Towards Interpretable Diabetic Retinopathy Detection: Combining Multi-CNN Models with Grad-CAM

Zakaria Said[1]*, Fatima-Ezzahraa Ben-Bouazza[2], Mounir Mekkour[3]
Mathematical Analysis and Applications Laboratory, University Sidi Mohamed Ben Abdellah, Fes, Morocco[1]
Faculty of Science and Technology, Hassan I University, Settat, Morocco[2]
Mohammed VI University of Health Sciences, Casablanca, Morocco[2]
LaMSN, La Maison Des Sciences Numériques, France[2]
Mathematical Analysis and Applications Laboratory, University Sidi Mohamed Ben Abdellah, Fes, Morocco[3]

Abstract—**Diabetic retinopathy (DR) is a leading cause of vision impairment and blindness, necessitating accurate and early detection to prevent severe outcomes. This paper discusses the utility of ensemble learning methodologies in enhancing the prediction accuracy of Diabetic Retinopathy detection from retinal images and the prospective utilization of Gradient-weighted Class Activation Mapping (Grad-CAM) to maximize model interpretability. Using a dataset of 1,437 color fundus images, we explored the potential of different pre-trained convolutional neural networks (CNNs), including Xception, VGG16, InceptionV3, and DenseNet121. Their respective accuracies on the test set were 89.27%, 91.44%, 89.06%, and 93.35%. Our objective was to improve the accuracy of diabetic retinopathy detection. We explored methods to combine predictions from these four models we began with weighted voting, which achieved an accuracy of 93.95%, and subsequently employed meta-learners, achieving an improved accuracy of 94.63%. These approaches surpassed individual models in distinguishing between non-proliferative and proliferative phases of DR. These findings underscore the potential of these approaches in developing robust diagnostic tools for diabetic retinopathy. Furthermore, techniques like Grad-CAM enhance interpretability, opening the door for further advancements in early-stage detection and clinical integration automatically while maximising accuracy and interpretability.**

Keywords*—*Diabetic retinopathy; retinal images; Grad-CAM; weighted voting; meta-learners

I. Introduction to Diabetic Retinopathy Detection

A. Background

Diabetes is one of the most prevalent diseases worldwide. It is a chronic condition characterized by elevated blood sugar levels (hyperglycemia) [1]. It involves the assimilation, utilization, and storage disorder of sugars in the diet.It occurs when the body fails to effectively utilize the insulin it produces or when the pancreas does not produce an adequate amount. Insulin is a hormone that is indispensable for the regulation of blood sugar levels by permitting glucose to access the body's cells. Excessive urine excretion, intense thirst, constant appetite, weight loss, impaired vision, and fatigue are among the most prevalent symptoms.In order to prevent long-term complications that affect various body systems, particularly nerves and blood vessels meticulous management is necessary.

According to a report by the World Health Organization (WHO) [2], more than 400 million people are suffering from diabetes in the world. It is anticipated that this figure will rise to 552 million in 2024. The World Health Organization also reports that diabetes is a significant cause of blindness, amputations, and mortality. One of the primary causes of blindness is diabetic retinopathy. More than 5 million people around the world with diabetes are blind. This number is expected to double by 2030. Research at a Jakarta government hospital in 2011 [3] indicated that the highest diabetes complication was neuropathy (54%), followed by diabetic retinopathy (33.4%) in second place.

The relationship between diabetes and diabetic retinopathy is both direct and causal. Diabetic retinopathy is an ocular complication caused directly by diabetes. The chronic hyperglycemia (high blood sugar levels) associated with diabetes damages the microscopic blood vessels in the retina [4]. This can lead to blood leakage, the formation of abnormal new blood vessels, and other alterations in the retina. The risk of developing diabetic retinopathy increases with the duration of diabetes and inadequate glycemic control. The longer the diabetes and the lesser the glycemic control, the higher the risk. Initially, diabetic retinopathy may be asymptomatic. However, as it progresses, it can cause mild to severe vision problems, even leading to blindness. A comprehensive eye examination by an optometrist can enable the early detection of diabetes and diabetic retinopathy, reducing the risk of visual loss. Early detection and effective management of diabetes, including excellent glycemic control, is essential to prevent or delay the progression of diabetic retinopathy. In summary, diabetic retinopathy is a direct and frequent consequence of diabetes, underscoring the significance of its detection and monitoring in preserving ocular health.

Deep learning algorithms have made significant advances in improving diabetic retinopathy detection in recent years, which can help physicians make informed decisions about the most effective treatment plan for each patient. This article will discuss the application of deep learning to detect diabetic retinopathy and the exploitation of some techniques that will help push the models performances in terms of diabetic retinopathy detection and other systems that will improve their results interpretability, It will also cover the performance metrics and validation techniques used to assess the efficacy of these models. Finally, we will examine how these results may influence future treatment procedures for diabetic retinopathy patients [5], [6].

The following sections of our paper are structured in a way that facilitates clarity and comprehensiveness. Section II provides a comprehensive description of the Dataset we used in our research. In Section III, we outline the approach we took to ensure accurate and reliable results. Section IV discusses the findings of our study and provides a comprehensive analysis of the used methodologies. We then delve deeper into the implications of our findings. Finally, in Section V, we present our conclusions and suggest potential avenues for future research.

B. Literature Review

Research on the classification of Diabetic Retinopathy (DR) has been extensive. Gondal et al. [7] introduced a CNN model with 93.6% sensitivity and 97.6% specificity , using Kaggle and DiaretDB1 datasets. Wang et al. [8] introduced a model that combined different networks and achieved AUC scores of 0.978 and 0.960 . Quelle et al. [9] worked on CNN models for binary classification and lesions detection. Chandrakumar and Kathirvel [10] achieved 94% accuracy on the DRIVE and STARE datasets using a CNN model with dropout regularization. Memon et al. [11] applied nonlocal mean denoising and brightness equalization, achieving a kappa score accuracy of 0.74. Pratt et al. [12] developed a CNN for five DR stages but struggled with mild stage classification due to dataset imbalance. Yang et al. [13] introduced a DCNN for normal and NPDR stages with lesion highlighting and grading. Garcia et al. [14] assessed CNN models (Alexnet, VGGnet16) on the Kaggle dataset, achieving 83.68% accuracy on VGG16. Dutta et al. [15] used the Kaggle dataset to assess three deep learning models, with the best training accuracy of 89.6% on a DNN. Recent advancements include Luo et al. [16], who proposed Multi-View DRD (MVDRNet) combining DCNNs and attention mechanisms, though it failed to train a network with lesion explanation. Chen [17] introduced a multi-scale shallow CNN model for early DR recognition, but it did not significantly improve classification precision. Martinez-Murcia et al. [18] created a CNN for routine DR diagnosis, which was not practical for clinical applications. Deepa et al. [19] created a Deep CNN (MPDCNN) for fundus image recognition, but it lacked advanced neural network architectures for more accurate detection. Das et al. [20] designed a deep learning architecture for DR categorization based on segmented fundus images, while Kalyani et al. [21] introduced a reformed capsule network for feature extraction from fundus images. Oh et al. [22] developed a novel DRD method using top fundus photography and deep learning techniques but did not effectively set an ROI for minimizing complexity. Erciyas and Barısci [23] applied deep learning techniques for automatic lesion detection through ROI extraction, but their method did not optimize system resource utilization.

