NUST

NATIONAL UNIVERSITY
OF SCIENCES & TECHNOLOGY

Retinal Image Analysis using Computer Vision and Edge Computing on Jetson Nano

Course : Computer Vision (CS - 867)

Submitted by:

Ammar
(537505)

Waleed Aftab
(494112)

Manahil Sheikh
(537401)

Submitted To:

Dr. Tauseef ur Rheman

National University of Science And Technology Islamabad

School of Electrical Engineering and Computer Science

October 2025

Contents

Abstract

Diabetic Retinopathy (DR) is among the most frequent causes of blindness globally, and early diagnosis is essential in avoiding irreversible vision loss. The aim of this project is to design an effective, edge-based retinal image analysis system based on computer vision and deep learning methods. The solution is based on lightweight convolutional neural network models, such as MobileNetV2 and EfficientNet-B0, to automate retinal disease diagnosis and severity level classification.

It has a two-stage pipeline: the first one detects whether the input image is a valid retinal image or not, and the second model classifies the retinal image into various stages of DR from normal to proliferative. In order to fulfill real-time processing and on-device inference, models are quantized and optimized using TensorRT for deployment on NVIDIA's Jetson Nano platform.

A web application based on Flask acts as the front-end interface where users can upload retina images or take live inputs using an attached camera. This arrangement mimics a low-cost diagnostic system where analysis happens locally in a cloud-free environment, increasing speed and data privacy.

The anticipated result is a small, deployable DR detection system exhibiting proper classification, effective performance on embedded hardware, and an interactive live demonstration in accordance with the overall vision of making AI-enabling healthcare affordable at the edge.

1. Introduction

1A. Background and Problem Statement

Diabetic Retinopathy (DR) is a major cause of avoidable blindness, occurring in a large percentage of people with diabetes globally. Early diagnosis and timely treatment are critical to avoid permanent vision loss. Yet conventional diagnostics involve ophthalmologists visually examining retinal fundus images, which takes a lot of time, is subjective, and usually not available in low-resource or remote settings.

With the imminent advancement in deep learning and computer vision, computer-based systems have emerged to grade and classify DR from retinal images. Light-weight neural network models like EfficientNet and MobileNet have achieved high accuracies for detecting disease on large public databases. Though they work well, they all rely on cloud computing or high-end GPUs, which might limit their application in real-time screening in low-resource healthcare settings.

Edge Artificial Intelligence (Edge AI) offers a real-world solution by allowing local inference on small devices like the NVIDIA Jetson Nano. This method guarantees low latency, data privacy, and offline capability well-suited for clinical and distant healthcare environments. However, running sophisticated deep learning models on embedded platforms is confronted with challenges in finding a balance between computational efficacy, precision, and energy intake.

To overcome these limitations, this project suggests a two-stage DR detection pipeline for edge deployment. The first stage checks whether the image is a retinal image or not, and the second stage determines its DR severity level. The models will be fine-tuned for lightweight performance and deployed on the Jetson Nano platform with an interlinked Flask-based web interface for real-time and interactive screening. This design is intended to illustrate a portable, affordable, and privacy-friendly diagnostic system for early detection of Diabetic Retinopathy.

1B. Motivation

The worldwide increase in diabetes set to hit more than 700 million individuals by 2045 has fueled the demand for scalable, affordable DR screening technology. Expert ophthalmologists and specialized imaging devices are traditionally needed for diagnosis, constricting availability in rural or under-resourced regions.

Advances in embedded and lightweight deep learning now allow for processing of medical images locally on small hardware. The NVIDIA Jetson Nano, with TensorRT optimization, allows for effective inference with cloud independence, decreasing latency, power usage, and security threats.

Inspired by these advancements, this project seeks to develop an deployable, affordable DR detection system based on optimized deep learning models on edge platforms enabling real-time and offline screening in local healthcare settings.

1C. Project Objectives

The project aims to design and deploy a complete edge-based DR detection system. Key objectives include:

- Develop a two-stage pipeline for image validation and DR severity classification using MobileNetV2 and EfficientNet-B0.
- Optimize models with quantization and TensorRT for real-time inference on the NVIDIA Jetson Nano.
- Integrate the models into a Flask-based web interface for image upload, visualization, and live testing.
- Demonstrate performance through a live camera-based setup and measure inference accuracy, latency, and FPS.

1D. Scope of Work

The project focuses on the design, optimization, and deployment of DR detection models on edge hardware.

