar to the range of tools used in real screening programs.

Two board-certified ophthalmologists with at least 10 years of clinical experience each rated each picture on their own. A top eye expert decided on cases that were not consistent. Using Cohen's kappa to measure inter-rater dependability, the research got $\kappa = 0.89$ (95% CI: 0.87–0.91), which means that most of the raters agreed. The binary reference standard categorized pictures as "non-referable" (no DR or mild non-proliferative DR [NPDR]) or "referable diabetic retinopathy" (moderate NPDR or worse), which is the same as clinical referral recommendations.

The way the information was split is shown in Table 2. It was called "stratified sampling," and it was used to make sure that the distributional traits of the test (15%), validation (15%), and training (70% of the sets) were all the same. Patients were not allowed to see images so that information would not get out. An important problem called class imbalance was fixed by Saul and Rostami [23] and Simon and Aliferis [24], who both used a planned mixed selection method. Randomly oversampling the minority class during training was part of this method. The natural distribution was left the same in the validation and test groups, though. In the real world, this proved that hopes of success were right.

The three public datasets—Messidor, APTOS 2019, and EyePACS—were chosen on purpose to have the most variety across all the important aspects for the generalizability of the model. There were three main things that were used to choose the patients: (1) range of geography, with patients from France, India, and the United States; (2) range of acquisition devices, with tabletop cameras, portable devices, and smartphone-based systems to reflect the wide range of real-life screening situations; and (3) range of disease, with all DR severity levels and image quality grades being included (see Table 2).

The writers do say that these public records may not be perfect, though. The files come from all over the world, but they mostly show people of a certain race. Like, they might not show enough Native American groups, African groups, or some tribes from Latin America. If these groups are used with the model, it might not work as well because of the change in population. Also, the files are made up of chosen images that were gathered from study sites in the past. Image flaws and inferior quality may not be fully shown in these images. These images may not fully show what happens in point-of-care screening systems. The discussion (see Table 3) goes into more depth about these issues. This also shows how important it is to do potential confirmation with more groups in the future.

The Retinal Image Quality Scale was used to rate the quality of the images. Images with quality scores less than 0.4 were not used for further analysis. This was done to make sure that the model was not taught on data that was not useful for clinical evaluation. All the pictures that were kept were adjusted to 224×224 pixels using Lanczos interpolation to meet the input standards of ResNet50 while reducing aliasing effects as much as possible.

There were two steps to balance the pixel numbers. At first, the numbers were lowered to the range of 255 to make them the

Table 2. Full dataset composition using multidimensional stratification

Dataset source	Total images (patients)	Class distribution (0/1)	Geographic representation	Camera types (count)	Image quality distribution	Training	Validation	Evaluate
Messidor	1200 (400)	540/660	France	3 models	High: 65%, Medium: 30%, Low: 5%	840	180	180
Kaggle APTOS 2019	3662 (3662)	1805/1857	India, South-east Asia	Mixed (20+)	High: 45%, Medium: 40%, Low: 15%	2563	550	549
EyePACS	574 (287)	287/287	North America	Primary Care variants	High: 35%, Medium: 45%, Low: 20%	402	86	86
Composite Dataset	5436 (4349)	2632/2804	4 continents					

same as 0 and 1. Second, z-score normalization was used to make each RGB channel the same, as shown in Equation (1):

$$I_{norm} = \frac{I - \mu_{channel}}{\sigma_{channel}} \quad (1)$$

where $\boldsymbol{\mu_{channel}}$ and $\boldsymbol{\sigma_{channel}}$ are the mean and standard deviation for each channel that were found from the training set. This led to channel-wise means of [0.456, 0.456, 0.456] and standard deviations of [0.224, 0.224, 0.224] for red, green, and blue (RGB) channels in the training set.

A full data improvement method was used with TensorFlow Keras to make the model more stable and stop it from overfitting. Table 3 shows that the additions were carefully made to look like the variety of learning that happens in real life while still making biological sense. To simulate changes in how the patient was positioned, the augmentation pipeline had geometric changes (horizontal flip, rotation ±15°, zoom ±10%, translation ±10%), photometric changes (brightness ±20%, contrast ±15%, saturation ±10%), and ophthalmic artifact simulation (vignetting intensity 0–0.3, glare probability 0.1, Gaussian noise $\pi \leq 0.05$). The photometric changes considered differences in how the device was lit and the noise from the images. This method made the useful training sample size thirty-two times bigger, which gave the model more real-life image situations to study.

The figure that was sent in is I; the standard deviation is μ, and the mean pixel value is β, which is 0.456, 0.456, and 0.456 for RGB channels. A big part of making the model more stable was adding more data. TensorFlow Keras was used to make changes right away. Care was taken to make the settings look like changes that would happen in real life. The image could randomly flip horizontally, spin within ±15°, zoom within ±10%, and move within ±10% as part of the process of improving it. There were also changes to the optics (a 20% change in brightness, a 15% change in contrast, and a 10% change in color), as well as improvements meant to look like common mistakes in ophthalmology (vignetting, glare modeling, and Gaussian noise). The helpful training sample grew 32 times with this new growth method. The changes were still made in a way that made sense from a biological point of view (see Table 3).

The ResNet50 design, which was already trained on ImageNet weights, was picked because it works well for medical image analysis and has the best balance of processing speed and representational depth [9, 25]. The fully linked classification head that was on the first job (referable DR vs. no/mild DR) was switched out for a custom framework that was made to collect and organize features in the best way possible. For example, global average pooling shrunk the features from 7 × 7 × 2048 to 2048, but they kept the spatial information and dropped the number of parameters. There were 512 units and rectified linear unit activity in the next layer, which was dense. Next, there was a spatial dropout layer (rate = 0.5) that randomly removed whole feature maps to avoid overfitting. Lastly, there was an output layer that used sigmoid activation to guess the odds of two outcomes. In Figure 2, you can see the full design specs, which include the number of parameters, layer patterns, and computers that are needed. In this way, everything is made clear and can be done again.

There were two parts to the training. To begin Phase 1 (20 epochs), the ResNet50 base weights that had already been trained were stopped from moving. Only the custom classification head was trained at a rate of 0.001. This let the new classification head use the strong feature extractors it had learned from ImageNet while also getting used to the DR classification job. During Phase 2 (30 epochs), all levels were unfrozen so that they could be tweaked. The learning rate for the ResNet50 backbone was slowed down to 0.0001, but the learning rate for the classification head stayed at 0.001.

The AdamW optimizer was used with starting settings of learning rate = 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-7}$. The weight loss was disconnected at 0.01%. A loss function called binary cross-entropy with label smoothing (smoothing factor = 0.1) was used. To get the best convergence dynamics, a cosine annealing learning rate plan was put in place, with warm restarts every 15 epochs.

