Journal of Health Informatics in Africa

Rapid Retinopathy Detection using Ablation-Guided Deep Learning

Mohammad Sharifur Rahman [a*], Khondoker Shaila Sharmin [b]

[a] Ulster University, Northern Ireland, United Kingdom
[b] Sir Salimullah Medical College, Dhaka, Bangladesh

Background and Purpose: The Prevalent cause of vision loss is diabetic retinopathy (DR) worldwide, impacting millions of people, and is projected to increase dramatically in countries in Asia and Africa. Early and prompt detection is crucial for reducing the likelihood of complications; However, manual detection is labour-intensive, resource-intensive, and have chance of human error. Deep learning (DL) and machine learning (ML) have improved automatic DR detection.; however, published literature also exists on ablation studies that compare combinations of DL feature extractors and traditional classifiers. The current study intends to systematically compare different DL models: VGG16, ResNet152V2 and Xception, combined with classical classifiers: Decision Tree (DT), k-Nearest Neighbours (k-NN), Artificial Neural Networks (ANN), and Random Forest (RF) to see which produces the most optimal model for DR detection. The study systematically compares different combinations to explore the potential use of hybrid DL-ML pipelines for improving DR classification performance and reducing computational costs, ultimately contributing to the development of an automated DR screening pipeline.

Methods: Using a dataset from EyePACS, 88.29GB in size, which included 35,126 labelled fundus images and 53,576 unlabelled fundus images. The images were sized for specific models (224×224 pixels for VGG16/ResNet152V2; 290×290 for Xception) and converted to greyscale. Three pretrained DL models (VGG16, ResNet152V2, Xception) were used for the feature extraction part of the process, and four different ML classifiers (RF, DT, K-NN, ANN). In total, there were twelve different combinations of models to extract features and classifiers to classify them. Using the available GPU power for the accelerator, the model was trained, validated, and tested in Google Colab. Confusion matrix-derived accuracy, recall, precision, and F1-score data were used to gauge the models' performance. The training and validation process involved optimising the hyperparameters and epochs (up to 30 epochs). An early stopping criterion was also added to all models to reduce the scope of overfitting.

Results: The hybrid Xception-RF model had the highest testing accuracy at 91.1% and the fully connected CNN models reached a maximum of 79.0%. VGG16-RF achieved a comparable accuracy (90.3%). Fully connected models achieved reasonably consistent accuracy, but hybrid models achieved improved accuracy over fully connected. Performance analysis suggests that class imbalance was an underlying advantage of the hybrid classification, particularly in the mild DR classes. It also suggests the advantages of transfer learning and RF-based classification in large-scale and imbalanced datasets.

Conclusions: In this study, it was shown that combining DL feature extractors with classical ML classifiers (Decision tree classifiers) could enhance the accuracy of DR detection. The best combination of the Xception-RF model was able to classify DR on a large, imbalanced dataset with advanced metrics. Using transfer learning with only a few epochs was sufficient and cost-effective. It is evident that adopting hybrid DL-ML pipelines can be highly effective for DR screening, especially in low-resourced healthcare systems, which allows for scalable practices. All in all, the next procedural study should make use of balanced datasets with wider population demographics and enhance further classification performance through different DL architectures.

Keywords: Diabetic Retinopathy, Deep Learning, Ablation Study, Convolutional Neural Networks, Transfer Learning, Random Forest

*Corresponding author address: Ulster University, Northland Road, BT48 7JL, Londonderry, Northern Ireland, United Kingdom. Email: Rahman-m11@ulster.ac.uk
 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

1 Introduction

Diabetes is a chronic illness that can initiate several complications, including damage to the nervous system, kidneys, eyes, feet, skin, and ears, as well as impairing memory and reasoning [1]. Complications associated with diabetes can be classified as acute and chronic complications. Recurring complications comprise microvascular disease and macrovascular disease. In conclusion, diabetes is a leading cause of sickness and death worldwide [2]. Recognising diabetes at an early stage and optimally managing diabetes allows the diabetic person to live longer and have better health [3]. It is estimated that one out of ten adults worldwide have diabetes (10.5%), with good expectation of this increasing to 12.2% by 2045 [4]. More than 54% of adult diabetics globally live in the continents Asia and Africa (21% and 33% respectively) [5]. This is a problem given the limitations in the diagnosis and management. It has been reported that nearly half (44.7%) of adult diabetics are undiagnosed and therefore, unaware of their affliction, and the risk of other problems associated [4]. There is evidence that the condition known as diabetic retinopathy may cause damage to the blood vessels in the retina [6]. Early detection and treatment of this retinopathy can prevent blindness and vision impairment [7]. The literature contains a number of techniques for analysing retinal pictures, such as DL and ML techniques, which aid in the diagnosis and treatment of diabetic retinopathy [3].[4]. Two subsets of artificial intelligence methods, machine learning (ML) and deep learning (DL), learn from data and have a variety of applications, such as object classification, detection, and segmentation. [10]. To categorise the degree of diabetic retinopathy, for instance, machine learning may utilise optical coherence tomography angiography (OCTA) pictures to extract features [6]. In another example, deep learning may use fundus images, in order to identify microaneurysms, haemorrhages and exudates and to measure the level of diabetic retinopathy as a continuum [7]. This means that these techniques potentially increase the accuracy, overall efficiency and reliability of the screening for diabetic retinopathy, whilst lessening the human effort and observation involved [8][9].

Initially, the diagnosis of diabetic retinopathy relied entirely on trained clinicians manually analysing retinal images, which was labour-intensive, time-consuming, and prone to subjective variability. Advances in digital imaging enabled automated analysis of fundus images, reducing clinical workload and improving consistency. However, early machine learning methods still required manual feature engineering. The emergence of deep learning, particularly convolutional neural networks, eliminated the need for handcrafted features and significantly improved diagnostic accuracy. As a result, hybrid approaches combining deep feature extraction with classical classifiers have become a promising direction for automated medical image analysis.