To summarize, research on DR classification may be categorized into binary and multi-class classification. Binary classification is limited in assessing the severity of Diabetic Retinopathy (DR). On the other hand, multi-class classification categorizes Diabetic Retinopathy (DR) into five distinct phases. Currently used models face challenges with the learning of the abstracted characteristics out of the different stages the diabetic retinopathy and accurately categorizing the stages of diabetic retinopathy in the inference process , which is crucial for successful therapy outcomes.

In response to this, our study aims to identify the stages of diabetic retinopathy accurately. We focused on maximizing the accuracy of the models and providing an interpretability approach to maximize the potential of our approach and assist the medical staff in making decisions based on concrete factors [24], [25], [26].

C. Research Contribution

Our contribution involves a focused effort on enhancing the performance of our models to detect diabetic retinopathy more effectively compared to others addressing the same issue. Our primary goal was to improve the effectiveness of our models in identifying diabetic retinopathy. We adjusted our strategy to enhance the model's ability to differentiate between the three main phases of diabetic retinopathy [27] NDR (No Diabetic Retinopathy), NPDR (Non-Proliferative Diabetic Retinopathy), and PDR (Proliferative Diabetic Retinopathy). Exploring multiple models for the task allowed us to evaluate their effectiveness and refine our approach for greater efficiency.

In the following section, we delve into interpretability and visual assistance concepts. This is relevant because diabetic retinopathy is caused by high blood sugar levels, which degrade capillary walls and result in leakage. This leads to the rupture and bursting of retinal vessels. Our goal is to equip medical professionals with a visualization technique using our advanced deep learning models. This method emphasizes crucial areas to help doctors detect abnormalities and potential lesions by focusing on the regions influencing the model's prediction.

The combination of these two methods enables precise detection and the creation of a comprehensive interpretive framework. This progress will enable the establishment of a system that greatly aids doctors in timely and accurate diagnoses, resulting in improved patient outcomes and a reduced risk of vision loss. Furthermore, this technology promises to streamline the diagnostic process, enabling prompt treatment and intervention as necessary.

II. Dataset and Preparation

To realize our project, we sought a representative database [28], [29] that would encompass the various stages and manifestations of diabetic retinopathy. After extensive research, we selected a comprehensive dataset comprising 1437 color fundus images, meticulously collected and classified by expert ophthalmologists. Further details about the dataset, including its composition, and the preprocessing steps undertaken to prepare the data for analysis, will be covered in the next subsections. This comprehensive approach ensures that our study is based on reliable and clinically relevant data, enhancing the accuracy and applicability of our findings.

A. Dataset Composition

The dataset "Fundus Images for the Study of Diabetic Retinopathy" [28] comprises 1437 color fundus images acquired at the Department of Ophthalmology, Hospital de

Clínicas, Facultad de Ciencias Médicas, Universidad Nacional de Asunción, Paraguay. Created by a team of researchers and ophthalmologists, this dataset was collected using the Visucam 500 camera from Zeiss, following clinical procedures. Expert ophthalmologists have meticulously classified the images into seven categories: No DR signs (711 images), Mild NPDR (6 images), Moderate NPDR (110 images), Severe NPDR (210 images), Very Severe NPDR (139 images), PDR (116 images), and Advanced PDR (145 images). This classification aids in the detection and study of Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR) at various stages. The dataset is a valuable resource for researchers and clinicians focusing on the early detection and management of diabetic retinopathy.

B. Dataset Preparation

This section provides a comprehensive overview of the steps undertaken to collect, analyze, and preprocess the data. This includes detailed descriptions of the procedures for data cleaning, formatting, and transformation. We also address any data quality issues encountered during the process and explain how they were resolved. Our primary objective is to ensure that our data preparation process is clearly and transparently documented, supporting the accuracy, reliability, and reproducibility of our research findings. Through this meticulous approach, we aim to establish a robust foundation for our subsequent analysis, ensuring that the data used is of the highest quality and integrity.

1) Data exploration: For this research section, we delved into our database and analyzed the data through diverse charts and summaries. Fig. 1 enabled us to have quantitative measures to evaluate our dataset qualitatively. This visual representation allow us to easily identify trends and patterns within the data, helping us make informed decisions moving forward. By analyzing both the count and percentage of each stage, we can gain a comprehensive understanding of the distribution of diabetic retinopathy stages in our dataset.

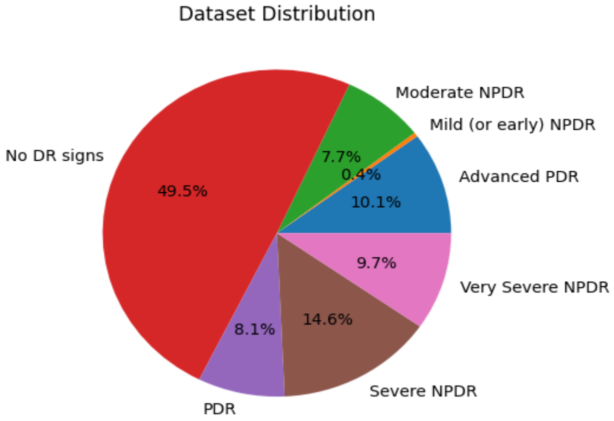


Fig. 1. Percentage distribution of diabetic retinopathy stages.

2) Data preprocessing: Analyzing the distribution graph reveals significant class imbalances. To address the primary stages of interest related to diabetic retinopathy, which include the no diabetic retinopathy (NDR) class, non-proliferative diabetic retinopathy (NPDR), and proliferative diabetic retinopathy (PDR), we merged the pathological sub-stages into these three main categories by assigning each sub-stage to its corresponding main category. This reorganization not only helps balance the distribution in our dataset but also aligns logically with the objectives of our research. The ultimate goal is to develop an automated diagnostic system that assists doctors in accurately predicting the stage of diabetic retinopathy, thereby improving clinical decision-making.

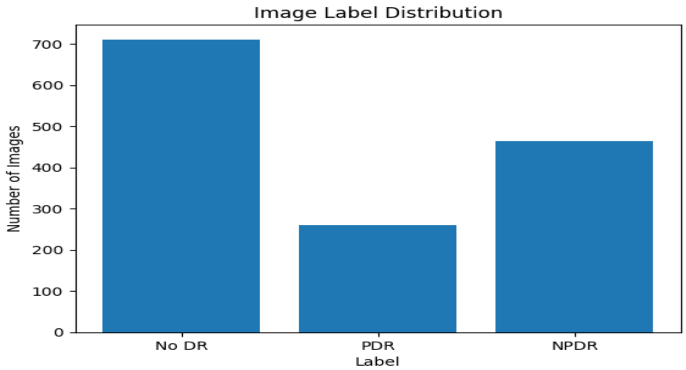


Fig. 2. Post-Merging label distribution of diabetic retinopathy stages.