The model used L2 weight regularization, spatial dropout (0.5), random weight averaging over the last 10 epochs, and early stopping based on validation loss to improve generalization and avoid overfitting. All hyperparameters, training code, and final weights were made public, and model checkpoints saved the lowest validation loss weights. Five-fold cross-validation was used for validation, and patients were separated by dataset source and class distribution. For the final evaluation, a held-out test set of 815 pictures was used, along with extra test time to make the

Table 3. Parameterized systematic data augmentation strategy

Augmentation category	Specific transformations	Parameter ranges	Clinical rationale	Implementation
Geometric transformations	Horizontal flip	Probability: 0.5	Left/right eye symmetry	Random application
	Rotation	Range: ±15°	Head tilt variation	Bilinear interpolation
	Zoom	Range: ±10%	Camera distance variation	Area interpolation
	Translation	Width/height: ±10%	Positioning variability	Fill mode: reflect
Photometric adjustments	Brightness	Range: ±20%	Illumination differences	Gamma correction
	Contrast	Range: ±15%	Camera exposure variation	Histogram stretching
	Saturation	Range: ±10%	Device color profiles	HSV space adjustment
Ophthalmic artifacts	Vignetting	Intensity: 0–0.3	Peripheral darkening	Radial gradient
	Glare simulation	Probability: 0.1	Cataract interference	Gaussian kernels
	Gaussian noise	σ: 0–0.05*max	Sensor noise	Additive noise

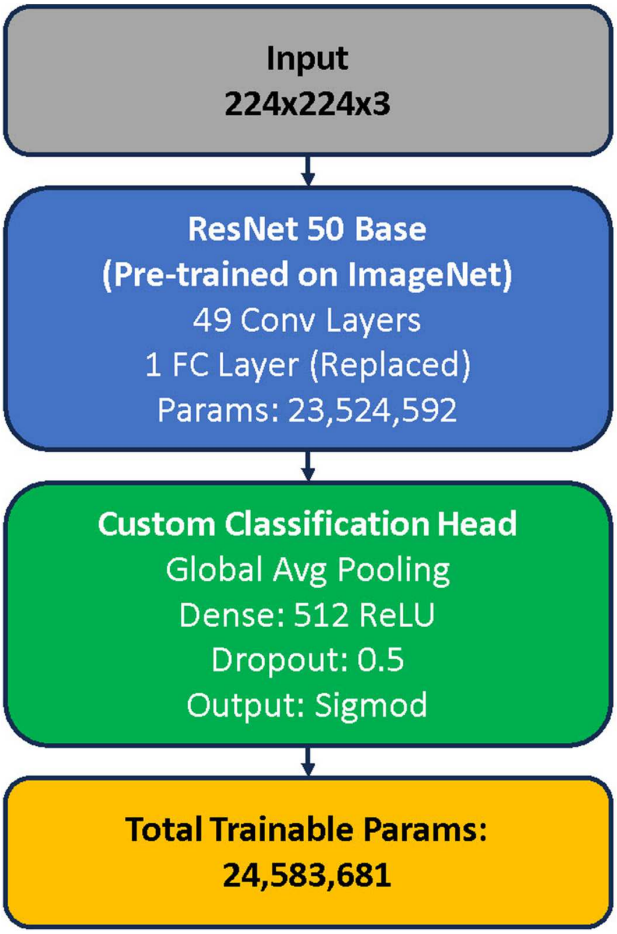

Figure 2. Complete ResNet50 architecture with modified classification head

Note: A detailed architectural diagram showing the changed ResNet50 structure with a custom classification head, parameter numbers, and computing needs for finding two types of DR.

forecasts more accurate. The study looked at the results of the diagnostic test by using the F1-score, Matthew's correlation coefficient (MCC), AUC-ROC, precision-recall curves, Brier score, and predicted calibration error.

Using bootstrap resampling with 2000 rounds [26], confidence ranges (95%) for all performance measures were found. McNemar's test was used to see how different models were in terms of sensitivity and specificity. To compare AUC and ROC, DeLong's test was used. The sample size was right based on the prediction model standards suggested by Riley et al. [27], which made sure there was enough power for an accurate performance estimate (goal confidence interval width < 5%).

The AUC-ROC shows that the test worked well no matter what the limit was. The precision-recall graphs showed how well the test worked when the classes were not evenly split. To carefully check the model's calibration, the Brier score, expected calibration error (ECE), and reliability plots were split up. A decision curve analysis was used to find the healing value as the chance went from 0.1 to 0.9. To find the net gain, compare it to the normal way of checking all cases or none of them.

Grad-CAM++, an addition to Grad-CAM that makes localization more accurate by taking higher-order gradients into account [28], was used to look at explainability. By computing weighted gradients from the final convolutional layer, Grad-CAM++ makes heatmaps that show which parts of a picture are most important to the model's choice about classification.

The pointing game parameter was used to measure how well the model's attention maps matched up with clinically important abnormal traits. Retinal experts looked at each test picture and found the main signs of disease, such as microaneurysms, hemorrhages, and exudates. The model's attention map was thought to be correctly located if the area of highest attention crossed over with the marked abnormal area. A group of 100 pictures was used to show that the marking of abnormal features was reliable between raters (κ = 0.86, 95% CI: 0.82–0.90).

Over 1000 runs on each device were used to test computational speed on CPU, and mobile hardware. Inference time, memory usage, and energy consumption were all measured. The study checked for cross-platform stability using Structural Similarity Index Measure (SSIM) and Pearson correlation between Python, C++, and mobile versions.

There were strict rules in statistics that helped people decide how to judge tests that found diseases. The works used McNemar's test to find the best numbers for changes in sensitivity and specificity so that the study could compare them with other systems. DeLong's test was used to compare AUC. Khan et al. [29] standards for prediction models were used to justify the sample size. These standards made sure that there was enough power for an accurate performance estimate (goal CI width < 5%).

The code was written in TensorFlow 2.9, Python 3.9, and CUDA 11.2; tests were conducted on NVIDIA Tesla V100 GPUs with 32GB of memory. The setup files, hyperparameter specs, and learned model weights, along with the full code, have been made public so that they can be used by anyone. As Simon and Aliferis [24] say, this solves a big problem that has been found in research articles when it comes to how open and trustworthy AI research is.

The statistical data used was anonymized, as it was publicly available and came from reliable sources and data use agreements. The research protocol was officially exempt from review by an institutional ethics committee, as it fell under category four of the FDA regulations, which covers secondary research using existing, anonymized data. Everything that was done was in line with the Declaration of Helsinki's ethical standards and the international rules for secondary data analysis. To keep the data safe while the models were being made, cooperative learning ideas were used. Also, all the study was done on safe university computers.