Comparing ML and DL for cervical cancer classification [24]: This study compared the ResNet-50 model with XGB, Support Vector Machine (SVM), and Random Forest (RF) models in order to identify four common disorders using cervicography images. According to this study, ResNet-50 outperforms classical classifiers in terms of accuracy, sensitivity, specificity, and F1 score. According to the research, in order to improve the model's interpretability, robustness, and reliability, it needs be verified on bigger, more varied datasets, such as clinical data and CT scans.

DL image classification data augmentation [25]: This paper presents data augmentation-based image style transfer for deep learning models in image classification. Using a dataset of ten animal categories, style transfer produced new images. The authors claim that the data augmentation produced greater accuracy for all four deep learning models (InceptionV3, ResNet50, VGG16, AlexNet) when analysed on the enlarged dataset. They advocated for expanding the use of data augmentation to other datasets and domains, considering further testing of style transfer methods, and examining how data augmentations influence different deep learning architectures. A review of DL and ML dermatoglyphics fingerprint pattern classification literature [26]: This review paper covers 50 DL/ML fingerprint pattern classification papers. The review paper covers skills based on DL using CNN, ML based on SVM, and optimisation methods. The review paper addressed research problems in fingerprint pattern classification, as well as potential future research directions, including handling noisy, unbalanced, and low-resolution patterns, as well as complex patterns.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

Machine Learning Fashion Image Classification Review [27]: To classify 60,000 Zalando Research images of various fashion categories into 10 different image classifications, this study examines four different deep learning methods (AlexNet, GoogleNet, VGG16, and ResNet50) and four different machine learning models (ANN, KNN, SVM, and RF). The study shows that the GoogleNet classifier has the most accuracy at 93.75% and ResNet50 is second at 92.5% and VGG16 at 91.25% and then AlexNet 90% accuracy. With an average accuracy CS (goal) of 88.71%, the study also demonstrates that Artificial Neural Networks (ANN) outperform machine learning (ML) continuous accuracies in terms of average accuracy. The leaning made some changes also to consider like the DenseNet and the InceptionV3 DL models, some changes to data rotating and flipping and also consider SIFT and SURF for the feature extraction models.

CNN/ML Similarity Image Classification [28]: This study compared the features and image classification methods in ML and CNN. The paper used 1000 Caltech101 images. CNNs had better accuracy and time performance than typical ML algorithms; the paper indicates that SIFT features with SVM scored 83%. The paper stated that it would use the CIFAR-10 and ImageNet datasets, LBP and HOG features, and tune the parameters of the ML and CNN algorithms.

Embedded Computer Vision [31]: Vora and Shrestha's research paper targeted diabetic retinopathy detection using embedded computer vision. The potential accuracy of the trained model in identifying diabetic retinopathy was close to 76%.

There were no other studies doing ablation model studies to determine diabetic retinopathy from retinal fundus images. There was no research on that topic, and with the estimated accuracy of between 60-80%, that is extremely likely to help an ophthalmologist with AI research. The following were the aims:

- To carry out ablation experiments with a series of DL models and finally a series of classifiers.
- To find the most accurate diabetic retinopathy detection using the combination and cooperation of the DL model with a classifier.

The paper is broken down into five sections. Section II provides a description of the methodology. Section III contains the experimental outcomes. Section IV presents a review of the study's findings and potential directions for future research. Section V contains a closing statement.

2 Methodology

2.1 Machine learning concepts

Particularly when utilising convolutional neural networks (CNNs), deep learning (DL) approaches outperform traditional machine learning techniques for feature extraction and picture classification.There are two primary parts to a CNN: a feature extractor, made up of convolutional and pooling layers, and a classifier made up of any machine learning (ML) method; for instance, in the process of developing a CNN for image classification, in order to calculate the actual classification, the CNN pulls "features" from each input image, such as edges and shapes. When elaborating on the classifier, the fully connected layer can be replaced with other classifiers, like SVM, RF, or DT, to improve overall classification performance.[11][17]

For instance, when constructing a CNN for image classification, the CNN's use of "features" such as edges and forms determines the actual categorisation for each input image. For example, while VGG16 contains only 16 layers for image classification, ResNet152V2 utilises identity mappings to facilitate learning. In contrast, Xception relies on depth-wise separable convolutions rather than regular convolutions for improved interpretation of the input image.[12][13][14][15][16].

Combining DL approaches with ANN, RF, DT, and k-Nearest Neighbours (k-NN) classifiers can further increase classification accuracy. ANN is very flexible, but can be computationally expensive. RF creates a robust ensemble of decision trees, DTs are simple and intuitive, and k-NN is an easy and powerful technique for classification applications.[18][19][20][21].

2.2 Deep learning's application to retinopathy

A growing number of ophthalmologists are using deep learning to detect and assess diabetic retinopathy (DR), a condition associated with diabetes that can result in vision loss. Lesions such as microaneurysms

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

and bleeding damage the retina's blood vessels in DR. The detection procedure makes use of retinal imaging, such as optical coherence tomography (OCT) and fundus images, and optical coherence tomography angiography (OCTA). Deep learning algorithms can automatically assess image quality, detect lesions, and classify DR using established grading criteria [7]. Figure 1 illustrates the various stages of retinopathy.

Ophthalmologists use retinal images from cameras (fundus camera) to determine the disease pathology. The four phases of proliferative diabetic retinopathy (PDR) and non-proliferative retinopathy are mild, moderate, and severe. Blindness or sight loss could result from blood vessel damage as the illness develops. [23].