After analyzing the Fig. 2 representing the distribution of labels post-merging, it is evident that the issue of class imbalance remains prevalent. This imbalance underscores the necessity of implementing data augmentation techniques to create a more balanced dataset, which is crucial for training our models effectively, as it helps prevent biases and improves the overall performance and generalization of the models. Therefore, we must incorporate appropriate data augmentation strategies to address this imbalance and enhance the robustness of our machine-learning models.

To tackle the class imbalance in our dataset, we employed several data augmentation techniques, including horizontal flip, vertical flip, 90-degree rotation, 180-degree rotation, 270-degree rotation, and zoom. These transformations were carefully selected to maintain the realistic characteristics of the original data while increasing the representation of underrepresented classes. By applying these augmentation techniques, we created an enhanced dataset that ensured each class had sufficient samples. The resulting distribution, depicted in Fig. 3, shows a significantly more balanced dataset. This balanced dataset facilitates more accurate and reliable model training, thereby improving performance and generalization.

The dataset was divided into 80% for training and 20% for testing. Subsequently, the training set was further split, allocating 80% for training and 20% for validation. The sample sizes for the three classes (NDR, NPDR, PDR) across the training, validation, and testing sets were as follows: training set - NDR: 1920, NPDR: 1785, PDR: 1001; validation set - NDR: 480, NPDR: 447, PDR: 251; testing set - NDR: 600, NPDR: 558, PDR: 314.

III. Proposed Approach

Developing different pathways to illustrate our models is the main objective of our research which aims at improving the

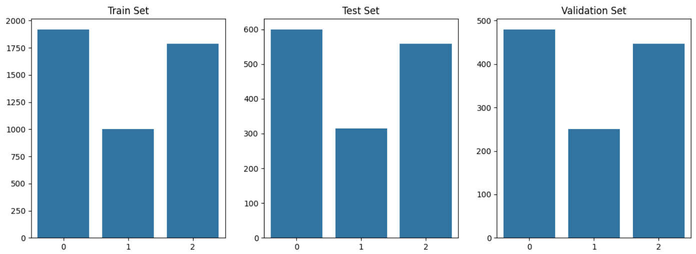


Fig. 3. Distribution of labels in training, testing, and validation sets after data augmentation.

performance and interpretability of predictive models. We will first train and investigate Gradient-weighted Class Activation Mapping (Grad-CAM) technique. It is a powerful visualization tool that allows us to understand how convolutional neural networks (CNNs) make decisions by showing which parts of the input image have the most influence on the model's predictions. After that we will use ensemble methods for enhancing model performance namely meta-learners for increased predictive accuracy. This approach not only reveals interpretative routes of our models but also boosts their prediction abilities significantly.

A. Models Training

1) Elaboration and evaluation of the base CNN model: In our initial training experiment, we designed the convolutional neural network (CNN) depicted in Fig. 4 using Keras to evaluate the model's performance on the retinal image dataset of images resized to 112x112x3. The network architecture included three convolutional layers, each with 64 filters, a 3x3 kernel size, and ReLU activation, followed by 2x2 max-pooling layers, and two fully connected layers with 128 units each, incorporating dropout regularization with a dropout rate of 0.5. The model was compiled with a set of hyper-parameters summarized in Table I below and to enhance the effectiveness of our training we used callbacks such as ReduceLROnPlateau to adjust the learning rate by a factor of 0.2 with a patience of 5 epochs (minimum learning rate of 0.0001), EarlyStopping to prevent overfitting with a patience of 10 epochs, and ModelCheckpoint to save the best model based on validation accuracy. Training was initially launched for 80 epochs with validation on a separate test set, providing insights into the model's generalization capabilities and guiding further optimization efforts.

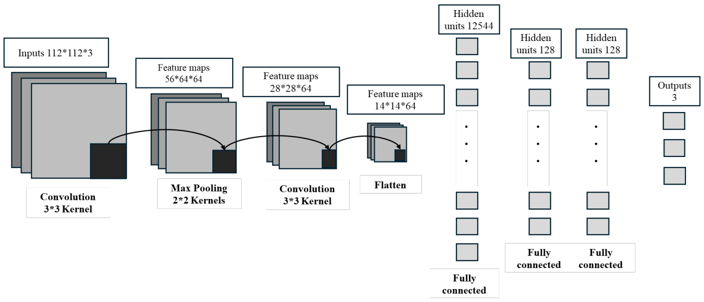


Fig. 4. Layered visualization of base convolutional neural network.

TABLE I. SUMMARY OF HYPERPARAMETERS USED DURING TRAINING

Hyperparameter	Value
Input Size	112×112×3
Batch Size	32
Number of Epochs	80
Learning Rate	0.001
Loss Function	Categorical Crossentropy
Optimizer	Adam

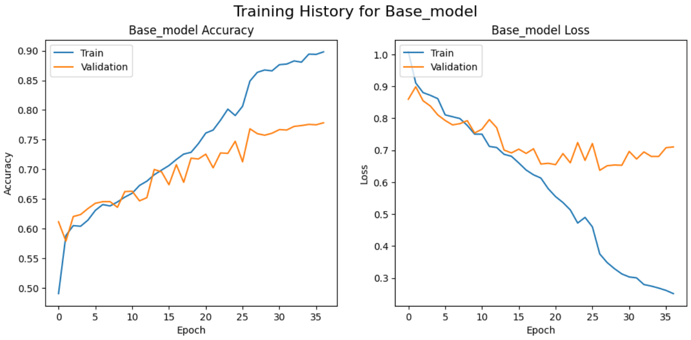


Fig. 5. Training and validation accuracy and loss for base CNN model.

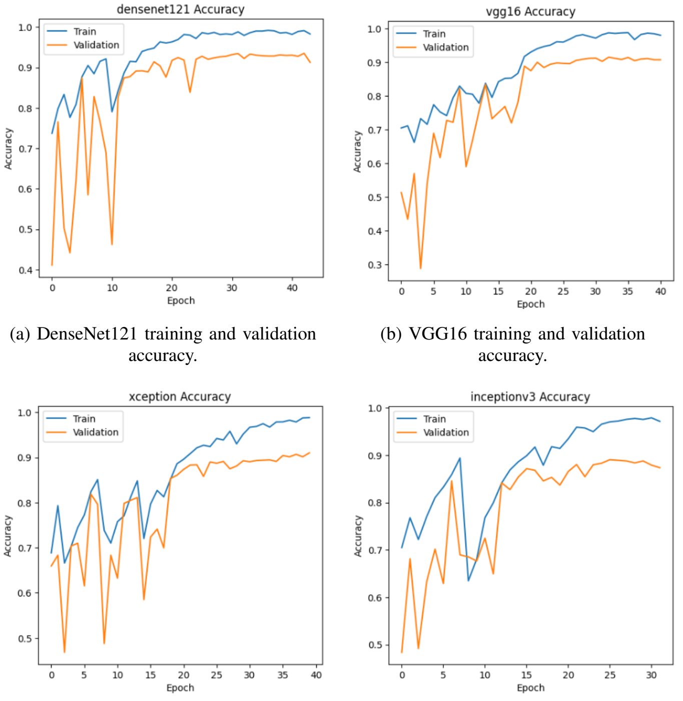
The results of our initial model training are depicted in Fig. 5, illustrating the performance evolution of our base model. The accuracy plot shows a consistent increase in training accuracy, surpassing 90% by the end of the training period, while the validation accuracy exhibits significant fluctuations and stabilizes around 75%. Concurrently, the loss plot indicates a steady decrease in training loss, reaching below 0.3, whereas the validation loss remains relatively high and variable, around 0.7. These observations suggest early signs of overfitting, possibly due to the limited size of our dataset.