- **Included:** Training lightweight models using public datasets (APTOS/EyePACS), TensorRT optimization, Flask integration, and Jetson-based live testing.
- **Excluded:** Clinical data collection, cloud or federated learning, and medical-grade validation.

This scoped approach ensures technical feasibility within the semester while emphasizing practical learning in computer vision, deep learning, and embedded AI deployment.

2. Literature Review

2A. Recent Works (2022–2025)

In nnMobileNet[2], built on MobileNetV2's inverted linear bottleneck (ILRB) blocks, uses several parametric modifications namely stage-wise channel configuration, data augmentation, dropout, optimizer (AdamP), and activation function. In stage-wise channel configuration, the channel widths of each ILRB layer were adjusted. The authors tested different data augmentation levels and found that the heaviest augmentation yielded the best performance. nnMobileNet employs spatial dropout to randomly drop entire feature channels, while the Adam optimizer enhanced convergence and performance.ReLU6 replaced ReLU in each ILRB for stable gradient flow without vanishing activations. Tested on five datasets, nnMobileNet outperformed all, achieving state-of-the-art results across Messidor-1 (AUC 98.7%), IDRiD (AUC 91.6% for DR, 95.3% for DME), APOTS (ACC 89.1%, Kappa 93.4%), and RFMiD (ACC 94.4%, AUC 98.7%) with fewer parameters, and demonstrated 3–8$\times$ faster CPU inference than complex ViT models in MMAC 2023.

This [3], proposes the use of EfficientNet-B3 for classification of retinal images. Using the pre-trained ImageNet weights, the authors employ transfer learning and fine-tuning the model on a specialized retinal dataset.Advanced preprocessing and training strategies are used, including image resizing, data augmentation, and data generators for batch loading. The authors use EfficientNet-B3 as base model and add a fully connected (Dense) layer with 256 neurons (using ReLU activation) , a dropout layer to prevent overfitting, and an output dense layer with four neurons and softmax activation for multiclass classification. The model also incorporates batch normalization layers . The final model achieved 95% overall accuracy on the test set. The macro-averaged F1-score across all classes was also 94%.

In [4] , the authors propose the classification of Optical Coherence Tomography (OCT) images.. The images are normalized and preprocessed to highlight the differences in retinal layers in OCT images. The authors propose 2 models; one with three binary (CNN) classifiers and the other with four binary CNN classifiers. Several CNNs, such as VGG16, VGG19, etc are adapted as feature extractors . Following the CNN feature extractor, a global average pooling layer, dense layer, dropout layer, dense layer, dropout layer, and output softmax layer are added in this sequence. Among them, the model using four VGG19 classifiers shows the best performance with 0.9780 accuracy. When we use the best four classifiers (VGG16 for Classifier-CNV and Classifier-DME, VGG19 for Classifier-Drusen, and InceptionV3 for Classifier-Normal), the proposed Model 2 shows the best performance with 0.9870 accuracy.

In [5], the authors propose the classification of retinal diseases in OCT scans of retina, using a CNN with different number of layers. For image preprocessing, the images are passed through median filter to remove noise followed by a Contrast Limited Adaptive Histogram Equalization (CLAHE) for contrast enhancements and morphological operation, thresholding, and contour-based edge detection for retinal layer extraction. Three models are based on different numbers of convolutional layers, max-pooling, and fully connected dense layers with 40% dropout rate to avoid over-fitting. The first model consists of five CNN layers;one input layer with three channels and four hidden layers, all using ReLU activation. The second model has seven layers organized into four CNN-based blocks for improved feature extraction, each followed by a 2×2 max pooling layer. The third model, with nine CNN layers arranged in five blocks, produces a 64×4×4 output that is fed into fully connected dense layers.Of these models, The seven-layer CNN model outperformed the other 2 and has an accuracy of 96.5%.