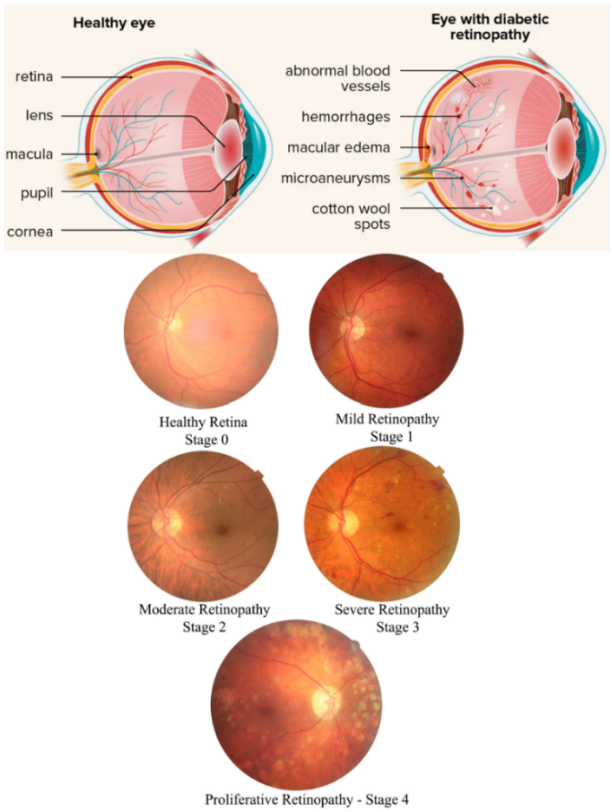

Figure 1. Diabetic Retinopathy Stages

2.3 Experimentation

This proposed ablation experiment will investigate image processing, feature extraction, classification, and performance measurement using four classifiers (ANN, RF, DT, and k-NN) and three deep learning models (VGG16, ResNet152V2, and Xception). Deep learning is ideal for extracting such complex features from large image datasets based on its ability to recognise patterns, scalability, and limited preprocessing requirements. Data from images was collected from EyePACS and stored on Kaggle. Processing was then conducted on Google Drive and Google Colab. Python was selected as the programming language due to its ease of use and extensive library.

Image classification is a supervised learning problem that assigns class labels to images into known categories. VGG16, ResNetV2, and Xception were applied as feature extractors, employing transfer learning to improve efficiency and machine learning classifiers ensure accurate categorisation.

The experiment produced a total of twelve model scenarios by combining features from a DL model with traditional classifiers and was applied to training, validation, and testing on new images. Performance of the models was measured using a confusion matrix.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

Google Colab is the cloud-based Jupyter Notebook environment used in the experiment. Processing was performed in the cloud using Google Cloud's large-scale image categorisation with its own GPUs (Tensor Processing Units). According to estimates, it contained 40 GB of GPU memory, 83 GB of RAM, and a premium GPU subscription to allow for plenty of capability.

2.4 Datasets and model evaluation metrics

EyePACs allowed us to utilise their Kaggle public image dataset, which comprised 88.29 GB of data. This dataset consisted of 15 zip files with 53,576 unlabelled images for testing and 35,126 labelled images for training. We divided the 35,126 photos into two training and validation folders, using 10,500 for validation and 24,626 for training. The image files were organised in order of retinopathy severity, where 0 was no retinopathy and 4 was the most severe retinopathy.

The study used a 3-step process. The first step was to convert all photographs to greyscale and resize them. The Xception model was used with a 290 x 290-pixel resolution, while VGG16 and ResNetV2 each used a 224 x 224-pixel resolution. Each model was uploaded to two separate folders. The second step involved training the models on the processed images, and the third step involved validating the remaining processed images.

The fact that both files included five subfolders classified by retinopathy indicates that the data was not fully balanced. However, this is the reality of the real world, where disease databases are more representative, as they generally have more samples that are considered disease-free.

We implemented a confusion matrix to show model performance, which compared actual class labels against predicted class labels (TABLE I). While simultaneously showcasing our model's accuracy, precision, recall, and other metrics, the confusion matrix enabled us to pinpoint the model's benefits and drawbacks.

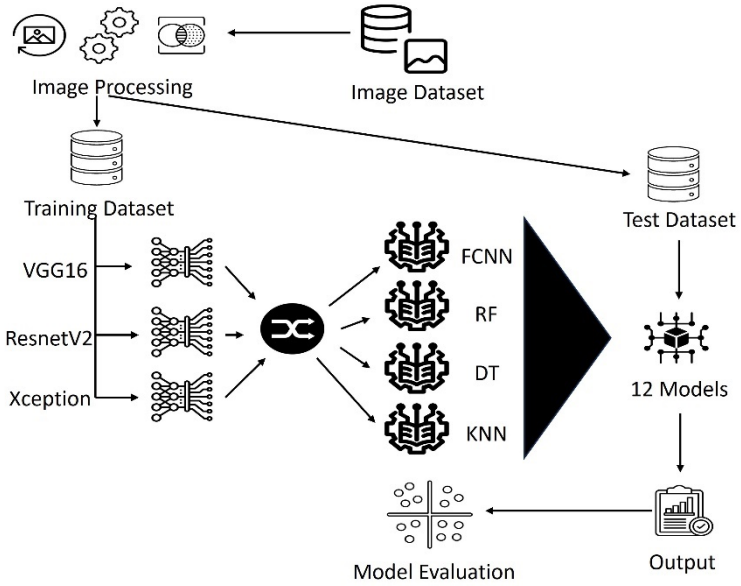

Figure 2. Methodology of the Ablation Experiment

Data augmentation and class balancing:

No explicit synthetic data augmentation (such as rotation, flipping, or colour jittering) was applied in this study. The models were trained on the original dataset distribution to preserve the real-world imbalance in diabetic retinopathy prevalence. Similarly, no oversampling or undersampling strategies were used for class balancing. Instead, robustness to class imbalance was implicitly handled by the Random Forest ensemble classifier, which demonstrated superior performance on minority classes compared to fully connected CNN models.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

Table 1. Confusion matrix metrics.

Metric	Formula	Significance
Accuracy	(TP + TN)/ (TP + TN + FP + FN)	The frequency with which a model or entity makes accurate predictions is a straightforward indicator of accuracy; however, if either the data points are imbalanced or there are more than two classes, the accuracy metric can be misleading.
Precision	TP/(TP+FP)	Precision is a statistic that shows how accurate a model or entity is; it determines the percentage of true positives among all anticipated positives. For spam or fraud detection, for example, precision is a helpful measure to have when you want to limit false positives at all costs.
Recall	TP/(TP+FN)	Recall is a metric that indicates how accurately an entity or model predicts; that is, it determines the ratio of actual positives to all anticipated positives. Recall is useful when you do not want to produce false negatives (e.g., medical diagnoses or loyal customers).
F1	2* (Precision * Recall) / (Precision + Recall)	Since the F1 score is the harmonic mean of precision and recall, it assigns equal weights to false positives and false negatives. The F1 score is a number between 0 and 1 that represents the model's matching prediction performance, where 0 indicates a complete failure and 1 indicates a perfect match. The F1 score benefits from unbalanced data and is highly exposed at the accuracy and coverage points.