To address this issue, we plan to leverage pretrained models for transfer learning. By utilizing a large, pretrained model, we can benefit from its learned representations, which are generally more robust and less prone to overfitting. Transfer learning allows us to adapt these prelearned features to our specific task, typically requiring less data and thereby reducing the risk of overfitting to our training set.

In the upcoming subsection, we will detail our implementation of transfer learning tailored to our task, aiming to maximize the accuracy of our models.

2) Exploring pretrained CNN models for improved performance: In the following section, we considered a selection of pretrained models for our project. The selection of the models was influenced by specific characteristics such as accuracy, architectural diversity, and efficiency as outlined in the table proposed by Keras [30] depicting models performances on the ImageNet validation dataset. Our choice included Xception [31], DenseNet121 [32], VGG16 [33], and InceptionV3 [34] for classifying diabetic retinopathy images, this choice is well-founded due to the high accuracy, architectural diversity, and efficiency of these models, which are essential for effective classification of diabetic retinopathy images. Xception's depth-wise separable convolutions ensure detailed feature extraction with reduced computational complexity, vital for detecting subtle variations in retinal images. DenseNet121's densely connected layers maximize information flow and feature reuse, enhancing the learning of complex patterns crucial for accurate diagnosis. VGG16's straightforward architecture contributes to robust performance, making it a practical choice for clinical applications due to its ability to deliver consistent results. InceptionV3 achieves a balance between high accuracy and efficiency by utilizing factorized convolutions and dimensionality reduction techniques, which enable effective analysis even with limited computational resources. Together, these models offer a comprehensive approach, combining high performance, diverse methodologies, and operational flexibility, essential for reliable and precise diabetic retinopathy classification.

In our transfer learning experiment, we retrained the four selected models, each adapted for a three-class classification task. The training protocol was meticulously designed to optimize model performance. Each model was initialized with ImageNet weights and fine-tuned by replacing the original classification head with a dense layer of size 3, followed by a softmax activation. The training was conducted using the Adam optimizer with a learning rate of 0.001, minimizing categorical crossentropy loss. The training was initialized for 80 epochs, with the following key hyperparameters that included a batch size of 32, a dropout rate of 0.5 to prevent overfitting, and L2 regularization with a lambda of 0.01. We employed ReduceLROnPlateau to reduce the learning rate by 20% if the validation loss plateaued over 5 epochs (minimum learning rate of 0.0001), EarlyStopping to halt training if validation loss stagnated for 10 epochs, and ModelCheckpoint to save the best-performing model based on validation accuracy. The models were evaluated on a the test validation set, with accuracy and loss monitored throughout the training to assess performance.



(a) DenseNet121 training and validation accuracy.

(b) VGG16 training and validation accuracy.

(c) Xception training and validation accuracy.

(d) Inception training and validation accuracy.

Fig. 6. Training and validation evolution of pre-trained models.

The Fig. 6 shows how the four models evolve during the training process on our Diabetic Retinopathy dataset. It illustrates a rapid increase in accuracy in the initial epochs, followed by stabilization at high levels. While the training accuracy of all models rapidly reaches near-perfect levels, the validation accuracy stabilizes slightly lower, typically around 0.85 to 0.90. During the early stages, noticeable fluctuations in both training and validation accuracies are observed, especially within the first 5 to 10 epochs of training. These initial fluctuations can be attributed to the fine-tuning process, where the models, pre-trained on a different task with a different set of images, are adjusting their weights to accommodate the new data. The fluctuations are likely a result of the models attempting to balance learning new features specific to the new dataset while retaining the generalized knowledge acquired from their pre-training. As training continues, the models adapt gradually, resulting in fewer fluctuations and increased stability in accuracy. The results obtained of this operation were as follows, VGG16 demonstrated strong performance, reaching a peak validation accuracy of 91.44% at epoch 33. Despite initial success, the model exhibited some fluctuations in validation accuracy, showing it was affected by changes in the learning rate. However, the overall stability and high accuracy make VGG16 a robust choice for our classification task. DenseNet121 showed impressive results, achieving the highest validation accuracy of 93.55% at epoch 43. This model exhibited consistent performance with minor variations in accuracy and loss, reflecting its ability to capture intricate features effectively. InceptionV3 achieved a peak validation accuracy of 92.80% at epoch 38, performing well on the complex data and competing effectively with other models. Xception achieved a peak validation accuracy of 92.35% at epoch 36, showcasing high accuracy and efficient performance, with slight variations in precise validation accuracy and loss metrics.

These results underscore the efficacy of transfer learning in leveraging pretrained models for specialized classification tasks, demonstrating significant potential for diabetic retinopathy classification task.

B. Enhancing Model Interpretability with Grad-CAM

After training models with enhanced accuracy in detecting diabetic retinopathy, it is crucial to implement methods that boost the visibility and interpretability of these deep learning predictions. Such methods enable practitioners to visually interpret model predictions, highlighting areas of interest that may require further examination. Integrating Gradient-weighted Class Activation Mapping (Grad-CAM) with our most precise model, the Densenet model, offers a robust solution.

Grad-CAM is a technique used to produce visual explanations for the decisions made by convolutional neural networks (CNNs). It works by utilizing the gradients of a target concept, such as diabetic retinopathy, flowing into a convolutional layer of the CNN. By calculating these gradients, Grad-CAM generates a heatmap that shows which regions of the input image are most influential in the model's decision-making process. These heatmaps effectively pinpoint the critical areas within retinal images that the model considers important for diagnosing diabetic retinopathy.

In our approach to generate these Grad-CAM heatmaps, we first constructed a gradient model that outputs the activations from a convolutional layer and the model's predictions. We employed TensorFlow's GradientTape to capture the gradients of the predicted class score with respect to these activations. By computing the mean intensity of the gradients for each output channel of the convolutional layer, we obtained the importance weights. These weights were then applied to the feature maps, resulting in a heatmap that highlights the regions in the image most relevant to the prediction.

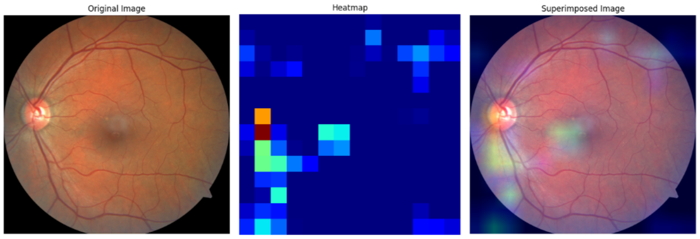


Fig. 7. Grad-CAM visualization for No Diabetic Retinopathy (NDR).