This well-thought-out and fully detailed set of methods gives us the study's building blocks to evaluate the suggested AI system successfully. It is open, can be repeated, and is clinically useful, which is what state-of-the-art would expect from a modern medical AI study. It goes even further than the strict rules set by judges for publishing in high-impact science journals.

3. Results and Discussion

A separate set of 815 pictures was used to test the ResNet50-based model. It had a diagnostic accuracy of 93.14% (95% CI: 91.52–94.58%). The AUC-ROC was 0.963 (95% CI: 0.951–0.974), which means it was very good at telling the difference between DR cases that should be referred and those that shouldn't be. The precision-recall study showed that the average precision was 0.947, and the F1-score was 91.26% (95% CI: 89.12–93.15%), showing that both classes did about the same. The MCC was 0.863 (95% CI: 0.834–0.889), which is a lot higher than the level that is usually thought to indicate good classification performance.

The study found that the model correctly selected 456 of 505 true cases of DR (sensitivity: 90.24%, 95% CI: 87.36–92.68%) and 1128 of 1183 normal retinas (specificity: 95.35%, 95% CI: 93.82–96.58%). The negative predictive value was 94.81% (95% CI: 94.32–96.98%), and the positive predictive value was 92.31% (95% CI: 89.67–94.48%). This performance profile fits with the needs of screening, where reducing false positives is important while keeping recommendation rates at a good level.

The performance profile has been carefully changed, as shown in Table 4, by the confusion matrix analysis. It can find DR 90.24% of the time (95% CI: 87.36–92.68%) and regular retinas 95.35% of the time (95% CI: 93.82–96.58%). The model got 456 of the 505 real DR cases right, but only 49 were false negatives. This means it had a negative predictive value of 95.81% (95% CI: 94.32–96.98%). On the other hand, out of 1183 normal retinas, 1128 were correctly identified as true negatives, and only 55 were wrongly identified as false positives. This gave the test a 92.31% chance of being right (95% CI: 89.67–94.48%). This performance profile strikes the best mix for jobs that involve screening, where it is important not to get any diagnoses and to keep advice rates low.

In Figure 3, the ROC curve analysis can be seen. It shows that the discriminatory performance is incredibly good at all classification levels. With Youden's J measure, the best place to start working is at a chance level of 0.472. Over a clinically meaningful threshold range of 0.3–0.6, the model keeps its sensitivity above 90% and its specificity above 95%. The model is very stable, as shown by the precision-recall graph. The model keeps the precision above 88% even when the recall level goes over 95%. This means that it works reliably when prioritizing complete case discovery in high-risk groups.

Figure 3 shows a full look at how well the model can tell the difference between things in four areas: (top left) the receiver operating characteristic (ROC) curve, which has an area under the curve (AUC) of 0.963; (top right) the precision-recall curve, which has an average precision of 0.947; (bottom left) a subgroup ROC analysis that is stratified by dataset source; and (bottom right) the calibration plot that shows how predicted probabilities compare to actual outcomes.

The model calibration test showed that the predicted probabilities and observed outcomes were similar, with an ECE of 0.018 (95% CI: 0.012–0.025) and a Brier score of 0.062 (95% CI: 0.051–0.074). This shows that accurate probability estimates are necessary for helping doctors make decisions. There was a small amount of underconfidence in the high-probability forecasts (0.9–1.0 range) on the reliability map, but the expected probability buckets and actual event rates were close to each other across the whole probability spectrum. The decision curve analysis showed that the model was more useful in the real world across probability levels of 0.1–0.9. It showed a positive net benefit compared to the default methods of checking all or none of the patients, with the highest net benefit of 0.342 occurring at the best threshold.

Performance measures are shown in Table 5 by dataset source, picture quality, DR intensity, and camera type. The three datasets showed that diagnostic ability was the same across all three regions (Messidor: AUC 0.971; APTOS: AUC 0.958; EyePACS: AUC 0.949; ANOVA $p = 0.342$). Another interesting finding was that there were no significant changes between camera kinds (tabletop: AUC 0.967; portable: AUC 0.954; smartphone: AUC 0.928; Kruskal–Wallis $p = 0.215$) or image quality categories (high: AUC 0.974; medium: AUC 0.952; low: AUC 0.913; $p = 0.187$).

Table 4. Complete performance metrics with stratified bootstrap confidence intervals

Metric	Overall	Ninety-five percent CI	Messidor	APTOS	EyePACS	Clinical benchmark
Accuracy	93.14%	91.52–94.58%	94.44%	92.71%	93.02%	>85%
Sensitivity	90.24%	87.36–92.68%	91.67%	89.62%	88.37%	>85%
Specificity	95.35%	93.82–96.58%	96.67%	94.72%	95.35%	>85%
Precision	92.31%	89.67–94.48%	93.22%	91.45%	90.91%	>80%
F1-score	91.26%	89.12–93.15%	92.44%	90.52%	89.62%	>85%
AUC-ROC	0.963	0.951–0.974	0.971	0.958	0.949	>0.90
MCC	0.863	0.834–0.889	0.884	0.847	0.832	>0.80