2.5 Image Processing Challenges and Solutions

During this study, JPEG images of varying sizes were scaled down and converted to greyscale (pixel range: 0-255). A standard CPU took 25 hours to process 35,126 images. The images were saved as NumPy arrays on a cloud drive, resulting in four arrays for the 224x224 and 290x290 pixel train and test datasets.

The project's next stage was to create three deep learning models to extract information from the images. Although normalisation is a normal practice, it was not applied during this study due to hardware constraints in Google Colab, which has 83 GB of RAM and 40 GB of GPU RAM. TensorFlow was set to utilise 70% of the GPU RAM, but at times it exceeded this limit, causing issues with the system.

The final stage of this analysis involved incorporating image features into machine learning classifiers, including a Random Forest classifier optimised with a grid search. The Random Forest classifier was optimised using grid search over key hyperparameters, including number of estimators (100 to 1000), maximum tree depth (None, 10, 20, 50), minimum samples per split (2, 5, 10), and maximum feature selection strategy (sqrt, log2). The final configuration selected the parameter set yielding the highest validation accuracy.

2.6 Mathematical representation of the experiment

Problem Setup and Approach

Let X be the set of images of the human eye retina, and let

$$Y = \{0,1,2,3,4\}$$

be the set of labels corresponding to the five categories of diabetic retinopathy. Our objective is to discover a function.

$$f: X \to Y$$

that can accurately classify any image $x \in X$ into its correct label $y \in Y$.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

Approach

Various deep-learning methods and machine-learning classifiers as building blocks were employed for the function f.
Each deep learning method D is a function that holds an image $x \in X$ and yields a feature vector:

$$D(x) \in \mathbb{R}^n$$

where n is the feature space's dimension.
Each machine learning classifier C is a function that holds a feature vector $v \in \mathbb{R}^n$ and outputs a label:

$$C(v) \in Y$$

Function Definition

The function is defined for any combination of a machine learning classifier C and a deep learning technique D:

$$f_{D,C}: X \to Y$$

by

$$f_{D,C}(x) = C(D(x))$$

Training and Evaluation

We trained each function's logs and tested them on a database of images and corresponding labels, measuring their performance using a variety of metrics, including Accuracy, precision, recall, and F1-score.

Some hyperparameters were also changed (Table II) and tested with varying numbers of epochs to derive the optimal condition for performance for each function $f_{D,C}$. Hyperparameters that influence the behaviour of learning algorithms include the number of hidden layers, batch size, and learning rate. Conversely, epochs specify the number of times the full training dataset is used to train the learning algorithm.

Table 2. Hyperparameters were used in the experiment.

Algorithm	Hyperparameter	Value used
VGG16	i. input_shape	(224, 224, 3)
	ii. include_top	FALSE
	iii. weights	imagenet
	iv. pooling	None
	v. classes	1000
Resnet152V2	i. input_shape	(224, 224, 3)
	ii. include_top	FALSE
	iii. weights	imagenet
	iv. pooling	None
	v. classes	1000
Xception	i. input_shape	(290, 290, 3)
	ii. include_top	FALSE
	iii. weights	imagenet
	iv. pooling	None
	v. classes	1000
Random Forest	i. n_estimators	1000
	ii. max_depth	None
	iii. criterion	Gini
	iv. min_samples_leaf	1
	v. min_samples_split	2
	vi. bootstrap	TRUE

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

	vii.	max_features	Sqrt
	viii.	n_jobs	-1
	ix.	N_splits	10
Decision Tree	i.	Criterion	Gini
	ii.	splitter	best
	iii.	max_depth	None
	iv.	min_samples_split	2
	v.	min_samples_leaf	1
	vi.	max_features	None
KNN	i.	n_neighbours	5
	ii.	weights	uniform
	iii.	algorithm	auto
	iv.	leaf_size	30
	v.	p	2
	vi.	metric	minkowski

The mathematical representation of the experiment can be explained below.
Given:

$$X, Y, \{D_1, D_2, D_3, D_4, D_5\}, \{C_1, C_2, C_3, C_4, C_5\}$$

For each:

$$D_i \in \{D_1, D_2, D_3, D_4, D_5\},\ C_j \in \{C_1, C_2, C_3, C_4, C_5\}$$

Define:

$$f_{D_i,C_j}(x) = C_j(D_i(x))$$

Train and Test:
Train and test f_{D_i,C_j} on a dataset of $(x, y) \in X \times Y$.

Measure:
Evaluate the performance of f_{D_i,C_j} using some metric M.

Optimize:
Tune hyperparameters and epochs of f_{D_i,C_j} to maximize M.

Report:
Identify and report the best function f_{D_i,C_j} and its corresponding performance.

To represent the whole experiment in one formula, Let ε represent the overall experiment, $\mathcal{M}$ represent the set of DL models, $\mathcal{C}$ represent the set of ML classifiers, $\mathcal{H}$ represent the set of hyperparameters, $\phi_{\mathcal{M}}$ represent the feature extraction function for model $\mathcal{M}, \gamma_{\mathcal{C}}$ represent the classification function for the classifier $\mathcal{C}$, Δ_{train} and Δ_{test} represent the training and testing datasets, $\boldsymbol{\eta}$ represents the hyperparameter vector, and Π represents the performance metric.

$$\varepsilon = \sum_{\mathcal{M}\in\mathcal{M}} \sum_{\mathcal{C}\in\mathcal{C}} \sum_{\boldsymbol{\eta}\in\mathcal{H}} \Pi\left(\gamma_{\mathcal{C}}\left(\sum_{x_i\in\Delta_{\text{test}}} \phi_{\mathcal{M}}(x_i, \boldsymbol{\eta}) + \lambda \cdot \nabla_{\boldsymbol{\eta}}\mathcal{L}_{\mathcal{M}}(\boldsymbol{\eta})\right), Y_{\text{test}}\right)$$

In this representation:
λ represents a regularisation parameter.
$\nabla_{\eta}\mathcal{L}_{\mathcal{M}}(\boldsymbol{\eta})$ is a representation of the gradient of a loss function $\mathcal{L}_{\mathcal{M}}$ in relation to the hyperparameters $\boldsymbol{\eta}$.