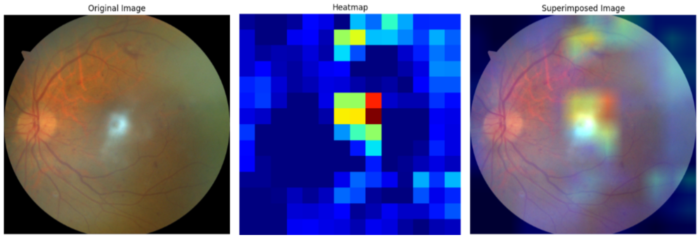


Fig. 8. Grad-CAM visualization for Non-Proliferative Diabetic Retinopathy (NPDR).

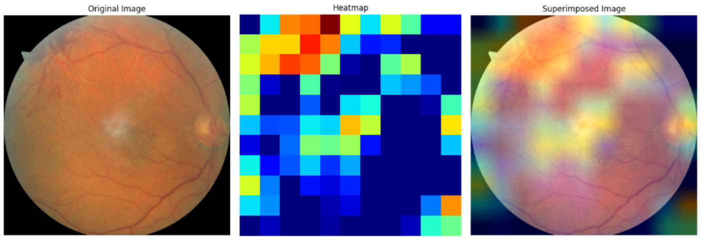


Fig. 9. Grad-CAM visualization for Proliferative Diabetic Retinopathy (PDR).

To visualize the highlighted regions, as depicted in Fig. 7, 8 and 9, which represent the output of our approach for different stages of diabetic retinopathy, we followed a systematic process. First, we loaded the original image and resized the heatmap to match its dimensions. The heatmap was then colored using the jet colormap, scaled appropriately, and superimposed onto the original image with a specified level of transparency. Finally, we displayed the original image, the heatmap, and the superimposed image side-by-side to provide a comprehensive visual analysis of the model's focus areas. This approach allows for a clear and detailed examination of the regions identified by the model, aiding in the interpretation and validation of its predictions.

The three figures illustrate examples of diabetic retinopathy classes: no diabetic retinopathy (NDR), non-proliferative diabetic retinopathy (NPDR), and proliferative diabetic retinopathy (PDR). In each figure, the image on the right shows the original retinograph, commonly used in medical practice. The middle image represents the heatmap, also known as class activation maps (CAM). These maps highlight critical regions in an image responsible for specific predictions made by a convolutional neural network (CNN), obtained by analyzing the flow of gradients in a CNN layer. These maps demonstrate how specific image regions influence the model's predictions.

The Fig. 7 shows a case without diabetic retinopathy, where the model uses the main blood vessels as key features to classify a patient as healthy and without diabetic retinopathy.

The Fig. 8 presents an example of non-proliferative diabetic retinopathy (NPDR), also known as early-stage diabetic retinopathy. This condition results in increased capillary permeability, microaneurysms, hemorrhages, exudates, and complications such as macular ischemia and edema. Our model accurately identifies these features in the three regions of interest, depicted in the left image by the overlay of the original image and the class activation map.

The Fig. 9 illustrates a case of proliferative diabetic retinopathy (PDR), which develops after NPDR and is more severe. It is characterized by the growth of new blood vessels, often accompanied by fibrous tissue growth in front of the retina. These new blood vessels can also form in the front part of the eye, including the iris, contributing to severe vision loss in proliferative retinopathy. These various symptoms are well detected by our model, justifying the extensive regions of interest given the presence of microvascularizations and clear lesions in the example image represented by Fig. 9.

Examining the examples clearly shows the value of using GradCAM to pinpoint specific areas of abnormality, enabling the development of targeted detection and treatment procedures. This advanced technology improves the accuracy and efficiency of detecting retinopathy, leading to better patient outcomes.

C. Boosting Model Performance through Ensemble Methods

In this section, our goal is to enhance the predictive capabilities of our models. To achieve this, we will investigate the collective capabilities of our trained models. Fig. 10 illustrates the conceptual diagram of our proposed approach, leveraging the potential of our four trained models and the generated GRAD-CAM to achieve a balance of precision and interpretability.

Our goal is to enhance prediction accuracy and dependability by leveraging the combined strengths of these models. Fig. 11 illustrates the conceptual diagram of the proposed ensemble learning prediction generation process. By adopting this strategy, we aim to utilize the variety and complementing qualities of each model, resulting in a stronger and more reliable prediction model.

This approach allows us to make informed decisions based on the insights provided by each model, leading to more robust predictions. Through this ensemble learning technique, we can maximize the potential of our models and improve overall performance.

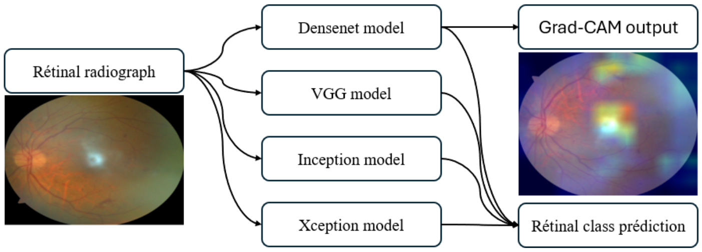


Fig. 10. Conceptual diagram of the proposed approach.

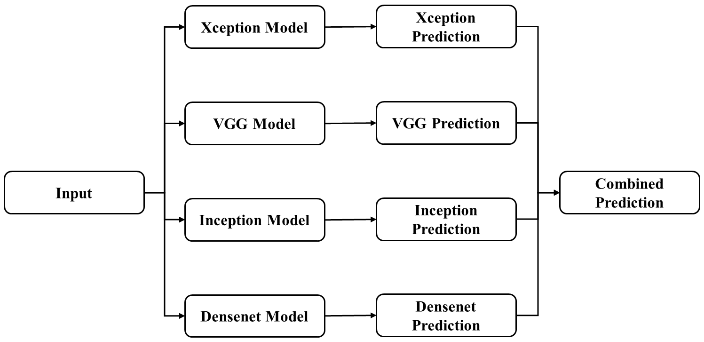


Fig. 11. Conceptual diagram of the proposed ensemble learning prediction.

1) Improving accuracy with weighted ensemble predictions: Within this section, we employed a weighted ensemble of our four trained models to enhance their performance in classifying diabetic retinopathy. The process operates as follows: each model generates predictions on the test samples, and we assess the correctness of each model. Subsequently, these accuracies are utilized to allocate weights to the models, with the more precise models being assigned larger weights. The weights are normalized to ensure that their sum is equal to one.

To get the ultimate forecast for each test sample, I employ a weighted majority voting approach. The predicted result of each model is multiplied by its corresponding weight, and the class with the greatest weighted vote is selected as the ultimate prediction. This approach leverages the advantages of each individual model, leading to a more resilient and precise total prediction.

Mathematically, this can be described as follows:

Let M_i denote the i-th model, and let p_{ij} be the prediction of model M_i for sample j. The accuracy a_i of model M_i is used as its weight w_i, where w_i is normalized so that $\sum_{i=1}^{N} w_i = 1$.

The weighted vote $v_j(k)$ for class k for sample j is given by:

$$v_j(k) = \sum_{i=1}^{N} w_i \cdot I(p_{ij} = k) \tag{1}$$

where, I is the indicator function, which is 1 if $p_{ij} = k$ and 0 otherwise.