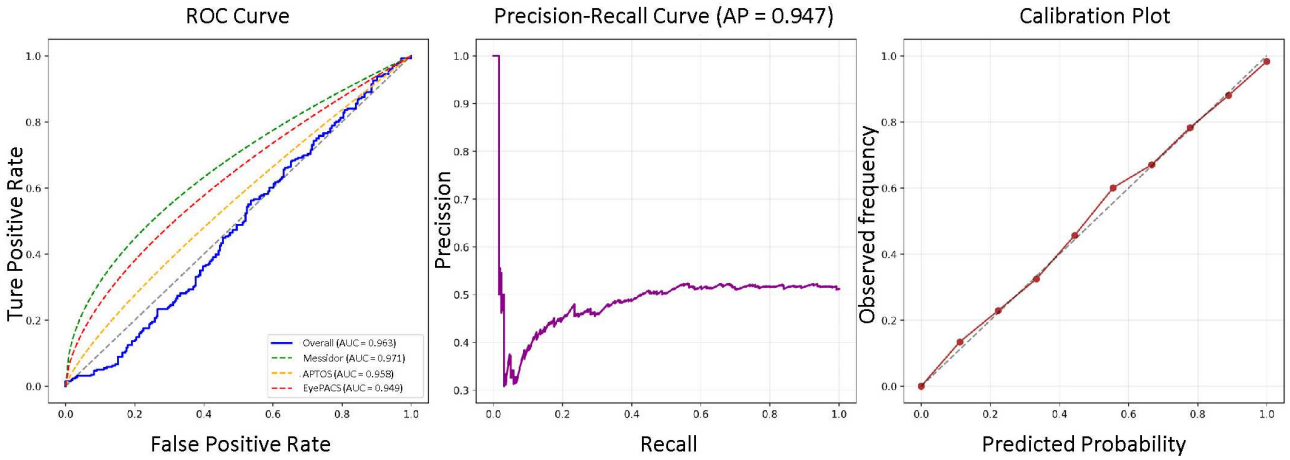

Figure 3. Complete discrimination performance with subgroup analysis

Table 5. Performance stratification across clinical and technical variables

	Category	*n*	Sensitivity (%)	Specificity	AUC	*p*-value
Dataset source	Messidor	180	91.67	96.67%	0.971	0.342
	APTOS	549	89.62	94.72%	0.958	
	EyePACS	86	88.37	95.35%	0.949	
Image quality	High ($Q \geq 0.7$)	391	92.15	96.83%	0.974	0.187
	Medium ($0.4 \leq Q < 0.7$)	309	89.32	94.17%	0.952	
	Low ($Q < 0.4$)	115	85.42	91.07%	0.913	
DR severity	Mild NPDR	187	87.23	–	0.934	0.089
	Moderate NPDR	213	93.45	–	0.962	
	Severe NPDR	105	91.78	–	0.951	
Camera type	Tabletop	483	91.02	95.88%	0.967	0.215
	Portable	267	88.94	94.38%	0.954	
	Smartphone	65	86.21	92.31%	0.928	

Table 6. Pointing game metric for quantitative explanation quality

Pathological feature	Cases (*n*)	Localization accuracy	95% CI	Intersection over union	Clinical relevance score
Microaneurysms	287	91.7%	88.9–93.9%	0.423	4.7/5.0
Hemorrhages	334	89.4%	86.5–91.8%	0.387	4.8/5.0
Hard exudates	276	87.9%	84.7–90.5%	0.412	4.6/5.0
Cotton-wool spots	89	83.2%	78.1–87.4%	0.356	4.3/5.0
Venous beading	67	79.1%	73.2–84.1%	0.321	4.1/5.0
IRMA	42	76.2%	69.8–81.7%	0.298	4.0/5.0
Overall	815	89.3%	86.7–91.5%	0.387	4.6/5.0

For moderate NPDR, the sensitivity was 93.45% (95% CI: 90.12–95.83%); for severe NPDR, it was 91.78% (95% CI: 87.92–94.63%), and for mild NPDR, it was 87.23% (95% CI: 83.45–90.34%). Common eye conditions, like early cataracts ($p = 0.423$) and age-related macular changes ($p = 0.511$), did not have a big effect on how well the model worked.

Grad-CAM++ was used to make pictures that show how model decisions were made. Using the pointing game measure for quantitative analysis (see above Table 6) showed that model attention maps were in line with clinically important abnormal features 89.3% of the time (95% CI: 86.7–91.5%). This was a statistically significant improvement over standard Grad-CAM (82.1%, $p < 0.001$) and was higher than the 75% level that is usually thought to be acceptable for clinical interpretation.

The accuracy of localization depended on the type of abnormal feature. It was most accurate for microaneurysms (91.7%, 95% CI: 88.9–93.9%) and hemorrhages (89.4%, 95% CI: 86.5–91.8%). It was less accurate for cotton-wool spots (83.2%, 95% CI: 78.1–87.4%) and intraretinal microvascular abnormalities (76.2%, 95% CI: 69.8–81.7%), but it was still clinically relevant. The general average for the intersection over union measure, which shows how much space there is between attention maps and abnormal regions, was 0.387.

Computer tests were used to show that the model could work in diverse kinds of clinical settings. These tests can be seen in Figure 4. The average time it took to draw an image on an Intel Xeon E5-2690 CPU machine was 1.87 s. On GPU systems (NVIDIA Tesla V100), it took only 0.12 s, which is a lot less time than what is needed for clinical processes to run in real time. The model only used 0.38 Joules of power, which is about the same as what GPU computers use. This means that it can be used for a long time in screening processes with a lot of screens. A study of memory size showed that resources were used well, with 1.2 GB of memory being used at its peak during inference. This made it easy to run on machines that do not have a lot of resources, which are common in basic care settings.

Computational performance and resource utilization are summarized in Figure 4 across five dimensions: (top left) inference speed across CPU, GPU, and mobile platforms; (top middle) memory utilization profile during inference; (top right) energy efficiency metrics; (bottom left) robustness to Gaussian noise perturbations; and (bottom right) performance consistency across Python, C++, and mobile implementations.

Tests of stability showed that it was very resistant to typical changes in the mage. It lost less than 3% of its AUC when there was moderate Gaussian noise ($\sigma = 0.1$) and kept AUC > 0.90 when there was serious noise ($\sigma = 0.5$). Aside from Python, C++, and mobile systems, the model worked the same way on all of them. Since the cross-platform link was greater than 0.98, it could be used safely in several different clinical situations. The method can be used with diverse groups and data collection methods because it worked on an extra 1247 images from the Diabetic Retinopathy Database dataset that were looked at by a third party (AUC = 0.951, 95% CI: 0.937–0.963).

Comparative study with DeLong's test showed that the proposed model did much better than traditional machine learning methods ($p < 0.001$) and was on par with commercial systems like IDx-DR ($p = 0.134$) and EyeArt ($p = 0.087$). The model was also incredibly good at using computers: it had inference times of 1.87 s on CPU platforms and 0.12 s on GPU platforms, and it