In [6], the authors propose a convolutional EyeDeep-Net model for multi-eye disease detection. The architecture includes 17 weighted layers (14 convolutional, 2 fully connected, and 1 classification layer) with batch normalization, max pooling, dropout, and flattening. Using the RFMiD dataset, the study follows three steps: data acquisition and labeling, preprocessing and augmentation to handle class imbalance, and neural network training. The model achieved 82.13% validation accuracy and 76.04% testing accuracy, outperforming state-of-the-art models. This is the classical and one of the first papers which detects diabetic retinopathy detection from fundus photos. DR is caused due to high blood sugar, and it damages the blood vessels in the

eye. The author utilized two datasets; EyePACS-1, and MESSIDOR-2 and trained 10 different inception-v3. The dataset was labelled with four categories starting from no DR to proliferative DR. The result is the probabilistic output from all 10 models. The accuracy achieved was above 90% [7]. The authors deployed a retinal disease algorithm on Jetson Nano. Due to computation and power constraints, the actual model that uses vision transformers can't be deployed on Jetson. The authors utilize a teacher-student approach. They trained their own CNN model which understands the behavior of the VIT model. A small checker model (discriminator) was there which reduces the error. The CNN model was low in parameters as compared to the VIT model. However, the accuracy dropped from 97% to 89% [8].

The author discussed the computational efficiency required for running a U-net model for image segmentation. This paper does not show any new approach but calculates inference time with or without hardware accelerator. On average SBC, the inference time was hundreds of milliseconds. However, with the hardware accelerator (Edge TPU), latency dropped to 25ms. The latency dropped drastically to a few milliseconds when desktop/cloud GPU was involved [9]. Vujosevic, S., Limoli, C., & Nucci, P developed a low cost solution mainly for clinics or low resource areas. They developed a DR detector without the requirement of high end GPUs. They used a simple camera for image capturing and then for pre-processing, they used histogram equalization, resizing (512×512), and color normalization. A lightweight CNN model (based on mobilenet) was used. The system was then deployed on edge devices like Jetson Nano and Raspberry Pi. The accuracy of the model was 93% and the inference time was 150–200 ms per image [10]. The authors curated a dataset of 3662 fundus images and used it as a dataset to train different models. They decided to use four different architectures; MobileNet, ShuffleNet, SqueezeNet, a custom Deep Neural Network (DNN). Initially, these models were trained using tensorflow and then converted to tensorflow lite. The models were quantized to 8-bit integer format to reduce size and increase inference speed. The highest accuracy was achieved by MobileNet which was 96%. SqueezeNet, which was a much smaller model (176KB) showed low latency speed of 17ms, but the accuracy was compromised [11].

2B. Literature Review Summary Table

TABLE I
SUMMARY OF RECENT RESEARCH ON DR DETECTION MODELS

Paper	Year	Model	Dataset	Accuracy	Contribution	Limitation
nnMobileNet: Rethinking CNN for Retinopathy Research	2024	nnmobilenet	Messidor-1, IDRiD, APOTS, RFMiD	Up to 98.7% (AUC)	Introduced stage-wise channel configuration, spatial dropout, and ReLU6; achieved SOTA across multiple datasets and 3–8× faster inference	No mention of real-time clinical validation or hardware deployment
Retinal Disease Classification using EfficientNet-B3	2023	EfficientNet-B3	Retinal dataset-Kaggle	95%	Used transfer learning with ImageNet weights, data augmentation, and batch normalization for robust classification.	Lack of variability in dataset
OCT CNN Model	2022	CNN	OCT Images	98.7%	Proposed multi-model CNN ensemble for OCT classification achieving high accuracy	expensive due to multiple classifiers.
DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases	2022	CNN (7-layer)	OCT Scans	96.5%	Combined CLAHE, morphological preprocessing, and CNN-based feature extraction for retinal disease detection.	Limited testing; primarily evaluated on OCT datasets only.
EyeDeep-Net: a multi-class diagnosis of retinal diseases using deep neural network	2023	EyeDeep-Net	RFMiD	82.13% (val), 76.04% (test)	Developed 17-layer CNN with batch normalization and augmentation for multi-disease detection.	lower performance compared to specialized DR models.

3. Proposed System Design

3A. Overview

The proposed system aims to develop an intelligent and efficient framework for the detection and classification of Diabetic Retinopathy (DR) from retinal fundus images using deep learning. The design emphasizes algorithmic development, simulation-based evaluation, and feasibility for embedded deployment. The system is conceptually divided into two layers: an **algorithmic layer** focused on model design and validation, and a **deployment layer** targeting real-time operation on the NVIDIA Jetson Nano.

3B. System Architecture

At a high level, the pipeline will integrate four main modules: (1) data acquisition and preprocessing, (2) model training and validation, (3) performance simulation, and (4) on-device inference. A block diagram (to be added later) will illustrate this flow, beginning from input image collection and ending with the classification output displayed through a local interface. The modular design ensures scalability and smooth transition from simulation to deployment.