The final prediction $\hat{y}_j$ for sample j is:

$$\hat{y}_j = \arg\max_k v_j(k) \tag{2}$$

The following Table II displays the precision of each individual model as well as the collective ensemble model. The findings reveal that the ensemble model obtained a greater accuracy than any of the individual models, proving the usefulness of the weighted ensemble technique.

TABLE II. PERFORMANCE METRICS OF INDIVIDUAL AND ENSEMBLE MODEL

Model	Precision	Recall	F1-Score	Accuracy
Xception Model	0.892542	0.892663	0.892164	0.892663
Vgg Model	0.914305	0.914402	0.914229	0.914402
Inception Model	0.890399	0.890625	0.890181	0.890625
Densenet Model	0.937100	0.935462	0.935564	0.935462
Ensemble (Weighted Voting)	0.939495	0.939538	0.939275	0.939538

By adopting this weighted ensemble technique, the combined model leverages the capabilities of each individual model, leading to increased classification performance, as indicated by the ensemble accuracy of 0.9395.

2) Advanced ensemble technique stacking generalization: Stacking generalization is a sophisticated ensemble learning technique designed to enhance predictive performance by integrating the outputs of numerous base models. Unlike simple averaging or majority voting, stacking involves building a meta-model to learn the best method to combine the predictions of base models. This approach leverages the capabilities of each individual model, leading to more accurate and robust predictions.

The stacking process begins with training multiple base models independently on the training data. These base models can be of various types, or they can be the same type with varying hyperparameters or training sets. Each model is trained to optimize its performance on the given data, creating a diverse set of models with unique strengths and limitations.

Once the base models are trained, the next stage is to generate meta-features. This involves using the trained base models to make predictions on the training set. The predictions from each base model are put together, and if each model outputs a probability for each class, these probabilities are used as meta-features. For instance, in our classification problem with three classes and four base models, each model will output three probabilities per sample, resulting in a total of 12 meta-features per sample (4 models × 3 classes).

The collected meta-features are then used to train a new model, known as the meta-model. The meta-model is trained on the meta-features generated from the training set, learning how to combine the predictions of the basis models to create the final prediction. Common choices for meta-models include logistic regression, random forests, or another neural network, etc. The meta-model aims to capture intricate relationships between the base model predictions that simple averaging or voting are unable to capture.

Finally, the trained meta-model is used to make final predictions based on the meta-features of the test set. The meta-model processes these meta-features and outputs the final prediction for each sample. This final phase ensures that the strengths of each base model are effectively combined to produce the most accurate predictions.

In our application of this approach, we have leveraged the four models we previously trained and utilized the meta-characteristics generated by these retrained models to train a set of meta-learners. Specifically, we investigated decision trees, multi-layer perceptrons (MLP), Naive Bayes, k-nearest neighbors (KNN), support vector machines (SVM), random forests, and logistic regression. The results obtained from these meta-learners are summarized in the table above. To analyze and determine the best model among those evaluated, we need to consider various factors such as accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. Precision is crucial when the cost of false positives is high, as it indicates the proportion of true positive predictions out of all positive predictions made by the model. Recall, also known as sensitivity, is essential in scenarios where false negatives are particularly costly, as it measures the proportion of actual positives that are correctly identified. The F1-score provides a balanced assessment by considering both precision and recall, making it particularly useful when dealing with imbalanced datasets. This comprehensive evaluation will ensure that we identify the most effective model for our specific application.

In order to acquire the findings represented in Table III, we extensively fine-tuned the hyperparameters of each meta-learner. By varying parameters such as learning rate, number of estimators, kernel types, etc. we improved the performance of each model to reach the best potential outcome. The fine-tuning process comprises iterative testing and validation to establish the highest performing meta-models.

For logistic regression, the optimal combination of hyperparameters was determined to be $C = 10$ and $penalty = l2$, achieving an accuracy of 0.9436. Here, C controls the inverse of the regularization strength, with smaller values indicating stronger regularization, and $penalty$ specifies the norm used in the penalization .

In the case of the random forest, the best performance was obtained with $max\_depth = 20$ and $n\_estimators = 10$, resulting in an accuracy of 0.9341. The max_depth parameter limits the number of levels in each decision tree to prevent overfitting, while $n\_estimators$ defines the number of trees in the forest.

For the support vector machine (SVM), using $C = 10$ and a $kernel = linear$ yielded the highest accuracy at 0.9450. The C parameter is a regularization parameter, and the $kernel$ parameter specifies the kernel type used in the algorithm, with “linear” indicating a linear kernel.

The k-nearest neighbors (KNN) model performed best with $n\_neighbors = 10$ and $weights = uniform$, achieving an accuracy of 0.9429. The $n\_neighbors$ parameter determines the number of neighbors to use, and the $weights$ parameter indicates how the influence of the neighbors is weighted, with “uniform” meaning all neighbors are weighted equally.

The naive Bayes model, which did not require any hyperparameter tuning, reached an accuracy of 0.9212. Naive Bayes models typically do not have tunable hyperparameters in their basic form.

Lastly, the multi-layer perceptron (MLP) showed optimal performance with $activation = relu$ and $hidden\_layer\_sizes = (100,)$, resulting in an accuracy of 0.9450. The $activation$ parameter specifies the activation function for the hidden layer, with “relu” standing for Rectified Linear Unit, and $hidden\_layer\_sizes$ defines the number of neurons in each hidden layer, with (100,) indicating one hidden layer with 100 neurons.

The decision tree model achieved an accuracy of 0.9307 with $max\_depth = 10$. The max_depth parameter limits the number of levels in the tree, helping to prevent overfitting.

The KNN meta-learner scored the greatest accuracy of 0.9463, with good precision and F1-Scores across all classes, indicating a well-balanced and robust performance. Logistic Regression and Random Forest followed closely with accuracies of 0.9436 and 0.9429, respectively, indicating equal performance in precision, recall, and F1-Score. Naive Bayes and Decision Tree models revealed lower accuracies of 0.9212 and 0.9192, with higher variability in precision between classes.

In summary, stacking is a potent strategy in ensemble learning that may considerably boost prediction performance. By training numerous base models and a meta-model to integrate their outputs, stacking effectively exploits the capabilities of each model as what was proved in our experience. the greatest accuracy of our base models was of 0.9355 and were pushed to 0.9463 performed by the KNN meta-learner resulting to a more precise and reliable predictions wich was the purpose of this investigation .