3 Results

There are twelve combinations based on three DL models and four classifiers. The sets of hyperparameter settings were held to optimal settings, and identical sets of parameters for model algorithms were used across all varieties. At the end of each epoch, a computation is performed to yield loss and accuracy values. It calculates the accuracy and loss to find out how well the model is working. The loss number shows how well the model predicts the correct outcome given the information at hand. Our goal is to minimise this

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

number as much as possible. The accuracy is a measure of how well the model can categorise the photos within the pre and post-test training sets. There is a connection between the ability to increase training accuracy, decrease loss, and increase testing accuracy. As testing and training accuracy increase, loss reduces. If the model is trained for too long, it will begin to overfit to the training data, meaning it will perform at the upper end of model accuracy on the data it was trained on, but will perform worse with new data. To avoid overfitting and underfitting during the training phase, it is crucial to choose the right number of epochs. This experiment compared the model's accuracy throughout training and testing across a maximum of 30 epochs.

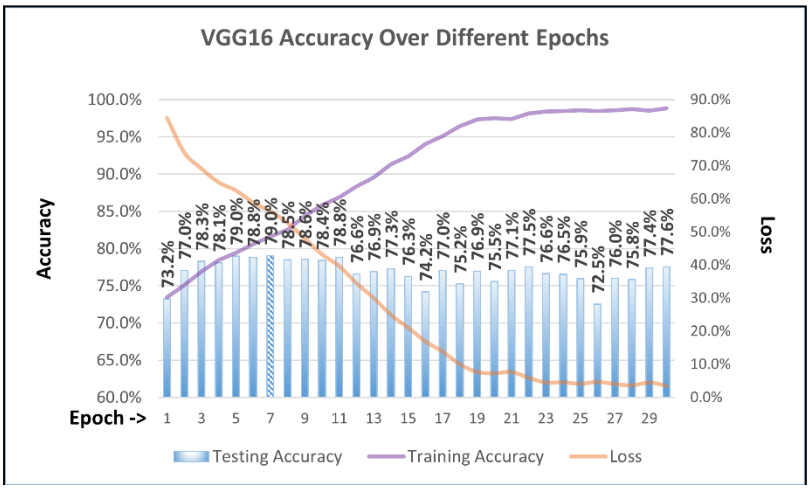

Figure 3. The fully connected VGG16 model's accuracy and loss during training and testing

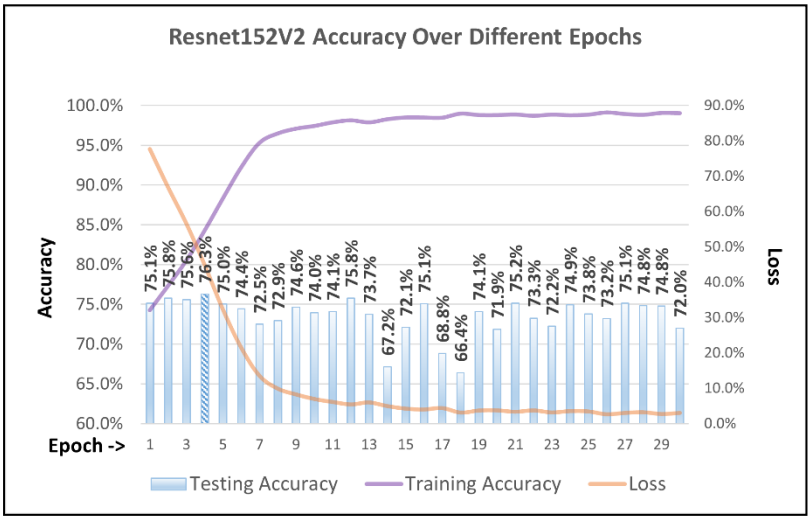

Figure 4. Accuracy and loss of a fully linked ResNet V2 model during training and testing

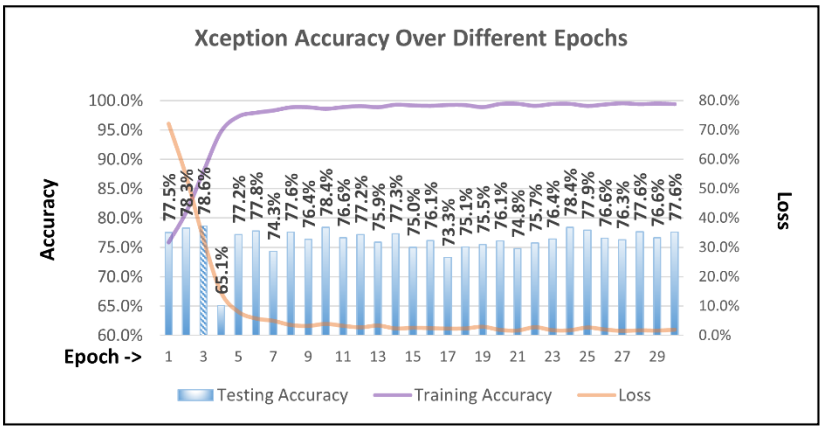

Figure 5. Xception model accuracy and degradation over various epochs

The performance of the fully connected VGG16 model is detailed in Figure 3. Training and testing adjustments were investigated over 30 epochs. The training accuracy peaked at 98.9% with a loss of 3.5% at epoch 30, indicating good convergence. Testing accuracy peaked earlier at approximately epochs 5 and 7 at about 79.0%, but gradually deteriorated thereafter. This indicates that testing accuracy was subject to overfitting effects at this point. The training accuracy at epoch 7 was still moderate at 81.6%, which suggests that some balance existed at this stage between the generalisation of the model and learning features with divergence occurring thereafter.