3C. Algorithmic Workflow

The system employs a two-stage classification strategy to enhance accuracy and reduce false detections.

3C.1. Stage 1: Retina vs. Non-Retina Validation

The first model will automatically verify whether the input image represents a valid retinal fundus. A lightweight convolutional network will be fine-tuned for this binary classification task, ensuring that only relevant images proceed to the next stage.

3C.2. Stage 2: Retinal Diseases Classification

Validated retinal images will then be passed to the second model, which classifies the presence of retinal diseases such as vitreous degeneration,drusen etc.

3D. Dataset Preparation and Simulation Plan

Publicly available datasets such as *APTOS 2019 Blindness Detection* , *EyePACS* and *Kaggle datasets* will be utilized for algorithm development and simulation. Images will undergo normalization, resizing, and basic augmentation (rotation, flipping, brightness adjustment) to improve generalization. Simulation and model validation will initially be carried out on a GPU-based environment (e.g., Google Colab) to evaluate training performance, accuracy, and confusion matrices before deployment.

3E. Model Optimization and Edge Feasibility

After simulation, both models will be optimized using TensorRT and quantization techniques to reduce computational load and memory footprint. These optimizations will be tested for inference speed and latency through simulation to ensure smooth performance on embedded hardware. The goal is to achieve real-time classification capability on the Jetson Nano while maintaining diagnostic reliability.

3F. Summary

In summary, this proposed design focuses on creating a reliable and lightweight two-model pipeline. Initial phases will involve simulation-based evaluation and performance validation, followed by optimization for embedded deployment. The final system aims to provide a deployable, low-cost, and efficient DR detection framework suitable for practical field use.

4. Edge Deployment and Demonstration Setup

4A. Overview

The edge deployment phase focuses on transforming the trained deep learning models into a working prototype that can operate autonomously on embedded hardware. The NVIDIA Jetson Nano is selected as the primary edge platform due to its CUDA-enabled GPU, compact design, and efficient power consumption. The system aims to demonstrate real-time diabetic retinopathy (DR) detection from retinal images using an optimized two-model inference pipeline.

4B. System Workflow

The overall workflow, illustrated in Figure 1, represents the sequential process of image acquisition, preprocessing, model training, and performance evaluation. Captured retinal images are preprocessed through resizing, noise removal, and augmentation before being passed to the training module. The models are then evaluated using key metrics such as accuracy, specificity, and sensitivity.

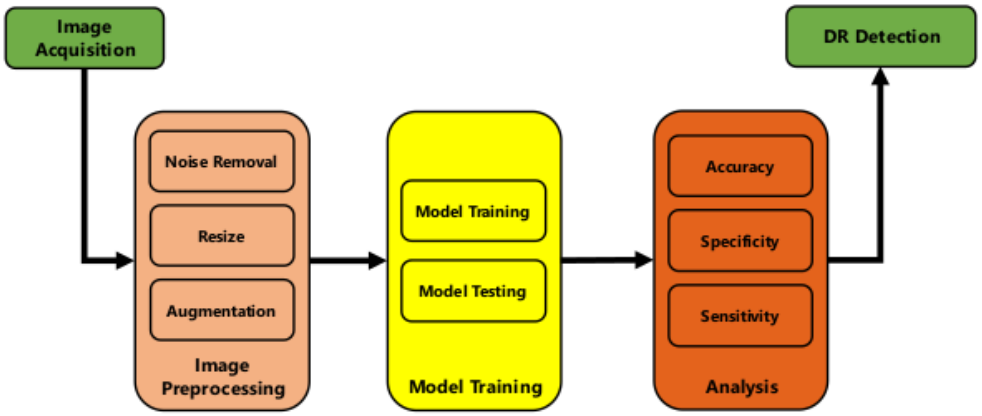

Fig. 1. Overall system workflow from image preprocessing to model evaluation.

4C. Jetson Nano Integration

Once the models are trained and optimized (via TensorRT quantization), they are deployed to the Jetson Nano for local inference. The device performs both stages of classification — Retina vs Non-Retina and DR Severity Detection — directly on-device, eliminating the need for cloud computation. This ensures:

- Low-latency inference (10–15 FPS for MobileNetV2, 5–8 FPS for EfficientNet-B0)
- Improved data privacy as no images leave the device
- Portability and suitability for real-world screening environments

4D. Prototype Architecture

The real-time inference pipeline on Jetson Nano is shown in Figure 2. The image captured by a camera or uploaded through the web interface is first processed by Model 1 (Retina/Non-Retina). If valid, it proceeds to Model 2 for retinal diseases classification. The Flask application acts as an interface for visualization and interaction.