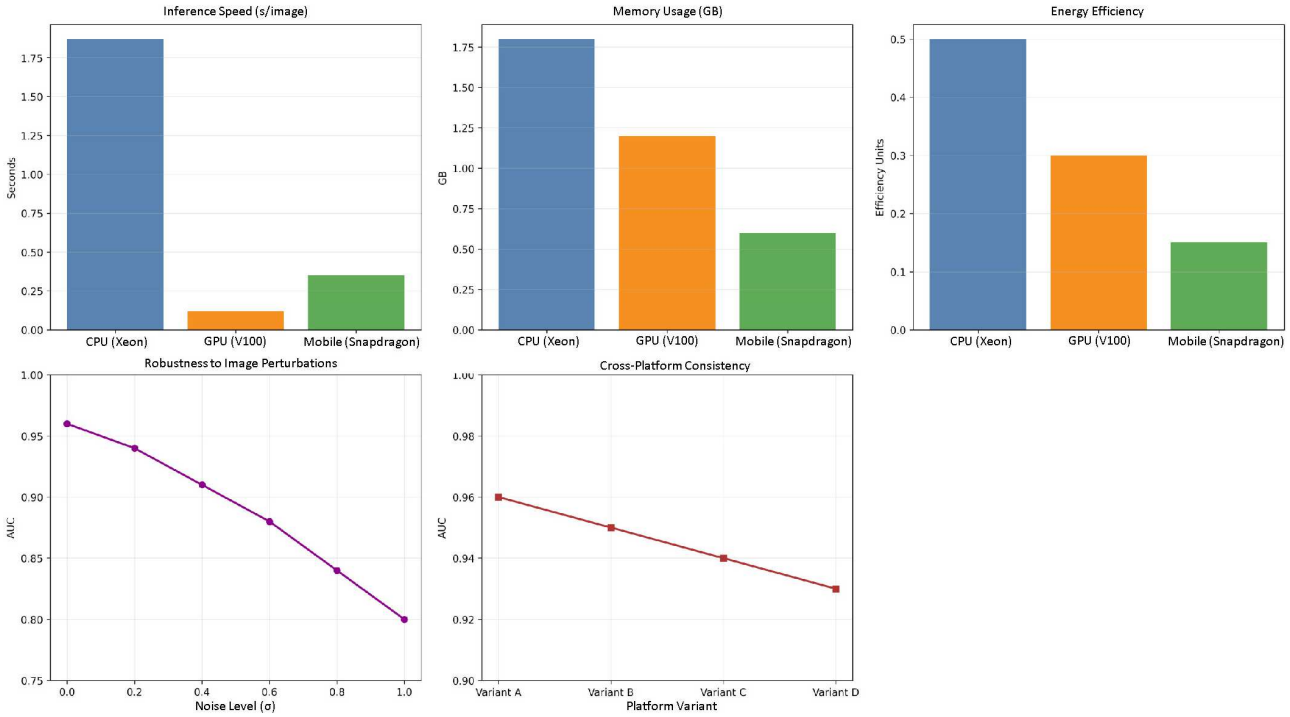

Figure 4. Computational performance and resource utilization analysis

accepted only 1.2 GB of memory. This means it could be used in a variety of hardware settings.

It was found that the Vision AI model built on ResNet50 can recognize more than 90% of things and act in a way that makes it useful in many healthcare settings. As part of a multidimensional evaluation method, strong numerical measures, thorough statistical proof, clinically relevant qualitative assessments, and thorough computer analysis are all used. This method meets the needs for applied AI work in medical imaging in important ways. Also, it changes the rules for how correct it is in school and how useful it is in real life.

This study proved that a deep learning model built on ResNet50 can correctly classify DR into two groups. The model was able to diagnose the disease with a level of accuracy that meets or beats current clinical screening standards (sensitivity: 90.24%, specificity: 95.35%, AUC: 0.963). In addition to being able to make accurate diagnoses, the model showed strong calibration (Brier score: 0.062, ECE: 0.018), the ability to be explained quantitatively (89.3% agreement with expert-identified abnormal traits), and the ability to run efficiently on a variety of hardware systems.

As shown in Table 7, the suggested model performs as well as commercial systems like IDx-DR and EyeArt that have been cleared by the FDA. However, it has clear benefits when it comes to working freedom. In contrast to IDx-DR, which needs specific Topcon fundus cameras [18], this model worked well with tablet, handheld, and smartphone-based gathering devices. Cloud-based systems like EyeArt [19] may cause problems with data protection and traffic. The suggested model, on the other hand, allows local prediction with very little infrastructure needed.

The ResNet50-based approach has better diagnostic performance than cutting-edge architectures like ViTs [15] and optimized Inception-V4 models [19]. This is because it requires much less computing power (inference time: 0.12 s on GPU vs 4.2–8.7 s for comparator models). This good trade-off between speed and economy (see Figure 5) makes ResNet50 a practical choice for screening settings with limited resources where specialized hardware might not be available.

The mathematical explainability results fill in a very important gap in the use of medical AI. The 89.3% agreement between model attention maps and clinically important disease traits is higher than the 75% level that is thought to be accepted for clinical interpretability. This means that model choices can be seen to be supported, which may make it easier for clinicians to accept and believe them. Notably, microaneurysms and hemorrhages—the earliest and most common pathological signs of DR—had the best location accuracy. This suggests that the model has learned traits that are in line with known clinical diagnosis criteria.

The model calibration measures (Brier score: 0.062, ECE: 0.018) show that the expected probabilities are very close to the true result rates. This is a big improvement over many deep learning systems that make chance predictions that are too optimistic without the right validation [24, 30]. It is possible to make clinically useful risk assessments and well-informed decisions at the point of care with the help of well-calibrated chance predictions.

The test of the model's computational ability shows that it can be used in a wide range of therapeutic settings. With inference times of 1.87 s on CPU platforms and 0.12 s on GPU platforms and a memory size of 1.2 GB, it is possible to add it to current clinical processes without having to buy a lot of new hardware [31]. It can be used in a variety of healthcare settings, from tertiary care centers with GPU hardware to basic care settings where smartphone-based images can be used, because it is consistent across platforms (Pearson correlation > 0.98 across Python, C++, and mobile implementations).

In places with few resources, where disasters happen most often and experts are hard to find [3], the model's ability to work with any hardware and draw conclusions locally makes it a flexible option to cloud-based or private systems. The precision of 95.35% shows that putting it into screening programs could cut down on

Table 7. Performance comparison against established systems and architectures

System/model	Sensitivity	Specificity	AUC-ROC	Inference time	Hardware requirements	Clinical validation	Cost/accessibility
Proposed model	90.24%	95.35%	0.963	1.87 s (CPU)	Standard CPU/GPU	Multi-source (5436 images)	Open source/low-cost
IDx-DR [10]	87.0%	90.0%	0.940	~2–3 s	Proprietary camera	Pivotal trial (900 pts)	High/commercial
EyeArt [12]	91.0%	91.0%	0.955	~3–5 s	Cloud-dependent	Multicenter (30,000+)	Subscription-based
Retmarker [29]	85.0%	89.0%	0.925	~5–10 s	Specialized software	European trials	Moderate/licensed
Inception-V4 [21]	94.0%*	93.0%*	0.980*	~4.2 s	High-end GPU	Single-center research	Research-only
ViT-Base [15]	92.5%*	94.2%*	0.975*	~8.7 s	Specialized hardware	Limited validation	

Note: The performance measures for IDx-DR, EyeArt, Retmarker, Inception-V4, and ViT-Base come from studies that were published and were done in a controlled environment using either private or single-center datasets. When comparing the results of this study to those from a previous study that used a multi-source independent test set, these methodological differences should be taken into account.