IV. Experimental Results and Comprehensive Discussion

A. Experimental Results

Our investigation first started by structuring the study area in terms of three classes of important relevance in clinical practice: NDR, NPDR and PDR, followed by the training of a basic convolutional neural network (CNN) using Keras to test its performance on retinal radiographs downsized to 112x112x3. The CNN design featured three convolutional layers with 64 filters, each employing a 3x3 kernel size and Rectified Linear Unit (ReLU) activation functions. This was followed by 2x2 max-pooling layers and two completely linked layers, each with 128 units. We applied dropout regularization with a dropout rate of 0.5 to avoid overfitting. The model was constructed using categorical crossentropy loss as the loss function and employed the Adam optimizer with a learning

TABLE III. Performance Metrics of Meta-Learners in Ensemble Approach

Meta-Learner	Meta-Model Accuracy	Precision (Class 0)	Precision (Class 1)	Precision (Class 2)	Recall	F1-Score
Logistic Regression	0.943614	0.962046	0.932886	0.929577	0.943614	0.937721
Random Forest	0.942935	0.954323	0.936242	0.934046	0.942935	0.937622
SVM	0.942255	0.958746	0.933333	0.929329	0.942255	0.937170
KNN	0.946332	0.960461	0.937086	0.936057	0.946332	0.941698
Naive Bayes	0.921196	0.962712	0.802778	0.955939	0.921196	0.912088
MLP	0.940897	0.954173	0.935811	0.929204	0.940897	0.935344
Decision Tree	0.919158	0.919094	0.944637	0.906195	0.919158	0.916664

rate set to 0.001. To improve the training process, we implemented callbacks such as ReduceLROnPlateau, EarlyStopping, and ModelCheckpoint. Training ran for 80 epochs, including validation on a unique test set to measure the model's generalization.

The results of our initial model training, depicted in Fig. 5, showed a consistent increase in training accuracy, surpassing 90% by the end of the training period. However, the validation accuracy showed notable fluctuations before stabilizing at around 75%. Concurrently, the training loss steadily decreased, while the validation loss remained relatively high and variable, suggesting early signs of overfitting due to the limited dataset size. To mitigate this, we utilized pretrained models for transfer learning by adapting Xception, VGG16, InceptionV3, and DenseNet121 for a three-class classification task. Each model was initialized with ImageNet weights and fine-tuned by replacing the original classification head with a dense layer of size 3, followed by a softmax activation. Training was conducted using the Adam optimizer with a learning rate of 0.001, a batch size of 32, a dropout rate of 0.5, and L2 regularization with a lambda of 0.01. We employed ReduceLROnPlateau, EarlyStopping, and ModelCheckpoint to optimize the training process.

The training evolutions of these models, depicted in Fig. 6, revealed that VGG16 achieved a peak validation accuracy of 91.44%, DenseNet121 reached the highest validation accuracy of 93.55%, InceptionV3 attained 92.80%, and Xception achieved 92.35%. These results underscore the efficacy of transfer learning in leveraging pretrained models for specialized classification tasks. Subsequently, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM) to improve the interpretability of the models predictions. Grad-CAM generates heatmaps that highlight regions of input images most influential in the model's decision-making process. Fig. 7, 8, and 9 show Grad-CAM outputs for various stages of diabetic retinopathy. These show where the model's focus is and help in understanding and confirming its predictions.

To further enhance the predictive capabilities of our models, we explored ensemble methods. We employed a weighted ensemble of our four trained models, allocating weights based on each model's accuracy. The combined model achieved an accuracy of 93.95%, outperforming the individual models. We explored stacking generalization, an advanced ensemble learning technique that combines the outputs of multiple base models through a meta-model. We trained several meta-learners, including decision trees, multi-layer perceptrons, Naive Bayes, k-nearest neighbors (KNN), support vector machines (SVM), random forests, and logistic regression. The KNN meta-learner achieved the highest accuracy of 94.63%, showcasing superior performance in precision and F1-scores across all classes. Logistic regression and random forest followed closely, with accuracies of 94.36% and 94.29%, respectively.

The workflow outlined in our study, as illustrated by the flowchart in Fig. 12, encapsulates the systematic approach we employed to optimize diabetic retinopathy classification. By combining data preprocessing, transfer learning with multiple deep learning models, and advanced ensemble methods, we were able to progressively enhance model accuracy and robustness. The flowchart also highlights the use of interpretability tools like Grad-CAM, which provided critical insights into the model's decision-making process. This visual representation underscores the complexity and integration of the methodologies discussed, offering a clear, step-by-step view of how each component contributed to the overall success of our approach. This structured methodology not only improved classification performance but also ensured that our models are interpretable and clinically relevant, paving the way for their potential application in real-world settings.

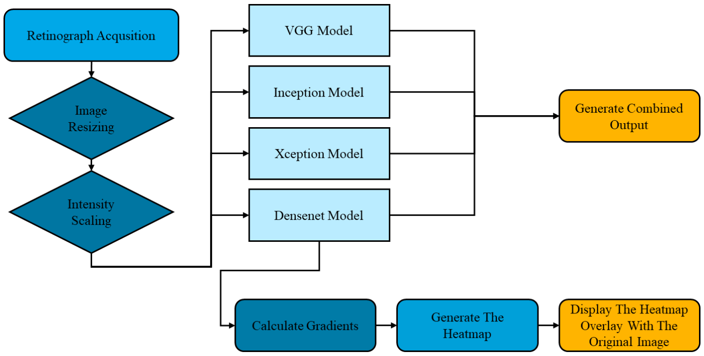


Fig. 12. Workflow for diabetic retinopathy classification using deep learning and ensemble techniques.

B. Discussion

The outcomes of our experiments were the materialization of our conceptual method illustrated in Fig. 10, aiding medical professionals in achieving both precise classification of diabetic retinopathy and visual assistance, thereby providing dual support for clinical decision-making.

Our experiments yielded several important discoveries that can be summarized as follows:

1) Transfer learning effectiveness: Pretrained models, particularly DenseNet121, substantially enhanced classification performance compared to the initial base model. This demonstrates the value of utilizing pretrained networks.

2) Enhanced performance through ensemble methods: Both weighted ensemble and stacking generalization techniques effectively boosted the predictive accuracy of our models. Notably, the KNN meta-learner achieved the highest performance, showcasing the power of combining multiple models to capture diverse patterns in the data.

3) Interpretability through Grad-CAM: The Grad-CAM visualizations provided meaningful insights into the decision-making process of the models, enhancing the interpretability and trustworthiness of the predictions.

V. Conclusion and Perspectives

In conclusion, our comprehensive approach integrated transfer learning, Grad-CAM for interpretability, and ensemble methods, resulting in significant improvements in the performance and reliability of our predictive models. Testing demonstrated that pre-trained models, advanced visualization techniques, and sophisticated ensemble strategies markedly enhance deep learning models for classifying diabetic retinopathy. This methodology not only improved model accuracy but also strengthen reliability.

Overall, our approach underscored the importance of leveraging diverse deep learning techniques to elevate the performance of predictive models in medical image classification. By incorporating these methods, we achieved substantial gains in accuracy and reliability for detecting diabetic retinopathy. The combination of these techniques not only enhances their applicability in identifying abnormalities but also paves the way for discovering new diseases and developing treatment strategies. Advanced visualization techniques empower medical professionals to visually cluster abnormalities, ensuring the development of robust and dependable systems that minimize flaws inherent in automated methods.

References

[1] M. Z. Banday, A. S. Sameer, and S. Nissar, *Pathophysiology of diabetes: An overview*, Avicenna Journal of Medicine, vol. 10, no. 4, pp. 174-188, Oct. 2020, doi: 10.4103/ajm.ajm_53_20.

[2] World Health Organization, *Diabetes*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: Jul. 5, 2024].