The ResNet152V2 model is demonstrated in Figure 4. The training accuracy peaked at a maximum of 99.1% with a minimum loss of 2.6% being achieved at epoch 26. The peak value for testing accuracy occurred much earlier, however, usually around epochs 4 with a reading of 76.3% with a concomitant training accuracy reading of 84.3%. This again indicates that while training was being improved upon the capability for generalisation of the model was being lost after early epochs of fitting, which is a classical symptom of overfitting. The Xception model also gave rise to interesting results shown in Figure 5. The peak training accuracy achieved was 99.5% with the lowest loss scored obtaining a figure of 1.5% at epoch 27. The peak value for testing accuracy on this occasion was recorded as 78.6% at epoch 3 followed by a plateau followed by some marginal deterioration thereafter. Again the striking difference in testing accuracy and the high training accuracy suggested that was an indication of the presence of overfitting when epochs increase beyond those at the earlier stages of training.

It is clear from examining all three CNN architectures detailed that the peak testing accuracies occurred within the first 5–7 epochs of training, followed by a drop off performance after further training epochs taken. This is indicative that for large data sets such as diabetic retinopathy classification, early stopping in this case after a few epochs is optimal in retaining values for generalisations and avoiding the appearance of overfitting. VGG16, ResNet152V2 and Xception architectures were applied in this study as feature extractors for downstream fully connected CNNs interfaced with Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbour (KNN) machine learning algorithms. The combination yielded twelve hybrid combinations, which were compared by means of 10-fold cross-validation studies, which showed considerable variations in accuracy results. The comparative charts suggest that the testing accuracies on these series all follow earlier epoch appearance styles in accuracy climb, indicating that CNN feature extraction with classical machine learning algorithms could lend itself to attaining appreciable amounts improvement values in visual patterns factoring as few training iterations as possible.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

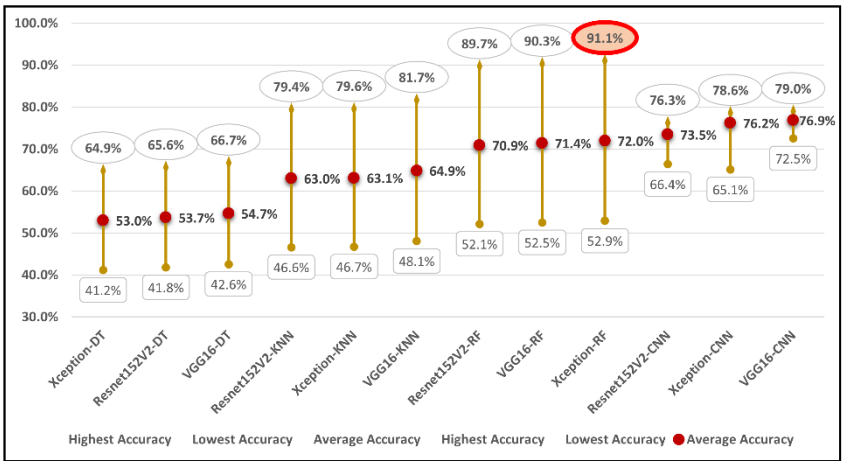

Figure 6. Testing the accuracy of twelve models

The combination of deep learning based feature extractor routines with traditional machine learning classifiers as illustrated in Figure 6 provided substantial increases in testing performance. The CNN architectures yielded testing performance in the region of 76.3-79.0% on outcomes. The hybrid models provide significant improvements with some yielding performance of over 90% accuracy. Of the twelve models employed, the best performing testing accuracy was exhibited by the Xception-Random Forest (Xception-RF) combo achieving performance of 91.1%, which was considerably better than all other combinations. The success of this model is due to the combined capability of representation of depth of the Xception model inused with the stability of the ensemble learning Random Forest classifier, producing effective generalisation performance. This particular model gave an average accuracy of about 72% over the epochs employed, which illustrates the consistent performance over the totality of the training process. The VGG16-Random Forest (VGG16-RF) achieved performance accuracy in the second best testing performance category producing 90.3% with an average accuracy of 71.4%, which was somewhat similar to that of the Xception-RF configuration. The further affirmation of strength in performance of Random Forest is therefore foreshadowed by these results the implemented hybrid structures giving stability in performance being somewhat modified, but demonstrated poorer performance where testing accuracy did plateau in the region of 76-79% during the testing session. This would no doubt indicate an ability of the structures of CNN architectures to give decent capabilities of learning features, but rather of the improved existence of hybrid learning structures (in other words, a coupling of the deep features with classifier of the decision type, rather indicating), of facilitated learning structure capability with discrimination power increasing and generality for robustness, of the models in employed structure. It may therefore ultimately conclude that classes of hybrid deep structures using CNN based routines of feature extraction in used with either heavier classifiers of adaptation as RF varieties or distance based properties either such as KNN classifiers shall give better returns in matters of testing accuracy, than the pure CNN's in used. The overall observations of the stability of average accuracy of epochs in the relevant studies gives a tendency for better convergence and correctabilities of overfitting phenomenon strongly, especially in those models involving RF as the superior classifier in used technology.

3.1 Computational efficiency:

From an empirical standpoint, hybrid models required significantly less training time than fully connected CNNs, as only feature extraction was performed by deep networks, while classification was performed using lightweight machine learning classifiers. Fully connected CNNs required end-to-end backpropagation across all layers for each epoch, whereas hybrid models converged faster due to reduced parameter optimisation. This indicates that hybrid DL–ML pipelines are computationally more efficient and scalable for large datasets.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

3.2 Sensitivity analysis:

Qualitative analysis of prediction behaviour indicates that hybrid models, particularly Xception-RF, exhibit higher sensitivity to mild and moderate DR classes than fully connected CNNs. The hybrid classifiers exhibited improved discrimination between adjacent severity levels (e.g., mild vs. moderate), which are typically more difficult to separate. This suggests that Random Forest is more robust to class overlap and noisy features, thereby improving sensitivity in clinically critical borderline cases.