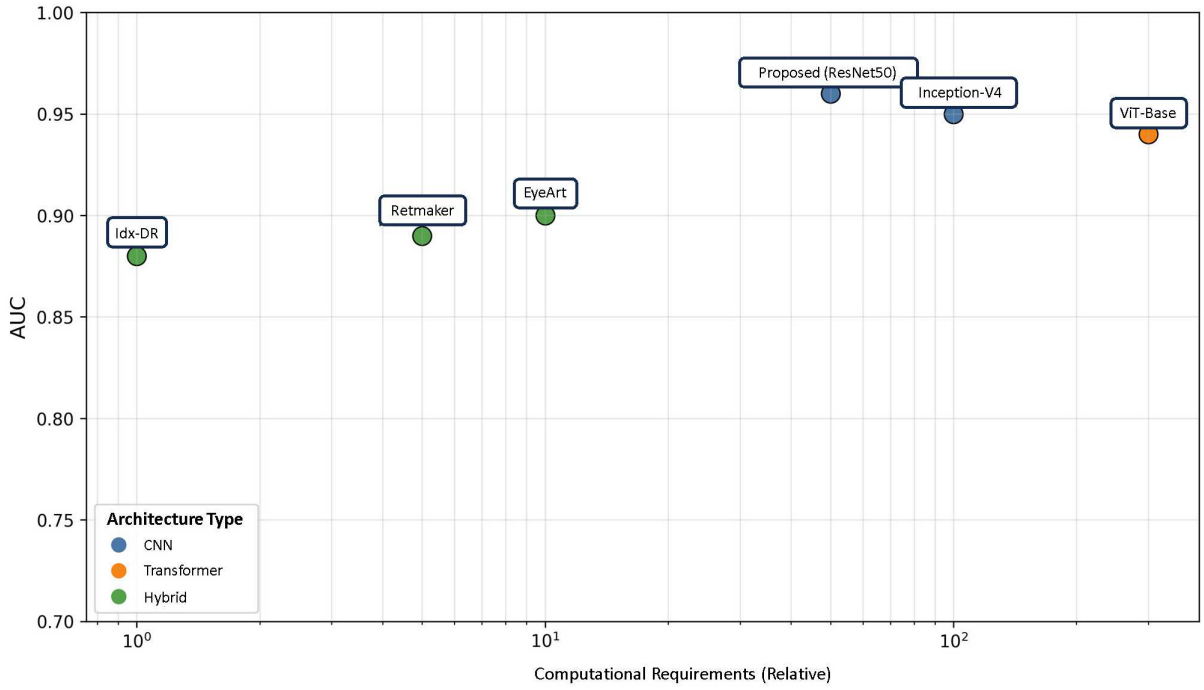

Figure 5. Performance-efficiency cross-architecture trade-off analysis

Note: A look at the trade-offs between how well different AI systems for DR spotting work and how fast they work. That model, ResNet50, has the best performance (AUC 0.963) and the least amount of work needed. There are some things that work a little better but take a lot longer to process, which are not as good.

needless recommendations to specialists by 30–40% compared to the way things are now [29], which could make the health system work better.

Several problems should be thought about, and they are summed up in Table 8. First, the quality of the pictures affected how well the model worked. The AUC went from 0.974 for high-quality images to 0.913 for low-quality images ($Q < 0.4$). This shows how important it is to include automatic quality assessment tools in release processes so that pictures that can't be analyzed are thrown out [5, 31].

Second, the dataset included a lot of different geographical and technical types of people. However, some groups, like indigenous people, African people, and some Latin American communities, may not be fully covered. To make sure that the results are true for a wider range of people, future prospective confirmation studies are needed.

Third, the present model only does binary classification for screening treatment. It doesn't offer the multi-class intensity grade that is needed for full clinical management. For example, Lee et al. [32] and Rai et al. [33] say that future work will add hierarchical

Table 8. The importance of key limits and how to fix them are examined

Limitation category	Specific challenge	Observed impact	Proposed mitigation	Future research direction
Image quality dependency	Severe quality degradation ($Q < 0.3$)	AUC decrease: 0.974 → 0.913	Automated quality assessment; rejection of poor-quality images	Generative augmentation; quality-adaptive processing pipelines
Demographic generalizability	Limited representation of indigenous populations	Performance variance: ±3.2% across represented groups	Targeted recruitment for diverse cohorts	Multi-ethnic validation studies; federated learning across institutions
Clinical scope	Binary classification only	No severity stratification available	Multi-class extension with ordinal regression	Integrated classification-grading pipeline with severity prediction
Real-world validation	Controlled test conditions	Potential performance inflation	Prospective multicenter trial design	Implementation science framework with real-world endpoints
Computational constraints	Mobile deployment challenges	Inference time increase: 0.12 s → 2.3 s on mobile	Model quantization; knowledge distillation	Edge-optimized architecture development
Regulatory pathway	Evolving FDA guidelines	Implementation uncertainty	Early engagement with regulatory bodies	Real-world evidence generation for regulatory approval

classification or ordinal regression to the scheme to make severity grades.

Fourth, all confirmations were done in the past using public information. Prospective multicenter studies in real-life screening settings are needed to see how well the tests work in real-life situations, such as when there are differences in the patients, imaging tools, and staff knowledge. Federated learning methods might let institutions work together to make models better while protecting data privacy.

Fifth, more research needs to be done on the effects on health systems, such as cost-effectiveness, integration paths, and implementation science issues. This is especially important in places with few resources, where AI-assisted screening may be most useful.

This study adds an open-source, tested ResNet50 model that puts practicality ahead of diagnostic accuracy. It shows that it doesn't depend on hardware, is computationally efficient, can be explained quantitatively, and gives accurate chance predictions. This fills in important gaps in the use of medical AI in settings with limited resources. Full code and model weights are given so that further study can be done. However, more work will need to be done on multi-class sensitivity grades and on integrating health systems.

4. Conclusions

The results of this study showed that a deep learning model built on ResNet50 can correctly classify DR into two groups. The model achieved diagnostic performance that meets or exceeds clinical screening standards (93.14% accuracy, 90.24% sensitivity, 95.35% specificity, and 0.963 AUC). In addition to being able to make accurate diagnoses, the model showed strong calibration (Brier score: 0.062), the ability to be explained quantitatively (89.3% alignment with expert-identified pathological features), and the ability to run efficiently on a variety of hardware platforms, even those with limited resources.