[3] P. Soewondo, A. Ferrario, and D. L. Tahapary, *Challenges in diabetes management in Indonesia: a literature review*, Global Health, vol. 9, p. 63, Dec. 2013, doi: 10.1186/1744-8603-9-63.

[4] National Eye Institute, *Diabetic Retinopathy*, 2023. [Online]. Available: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy. [Accessed: Jul. 6, 2024].

[5] O. Manchadi, F. E. Ben-Bouazza, and B. Jioudi, *Predictive Maintenance in Healthcare System: A Survey*, IEEE Access, vol. 11, pp. 61313-61330, 2023, doi: 10.1109/ACCESS.2023.3287490.

[6] O. Manchadi, F. E. Ben-Bouazza, Z. El Otmani Dehbi, A. Edder, I. Tafala, M. Et-Taoussi, and B. Jioudi, *An Internet of Things-based Predictive Maintenance Architecture for Intensive Care Unit Ventilators*, International Journal of Advanced Computer Science and Applications, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150294.

[7] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch, *Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images*, in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 2069–2073, Sep. 2017.

[8] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, *Zoom-in-net: Deep mining lesions for diabetic retinopathy detection*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 267–275, Springer, Berlin, Germany, 2017.

[9] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, *Deep image mining for diabetic retinopathy screening*, Medical Image Analysis, vol. 39, pp. 178–193, Jul. 2017.

[10] T. Chandrakumar and R. Kathirvel, *Classifying diabetic retinopathy using deep learning architecture*, International Journal of Engineering Research and Technology, vol. 5, no. 6, pp. 19–24, Jun. 2016.

[11] W. R. Memon, B. Lal, and A. A. Sahto, *Diabetic retinopathy*, The Professional Medical Journal, vol. 24, no. 2, pp. 234–238, 2017.

[12] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, *Convolutional neural networks for diabetic retinopathy*, Procedia Computer Science, vol. 90, pp. 200–205, Dec. 2016.

[13] D. Yang, X. Lin, Y. Gao, and X. Wu, *Lesion detection for diabetic retinopathy based on convolutional neural networks*, in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2630–2634, Sep. 2017.

[14] P. Garcia, C. Rodriguez, and S. Martinez, *Detection of diabetic retinopathy using a convolutional neural network*, Journal of Computer and Communications, vol. 5, pp. 1–7, 2017.

[15] S. Dutta, B. C. Manideep, S. M. Basha, R. D. Caytiles, and N. C. S. N. Iyengar, *Classification of diabetic retinopathy images by using deep learning models*, International Journal of Grid and Distributed Computing, vol. 11, no. 1, pp. 89–106, Jan. 2018.

[16] J. Luo, H. Zhang, and Y. Li, *Multi-View Diabetic Retinopathy Detection: A Method Combining DCNNs and Attention Mechanisms*, IEEE Transactions on Medical Imaging, vol. 40, no. 2, pp. 407–418, 2021.

[17] L. Chen, *Early recognition of diabetic retinopathy using a multi-scale shallow convolutional neural network*, Journal of Healthcare Engineering, vol. 2020, pp. 1–9, 2020.

[18] F. J. Martinez-Murcia and J. Ortuno, *Regular CNN for Routine Diabetic Retinopathy Diagnosis: Challenges and Solutions*, Computers in Biology and Medicine, vol. 131, p. 104245, 2021.

[19] P. Deepa, G. Selvaraj, and R. T. Selvi, *Efficient Deep Convolutional Neural Network (MPDCNN) for Fundus Image Recognition*, Journal of Digital Imaging, vol. 35, pp. 312–325, 2022.

[20] A. Das, K. Roy, and K. Biswas, *Segmented fundus images for the detection of diabetic retinopathy using a deep learning architecture*, Journal of Digital Imaging, vol. 34, pp. 735–746, 2021.

[21] D. Kalyani and G. Rao, *A reformed capsule network for feature extraction from fundus images*, IEEE Transactions on Medical Imaging, vol. 42, no. 4, pp. 1032–1043, 2023.

[22] S. Oh, J. Kim, and K. Lee, *Fundus photography and deep learning for diabetic retinopathy detection*, IEEE Transactions on Medical Imaging, vol. 40, no. 10, pp. 2472–2482, 2021.

[23] K. Erciyas and N. Barısci, *ROI extraction and deep learning techniques for automatic lesion detection*, Journal of Digital Imaging, vol. 34, pp. 1236–1245, 2021.

[24] S. Azeroual, F. E. Ben-Bouazza, A. Naqi, and R. Sebihi, *Predicting disease recurrence in breast cancer patients using machine learning models with clinical and radiomic characteristics: a retrospective study*, J Egypt Natl Canc Inst, vol. 36, no. 1, p. 20, Jun. 2024, doi: 10.1186/s43046-024-00222-6.

[25] I. Tafala, F.-E. Ben-Bouazza, O. Manchadi, M. Et-Taoussi, and B. Jioudi, *Cephalometric Landmarks Identification Through an Object Detection-based Deep Learning Model*. [Online]. Available: https://api.semanticscholar.org/CorpusID:268418674.

[26] A. Edder, F.-E. Ben-Bouazza, and B. Jioudi, *SkinNet: Enhancing Dermatological Diagnosis Through a New Deep Learning Framework*, in *International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023)*, M. Ezziyyani, J. Kacprzyk, and V. E. Balas, Eds., Springer Nature Switzerland, Cham, pp. 173–188, 2024, doi: 10.1007/978-3-031-52388-5-17.

[27] Y. Guex-Crosier and F. Behar-Cohen, *Ophtalmologie: Rétinopathie diabétique : nouvelles possibilités thérapeutiques*, Rev Med Suisse, vol. 11, no. 45657, pp. 101–107, 2015, doi: 10.53738/REVMED.2015.11.456-57.0101.

[28] V. E. Castillo Benítez, I. Castro Matto, J. C. Mello Román, J. L. Vázquez Noguera, M. García-Torres, J. Ayala, D. P. Pinto-Roa, P. E. Gardel-Sotomayor, J. Facon, and S. A. Grillo, *Dataset from fundus images for the study of diabetic retinopathy*, Data in Brief, vol. 36, p. 107068, 2021.

[29] V. E. Castillo Benítez, I. Castro Matto, J. C. Mello Román, J. L. Vázquez Noguera, M. García-Torres, J. Ayala, D. P. Pinto-Roa, P. E. Gardel-Sotomayor, J. Facon, and S. A. Grillo, *Dataset from fundus images for the study of diabetic retinopathy*, Zenodo, 2021, doi: 10.5281/zenodo.4647952.

[30] F. Chollet *et al.*, *Keras Applications API*, 2023. [Online]. Available: https://keras.io/api/applications/. [Accessed: May 22, 2024].

[31] F. Chollet, *Xception: Deep Learning with Depthwise Separable Convolutions*, 2017. [Online]. Available: https://arxiv.org/abs/1610.02357.

[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely Connected Convolutional Networks*, 2018. [Online]. Available: https://arxiv.org/abs/1608.06993.

[33] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, 2015. [Online]. Available: https://arxiv.org/abs/1512.00567.