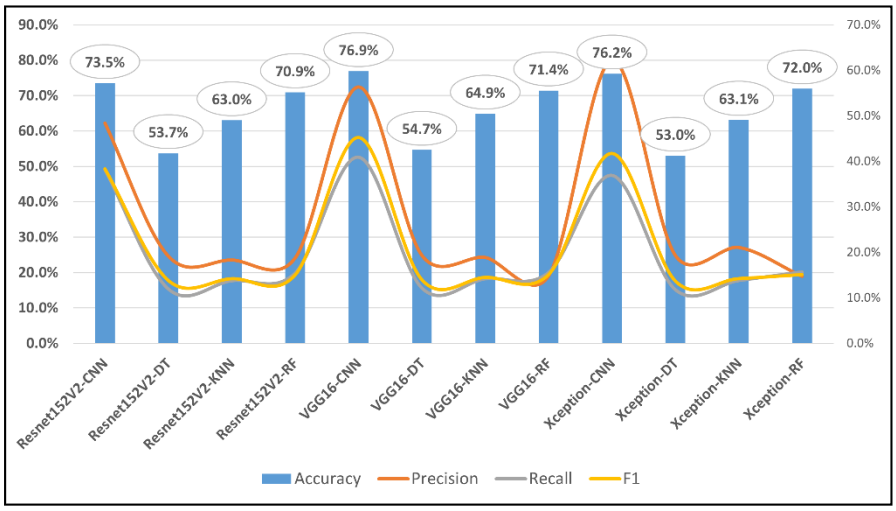

Figure 7. Twelve models' performance graph using confusion matrix measurements

Figure 7 displays the evaluation's findings in terms of metrics from the confusion matrix. The four models with the highest average accuracies were the fully connected Resnet152V2, VGG16, Xception, and hybrid model Xception-RF. Only fully linked models, however, outperformed the other three metrics—F1, Precision, and Recall—in a comparable manner. Although none of the models attained the highest testing accuracy, the fully connected Xception and VGG16 models outperformed the others. The model's performance suffered due to the discrepancy in class counts in categorical picture data.

3.3 Ablation summary of model combinations:

Across the twelve evaluated combinations, Random Forest consistently outperformed ANN, DT, and KNN when paired with deep feature extractors. The ranking of classifiers by overall performance was RF > ANN > KNN > DT. Among feature extractors, Xception produced the most discriminative representations, followed by VGG16 and ResNet152V2. The best overall configuration was Xception-RF (91.1%), while the weakest was Xception-DT (41.2%), demonstrating the dominant role of classifier choice in hybrid pipelines.

4 Discussion

There were only twelve combinations of classifiers (ANN, RF, DT, KNN) and deep learning algorithms (VGG16, Resnet152V2, and Xception) due to time and resource constraints. Experiments with numerous other completely connected CNNs were not possible for this study. This could be a possibility for future research. A fully connected DL model, VGG16, had the best prediction accuracy at the same time, at 79%; but during the first epoch of the experiment, Xception had the worst prediction accuracy at 65.1%. The total measurement range of variation for this type was 13.9%, so for Prediction (Ev) training and testing it can be considered reasonable if say, one model has 85% at Epoch 10 and 80% at Epoch 15, and other model is +/- 10% with the other model being +/- 5% on prediction testing. In terms of training and testing models for comparison of accuracy in classification, the Xception-Rf model (Xception-Based Deep Learning) produced an overall predicted accuracy of 91.1%. Xception-DT is the model in this group that is worse at

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

classification compared to other classification models. The predictions with Xception-DT for the correct classification category were 41.2%, which varied more than those of other models, and this difference increased by 49.9% when most models predicted annual change measurements.

In the training of fully connected models, as indicated, the number of epochs is usually 30. However, the performance of a modified classification model using feature training stopped improving after the 7th epoch. The performance levels for testing anonymous data began to decline, as observed in all cases, and performance dropped over time. In this work [32], which uses deep learning to diagnose diabetic retinopathy, I mention that the worst performance for a model, as indicated by the prediction change (Ev) values, was around the 15th training epoch for all models. For the performance test of a model, it is possible to have different hyperparameters set to different values for features and classifiers. This represents a significant area for future research. The poor performance results regarding the classification of this type of data are also due to the class imbalance of the images reported within the dataset itself. In the future, tests like this may be conducted using more representative data. Other important information was also missing from the images (e.g., patient demographics). The impacts of age and other eye conditions, for instance, can alter the retina and are not considered here, but they can be considered significant considerations in the future.

Random Forest outperforms other classifiers primarily due to its ensemble nature, which reduces variance and improves robustness to noisy and high-dimensional feature spaces. Deep learning feature extractors generate complex and redundant representations, which are effectively handled by Random Forest through random subspace sampling and majority voting. In contrast, ANN and KNN are more sensitive to noise and class imbalance, while Decision Trees tend to overfit high-dimensional data.

4.1 Statistical interpretation:

Although formal statistical significance testing (e.g., paired t-tests) was not performed due to computational constraints and the absence of multiple independent training runs, the observed performance gaps between top-performing models (e.g., Xception-RF vs. fully connected CNNs) were sufficiently large (>10%) to suggest practical and meaningful performance differences.

4.2 Generalizability considerations:

While the EyePACS dataset provides a large and diverse sample, the generalisability of the proposed models to other clinical datasets (e.g., Messidor, APTOS, DDR) remains an important consideration. Differences in camera type, illumination, patient demographics, and grading standards may affect model performance. However, since the proposed framework relies on transfer learning and generic feature representations, the hybrid DL–ML approach is expected to generalise reasonably well with minimal fine-tuning.

Recent state-of-the-art DR detection systems using end-to-end deep learning architectures report accuracies in the range of 80–88% on public datasets such as EyePACS and Messidor. The proposed Xception-RF hybrid model achieves a competitive accuracy of 91.1% while requiring fewer training epochs and reduced computational cost. This highlights the effectiveness of hybrid architectures as a lightweight alternative to fully end-to-end deep learning systems.