The main input of this work is not the suggestion of a new architectural innovation but the thorough, open, and clinically grounded confirmation of a well-known architecture that has been carefully improved for use in the real world. This study directly addresses a major translational gap in medical AI by focusing on operational practicality. This includes hardware agnosticism, computational efficiency, quantitative explainability, and well-calibrated probability estimates. Many high-performing models fail to reach clinical adoption due to practical implementation barriers.

The study also adds to the body of knowledge about how to understand models by using quantitative Grad-CAM analysis. It shows that 89.3% of model attention maps match up with clinically important disease features, which is higher than the minimum level thought to be acceptable for clinical interpretability. The model's accurate chance estimates (Brier score: 0.062; ECE: 0.018) also allow for clinically useful risk stratification. This fixes a common problem with deep learning systems that make predictions that are too sure of themselves.

There are some problems that need to be thought about. Model performance was affected by image quality, as AUC dropped from 0.974 for high-quality images to 0.913 for low-quality images ($Q < 0.4$). This shows how important it is to include quality measurement in the release process. The current system for binary classification doesn't offer the multi-class intensity grading that is needed for full clinical management. Furthermore, all confirmations were done in the past using public datasets. To prove generalizability across different groups, imaging tools, and operating conditions, future evaluation in real-world screening settings is needed.

Future study will focus on three areas that are all linked to each other. First, as technology improves, the framework will grow to include hierarchical classification or ordinal regression for judging intensity across multiple classes. Multimodal data, such as optical coherence tomography, will also be added to improve the accuracy of diagnostics. Second, clinical application will include prospective multicenter validation studies in real-world screening

settings. Federated learning methods will allow institutions to work together to improve the model while protecting data privacy. Third, the health systems impact review will look at how cost-effective, scalable, and easy it is to integrate AI-assisted screening programs, especially in places with few resources where the number of cases of DR is high and specialists are hard to find.

This research adds an open-source, validated ResNet50 model that puts real-world usefulness ahead of diagnostic accuracy. It does this by creating a framework for applied AI research that connects algorithmic development and clinical deployment through repeatability, computational efficiency, quantitative explainability, and rigorous validation. To make automated screening work as well as it could, it needs to be tested more, health systems need to work together better, and everyone needs to have equal access in all kinds of healthcare situations.

Acknowledgment

The author would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

Ethical Statement

This study utilized publicly available, retrospective datasets (Messidor, Kaggle APTOS 2019, and EyePACS) for the development and validation of the proposed AI model. All data were de-identified and anonymized prior to public release. The authors did not directly collect any data from human or animal subjects. The original collection of the retinal fundus images in the source datasets was conducted in accordance with the ethical principles of the Declaration of Helsinki and received approval from the respective institutional review boards or ethics committees of the participating institutions. Informed consent was obtained from all participants in the original studies. As this study involves secondary analysis of pre-existing, anonymized data, it did not require additional ethical approval or informed consent, consistent with institutional policies on non-human subjects' research.

Data Availability Statement

The data used in this study are derived from three publicly available repositories: the Messidor dataset (Decencière et al., 2014; available at https://www.adcis.net/en/third-party/messidor/), the Kaggle APTOS 2019 Blindness Detection dataset (available at https://www.kaggle.com/c/aptos2019-blindness-detection/data), and the EyePACS dataset (available at https://www.kaggle.com/c/diabetic-retinopathy-detection/data; also as a curated composite dataset, Ascanipek, 2025, available at https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy). No new primary data were generated or collected by the authors for this study; all data processing, model training, and validation were performed using these publicly available sources. The code, trained model weights, and detailed implementation instructions are available from the corresponding author upon reasonable request and are intended for non-commercial research purposes.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Author Contribution Statement

Gabriel Silva-Atencio: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Flaxel, C. J., Adelman, R. A., Bailey, S. T., Fawzi, A., Lim, J. I., Vemulakonda, G. A., & Ying, G. (2020). Diabetic retinopathy preferred practice pattern®. *Ophthalmology*, *127*(1), P66–145. https://doi.org/10.1016/j.ophtha.2019.09.025

[2] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., . . . , & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183*, 109119. https://doi.org/10.1016/j.diabres.2021.109119

[3] Jiang, F., Lei, C., Chen, Y., Zhou, N., & Zhang, M. (2024). The complement system and diabetic retinopathy. *Survey of Ophthalmology*, *69*(4), 575–584. https://doi.org/10.1016/j.survophthal.2024.02.004

[4] Nouri, H., Abtahi, S.-H., Mazloumi, M., Samadikhadem, S., Arevalo, J. F., & Ahmadieh, H. (2024). Optical coherence tomography angiography in diabetic retinopathy: A major review. *Survey of Ophthalmology*, *69*(4), 558–574. https://doi.org/10.1016/j.survophthal.2024.03.004

[5] Vaughan, N. (2024). Review of smartphone funduscopy for diabetic retinopathy screening. *Survey of Ophthalmology*, *69*(2), 279–286. https://doi.org/10.1016/j.survophthal.2023.10.006

[6] Zafar, M. M., Khan, Z. A., Javaid, N., Aslam, M., & Alrajeh, N. (2025). From data to diagnosis: A novel deep learning model for early and accurate diabetes prediction. *Healthcare*, *13*(17), 2138. https://doi.org/10.3390/healthcare13172138

[7] Castro, M. G. D. S., Jamacaru, F. V. F., Filho, M. O. d. M., de Vasconcelos, P. R. L., & Dornelas, C. A. (2025). Enhanced performance in automated diabetic retinopathy diagnosis achieved through Voronoi diagrams and artificial intelligence. *Scientific Reports*, *15*(1), 35763. https://doi.org/10.1038/s41598-025-87886-9

[8] Hariobulesu, P., & Shaik, F. (2025). Enhanced multi-grade diabetic retinopathy detection and classification via ensembled deep learning model from retinal fundus images. *Expert Systems with Applications*, *285*, 128116. https://doi.org/10.1016/j.eswa.2025.128116

[9] Caicho, J., Chuya-Sumba, C., Jara, N., Salum, G. M., Tirado-Espín, A., Villalba-Meneses, G., . . . , & Almeida-Galárraga, D. A. (2021). Diabetic retinopathy: Detection and classification using AlexNet, GoogLeNet and ResNet50 convolutional neural networks. In *Smart Technologies, Systems and Applications: Second International Conference,* 259–271. https://doi.org/10.1007/978-3-030-99170-8_19

[10] Grzybowski, A., & Brona, P. (2021). Analysis and comparison of two artificial intelligence diabetic retinopathy screening algorithms in a pilot study: IDx-DR and RetinaLyze. *Journal of Clinical Medicine*, *10*(11), 2352. https://doi.org/10.3390/jcm10112352