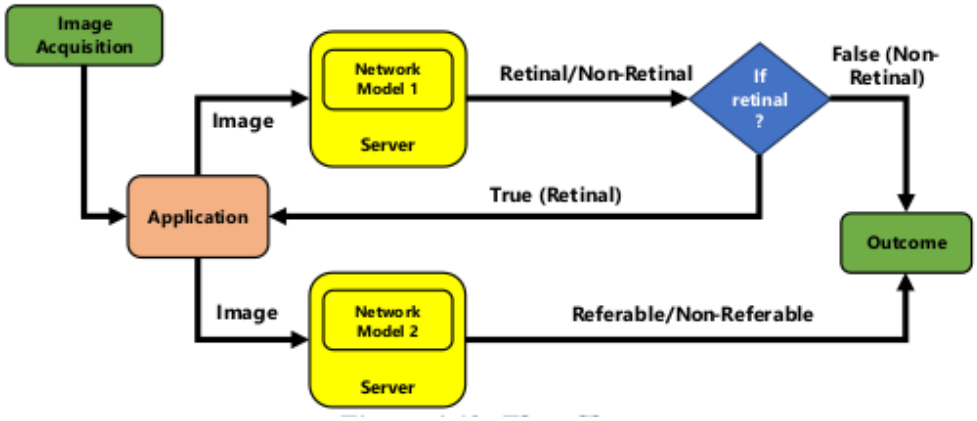

Fig. 2. Two-stage classification workflow integrated with Jetson Nano and Flask application.

4E. Flask Web Interface and Demonstration Setup

The Flask-based interface provides an accessible demonstration layer for the prototype. It allows:

- Image upload or live camera capture
- Real-time display of prediction labels and confidence scores
- Visualization of classification workflow running locally on Jetson Nano

The live demonstration setup includes the following hardware:

- NVIDIA Jetson Nano Developer Kit
- USB or Pi Camera Module (with macro lens for close-up retinal capture)
- Display monitor and local server interface

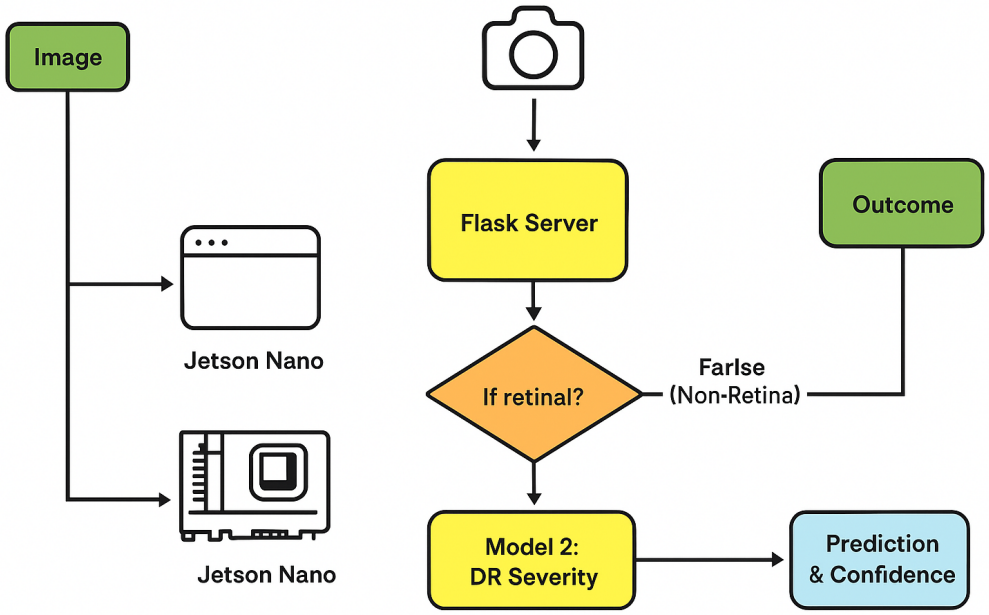

Fig. 3. Top-level system architecture showing Jetson Nano as the central node hosting the Flask server. The user can either upload a retinal image or capture one through the connected camera. The Flask application processes the input, passes it to the two-stage deep learning models (Retina validation and DR classification), and displays the diagnostic output in real time through a web interface.