5 Conclusion

This work greatly benefited from the use of deep learning for feature extraction and conventional classifier techniques for picture classification; the combination of Random Forest and Xception was able to determine the stage of retinopathy with a success rate of more than 91%. All performance-measuring criteria were impacted by the small number of cases in the moderate retinopathy group, which is frequently ignored as having no retinopathy. The best classifier was Random Forest, however optimised RF might produce even better outcomes. When it came to detecting diabetic retinopathy in retina fundus images, using a fully connected CNN model with a large number of training epochs did not significantly improve the results.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

Seven epochs were enough to train the fully connected models on this kind of image when employing the transfer learning technique. RF was the top classifier, and VGG16 was the best feature extractor. Xception-RF was the best model we could create.

References

[1] K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, and D. Papazoglou, "Complications of diabetes," J. Diabetes Res., vol. 2015, p. 189525, 2015.

[2] K. Kantawong, S. Tongphet, P. Bhrommalee, N. Rachata, and S. Pravesjit, "The methodology for diabetes complications prediction model," in 2020 Joint International Conference on Digital Arts, Media, and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), 2020, pp. 110–113.

[3] S. Colagiuri and D. Davies, "The value of early detection of type 2 diabetes," Curr. Opin. Endocrinol. Diabetes Obes., vol. 16, no. 2, pp. 95–99, 2009.

[4] "IDF diabetes atlas 2021," Diabetesatlas.org. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/. [Accessed: 29-Apr-2023].

[5] "Diabetes," Who. int. [Online]. Available: https://www.who.int/news-room/facts-in-images/detail/diabetes. [Accessed: 29-Apr-2023].

[6] N. Gundluru et al., "Enhancement of detection of diabetic retinopathy using Harris hawk's optimisation with a deep learning model," Comput. Intell. Neurosci., vol. 2022, p. 8512469, 2022.

[7] L. Dai et al., "A deep learning system for detecting diabetic retinopathy across the disease spectrum," Nat. Commun., vol. 12, no. 1, p. 3242, 2021.

[8] G. Ryu, K. Lee, D. Park, S. H. Park, and M. Sagong, "A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography," Sci. Rep., vol. 11, no. 1, p. 23024, 2021.

[9] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey," IEEE Access, vol. 10, pp. 28642–28655, 2022.

[10] A. K. Gangwar and V. Ravi, "Diabetic retinopathy detection using transfer learning and deep learning," in Evolution in Computational Intelligence, Singapore: Springer Singapore, 2021, pp. 679–689.

[11] Q. Lv, S. Zhang, and Y. Wang, "Deep learning model of image classification using machine learning," Adv. Multimed., vol. 2022, pp. 1–12, 2022.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv [cs.CV], 2014.

[13] H. H. N. Alrashedy, A. F. Almansour, D. M. Ibrahim, and M. A. A. Hammoudeh, "BrainGAN: Brain MRI image generation and classification framework using GAN architectures and CNN models," Sensors (Basel), vol. 22, no. 11, p. 4297, 2022.

[14] P. Dou, H. Shen, Z. Li, X. Guan, and W. Huang, "Remote sensing image classification using deep–shallow learning," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 3070–3083, 2021.

[15] X. Wu, R. Liu, H. Yang, and Z. Chen, "An xception-based convolutional neural network for scene image classification with transfer learning," in 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 2020, pp. 262–267.

[16] H. Benbrahim and A. Behloul, "Fine-tuned Xception for Image Classification on Tiny ImageNet," in 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), 2021, pp. 1–4.

[17] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1–6.

[18] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 72–78.

[19] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," IEEE Trans. Image Process., vol. 27, no. 8, pp. 4012–4024, 2018.

[20] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Trans. Syst. Man Cybern., vol. 21, no. 3, pp. 660–674, 1991.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615

[21] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of the nearest neighbour algorithm for learning and classification,” in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255–1260.
[22] R. P, Poonkodi, F. D. Shadrach, Anitha, E. Mary, and R. Nirmalan, “Diabetic retinopathy detection using retinal fundus image and image enhancement using fuzzy clustering,” in 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022, pp. 1–7.
[23] C. Wint and C. Guthrie, “Diabetic retinopathy,” Healthline, 11-Feb-2022. [Online]. Available: https://www.healthline.com/health/type-2-diabetes/retinopathy. [Accessed: 29-Apr-2023].
[24] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, “Comparison of the machine and deep learning for the classification of cervical cancer based on cervicography images,” Sci. Rep., vol. 11, no. 1, p. 16143, 2021.
[25] A. Mikolajczyk and M. Grochowski, “Data augmentation for improving deep learning in an image classification problem,” in 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, pp. 117–122.
[26] M. A. Bhimrao and B. Gupta, “An empirical study of dermatoglyphics fingerprint pattern classification for human behaviour analysis,” Soc. Netw. Anal. Min., vol. 13, no. 1, 2023.
[27] B. Samia, Z. Soraya, and M. Malika, “Fashion image classification using machine learning, deep learning, and transfer learning models,” in 2022 7th International Conference on Image and Signal Processing and their Applications (ISPA), 2022, pp. 1–5.
[28] G. K. Devipriya, E. Chandana, B. Prathyusha, and T. S. Chakravarthy, “Image Classification using CNN and Machine Learning,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 575–580, 2019.
[29] V. Nithyashree, “Image Classification using Machine Learning,” Analytics Vidhya, 20-Jan-2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/01/image-classification-using-machine-learning/. [Accessed: 30-Apr-2023].
[30] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, “Transfer learning for image classification using VGG19: Caltech-101 image data set,” J. Ambient Intell. Humaniz. Comput., vol. 14, no. 4, pp. 3609–3620, 2023.
[31] P. Vora and S. Shrestha, “Detecting diabetic retinopathy using embedded computer vision,” Appl. Sci. (Basel), vol. 10, no. 20, p. 7274, 2020.
[32] M. S. Rahman, "Diabetic Retinopathy Detection Using Deep Learning and Ablation Experiment," unpublished manuscript, School of Computing, Engineering, and Intelligent Systems, Ulster University, Northern Ireland, UK, 2023.

 J Health Inform Afr. 2026;13(1):17-31. DOI: 10.12856/JHIA-2026-v13-i1-615