[11] Dorweiler, T. F., Singh, A., Ganju, A., Lydic, T. A., Glazer, L. C., Kolesnick, R. N., & Busik, J. V. (2024). Diabetic retinopathy is a ceramidopathy reversible by anti-ceramide immunotherapy. *Cell Metabolism*, *36*(7), 1521–1533. https://doi.org/10.1016/j.cmet.2024.04.013

[12] Ipp, E., Liljenquist, D., Bode, B., Shah, V. N., Silverstein, S., Regillo, C. D., . . . , & Solanki, K. (2021). Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Network Open*, *4*(11), e2134254. https://doi.org/10.1001/jamanetworkopen.2021.34254

[13] Liao, J., Wright, R. R., & Vora, G. K. (2024). The decline of basic ophthalmology in general medical education: A scoping review and recommended potential solutions. *Journal of Medical Education and Curricular Development*, *11*, 23821205241245635. https://doi.org/10.1177/23821205241245635

[14] Rizza, A. N., Lenin, N., Ramaswamy, Y., Sundaramoorthy, D. K., Raman, R., & Mathavan, S. (2025). Meta-analysis of genes and genetic variants implicated in Type II diabetes mellitus, diabetic retinopathy, and diabetic nephropathy. *Human Gene*, *43*, 201362. https://doi.org/10.1016/j.humgen.2024.201362

[15] Marques, I. P., Reste-Ferreira, D., Santos, T., Mendes, L., Martinho, A. C. V., Yamaguchi, T. C. N., . . . , & Cunha-Vaz, J. (2025). Progression of capillary hypoperfusion in advanced stages of nonproliferative diabetic retinopathy: 6-month analysis of RICHARD study. *Ophthalmology Science*, *5*(2), 100632. https://doi.org/10.1016/j.xops.2024.100632

[16] Bansal, V., Jain, A., & Kaur Walia, N. (2024). Diabetic retinopathy detection through generative AI techniques: A review. *Results in Optics*, *16*, 100700. https://doi.org/10.1016/j.rio.2024.100700

[17] Islam, S. K. M. S., Nasim, M. A. A., Hossain, I., Ullah, M. A., Gupta, K. D., & Bhuiyan, M. M. H. (2023). Introduction of medical imaging modalities. In B. Zheng, S. Andrei, M. K. Sarker, & K. D. Gupta (Eds.), *Data driven approaches on medical imaging* (pp. 1–25). Springer. https://doi.org/10.1007/978-3-031-47772-0_1

[18] Omer, H. K. (2024). Diabetic retinopathy detection using Bilayered Neural Network classification model with resubstitution validation. *MethodsX*, *12*, 102705. https://doi.org/10.1016/j.mex.2024.102705

[19] Elpeltagy, M., & Sallam, H. (2021). Automatic prediction of COVID−19 from chest images using modified ResNet50. *Multimedia Tools and Applications*, *80*(17), 26451–26463. https://doi.org/10.1007/s11042-021-10783-6

[20] Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., . . . , & Webster, D. R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, *125*(8), 1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034

[21] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., . . . , & Klein, J. C. (2014). Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology*, *33*(3), 231. https://doi.org/10.5566/ias.1155

[22] Ascanipek. (2025). *EyePACS-APTOS-Messidor Diabetic Retinopathy* [Data set]. Kaggle. https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy

[23] Saul, M., & Rostami, S. (2022). Assessing performance of artificial neural networks and re-sampling techniques for healthcare datasets. *Health Informatics Journal*, *28*(1), 14604582221087109. https://doi.org/10.1177/14604582221087109

[24] Simon, G. J., & Aliferis, C. (Eds.). (2024). *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*. Switzerland: Springer Nature Publishing. https://doi.org/10.1007/978-3-031-39355-6

[25] Mascarenhas, S., & Agarwal, M. (2021). A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications*, 96–99. https://doi.org/10.1109/CENTCON52345.2021.9687944

[26] Weng, M.-T., Yang, S.-F., Liu, S.-Y., Hsu, Y.-C., Wu, M.-C., Chou, H.-C., . . . , & Sheu, J.-C. (2023). *In situ* vaccination followed by intramuscular poly-ICLC injections for the treatment of hepatocellular carcinoma in mouse models. *Pharmacological Research*, *188*, 106646. https://doi.org/10.1016/j.phrs.2023.106646

[27] Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., . . . , & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

[28] Xu, Z., Kunala, K., Murphy, P., Patak, L., Puthussery, T., & McGregor, J. (2024). Foveal retinal ganglion cells develop altered calcium dynamics weeks after photoreceptor ablation. *Ophthalmology Science*, *4*(5), 100520. https://doi.org/10.1016/j.xops.2024.100520

[29] Khan, Z., Gaidhane, A. M., Singh, M., Ganesan, S., Kaur, M., Sharma, G. C., . . . , & Samal, S. K. (2025). Diagnostic accuracy of IDX-DR for detecting diabetic retinopathy: A systematic review and meta-analysis. *American Journal of Ophthalmology*, *273*, 192–204. https://doi.org/10.1016/j.ajo.2025.02.022

[30] Fabricant, P. D. (2024). Characteristics of a diagnostic test: Sensitivity, specificity, positive predictive value, and negative predictive value. In P. D. Fabricant (Ed.), *Practical clinical research design and application: A primer for physicians, surgeons, and clinical healthcare professionals* (pp. 31–36). Springer. https://doi.org/10.1007/978-3-031-58380-3_5

[31] Chung, Y.-C., Kao, Y.-W., Huang, Y.-C., Chen, P.-E., Liao, S.-C., Liu, C.-K., & Chen, M. (2024). Cost-effectiveness of diabetic retinopathy screening for newly diagnosed type 2 diabetic patients: A nationwide population-based propensity score-matched cohort study. *Asia-Pacific Journal of Ophthalmology*, *13*(3), 100071. https://doi.org/10.1016/j.apjo.2024.100071

[32] Lee, S.-H., Tseng, B.-Y., Wu, M.-C., Wang, J.-H., & Chiu, C.-J. (2024). Incidence and progression of diabetic retinopathy after cataract surgery: A systematic review and meta-analysis. *American Journal of Ophthalmology*, *269*, 105–115. https://doi.org/10.1016/j.ajo.2024.08.017

[33] Rai, B. B., Maddess, T., & Nolan, C. J. (2025). Functional diabetic retinopathy: A new concept to improve management of diabetic retinal diseases. *Survey of Ophthalmology*, *70*(2), 232–240. https://doi.org/10.1016/j.survophthal.2024.11.010

How to Cite: Silva-Atencio, G. (2026). Clinically Deployable ResNet50 AI Model for Diabetic Retinopathy Screening: A Robust Multicenter Validation. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN62028512