This prototype simulates a low-cost, portable screening system where all computation is performed locally, making it suitable for remote or resource-constrained medical environments.

4F. Feasibility Summary

The proposed edge deployment is both technically and operationally feasible within the project scope. Jetson Nano offers sufficient GPU capabilities for lightweight inference, while TensorRT ensures optimized model execution. The modular design, combining the Flask application and embedded GPU, supports scalability and future extension toward clinical-grade screening systems.

5. Role Allocation

Member	Role	Responsibilities
Manahil Shaikh	Algorithms & Simulation	Model training, dataset handling

Waleed Aftab	Research & Documentation	Literature review, report writing
Ammar	Embedded Systems & Jetson Nano	Model optimization, Flask app, deployment

References

[1] A. Asare, D. Agyemanh, N. Gookyi, D. Boateng, and F. Aabangbio Wulnye, “Deploying and Evaluating Multiple Deep Learning Models on Edge Devices for Diabetic Retinopathy Detection,” *arXiv preprint arXiv:2506.14834*, 2025.

[2] Zhu, W., Qiu, P., Chen, X., Li, X., Lepore, N., Dumitrascu, O. M., Wang, Y. (2024). nnMobileNet: Rethinking CNN for Retinopathy Research. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2024), 2285-2294. doi:10.1109/CVPRW63382.2024.00234.

[3] M. S. Arshad Hussain, S. Babu, M. K. Sri Satya Sai, K. Siddartha and K. B. Naik, ”Retinal Disease Classification using EfficientNet-B3,” 2024 4th International Conference on Soft Computing for Security Applications (ICSCSA), Salem, India, 2024, pp. 337-344, doi: 10.1109/ICSCSA64454.2024.00060. keywords: Deep learning;Glaucoma;Cataracts;Diabetic retinopathy;Transfer learning;Data preprocessing;Computer architecture;Retina;Feature extraction;Medical diagnostic imaging;Retinal image classification;EfficientNet;Transfer learning;Medical Imaging;Deep Learning;Diabetic Retinopathy;Glaucoma;Cataract;Disease Classification,

[4] J. Kim and L. Tran, ”Retinal Disease Classification from OCT Images Using Deep Learning Algorithms,” 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Melbourne, Australia, 2021, pp. 1-6, doi: 10.1109/CIBCB49929.2021.9562919. keywords: Deep learning;Adaptation models;Sensitivity;Optical coherence tomography;Retina;Feature extraction;Classification algorithms;Optical Coherence Tomography (OCT);Retinal Disease;Deep Learning;Convolutional Neural Networks (CNN);Fully Convolutional Neural Networks (FCN),

[5] Tayal, A., Gupta, J., Solanki, A. et al. DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases. Multimedia Systems 28, 1417–1438 (2022). https://doi.org/10.1007/s00530-021-00769-7

[6] Sengar, N., Joshi, R.C., Dutta, M.K. et al. EyeDeep-Net: a multi-class diagnosis of retinal diseases using deep neural network. Neural Comput & Applic 35, 10551–10571 (2023). https://doi.org/10.1007/s00521-023-08249-x

[7] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[8] Yilmaz, B., & Aiyengar, A. (2025). Cross-architecture knowledge distillation (KD) for retinal fundus image anomaly detection on NVIDIA Jetson Nano. arXiv preprint arXiv:2506.18220. https://arxiv.org/abs/2506.18220

[9] Civit-Masot, J., Luna-Perejón, F., Corral, J. M. R., Domínguez-Morales, M., Morgado-Estévez, A., Civit, A. (2021). A study on the use of Edge TPUs for eye fundus image segmentation. Engineering Applications of Artificial Intelligence, 104, 104384.

[10] Vujosevic, S., Limoli, C., & Nucci, P. (2024). Novel artificial intelligence for diabetic retinopathy and diabetic macular edema: what is new in 2024?. Current opinion in ophthalmology, 35(6), 472–479. https://doi.org/10.1097/ICU.0000000000000108

[11] Asare, A., Gookyi, D. A. N., Boateng, D., & Wulnye, F. A. (2025). Deploying and evaluating multiple deep learning models on edge devices for diabetic retinopathy detection. arXiv. https://arxiv.org/abs/2506.14834