Deep Learning-Based Diabetic Retinopathy Detection and Classification

Vijay V. Chakole [1], Prasad R. Shastrakar [2], Priyanka N. Khodake [3], and Mohit R.Mohare[4]
[1]*Professor, Department of Electronics and Telecommunication Engineering, KDK College of Engineering, Nagpur, Maharashtra, India*
[2,3,4]*Student, Department of Electronics and Telecommunication Engineering, KDK College of Engineering, Nagpur, Maharashtra, India*

***Abstract*— One of main causes of blindness in world, Diabetic Retinopathy (DR), predominantly affects people with diabetes. Preventing eyesight loss requires early identification and prompt treatment, but manual screening techniques are frequently laborious and prone to mistakes. Rapid developments in Deep Learning, particularly Convolutional Neural Networks, have demonstrated significant potential for enhancing also automating diagnosis for diabetic retinopathy from retinal pictures. With an emphasis on comparative model performance and accessibility for medical professionals, this study attempts to improve the diagnostic procedure for diabetic retinopathy by developing and integrating DL models. To diagnose diabetic retinopathy, we compare seven pre-trained deep-learning models. Popular designs including ResNet50, VGG16, InceptionV3, DenseNet121, Xception, and EfficientNetB0 are included in these models. Each model is refined using a varied data-set of diabetic retinopathy that includes images classified into sub classes including normal, cataract, glaucoma, and various phases of the disease. This is accomplished by utilizing transfer learning. We analyze the models' robustness and applicability for diagnosing diabetic retinopathy employing a variety of performance indicators, such as accuracy, precision, recall, F1-score, as well as AUC-ROC.**

Keywords — Convolutional Neural Networks (CNNs), DenseNet121, Image Processing, InceptionV3, ResNet50V2, VGG16, VGG19.

I. INTRODUCTION

Among the most prevalent deleterious complications of diabetes mellitus is diabetic retinopathy (DR), which remains one of the principal causes of blindness and vision loss, especially among those of productive age. Diabetes- associated retinopathy is public health concern that warrants early identification and treatment to prevent permanent loss of vision as diabetes prevalence increases globally. Manual screening methods such as fundus photography, where trained ophthalmologists examine retinal photographs to identify early signs of the disease, have traditionally been employed to diagnose by DR Abdur Rahman Onik[1]. This approach is difficult to scale and susceptible to human error because of the large number of patients and the unavailability of trained experts. In addition, it is not possible to achieve uniform accuracy in diagnosis among different healthcare settings because manual evaluations are subjective and retinal images are complex. With such challenges, more efficient, quantifiable, and accurate methods for DR diagnosis are urgently needed, especially in under-resourced areas and resource-constrained situations. Medical image analysis has seen a revolution in past years because of DL, especially CNNs, which provide potential solutions for automating the diagnosis of several disorders, including diabetic retinopathy. By learning hierarchical representations of patterns in the input images, CNN's—a kind of deep learning model—perform very well when processing image data by Jordi Piera-Jiménez[2]. These models surpass traditional ML techniques in terms of precision as well as stability, displaying outstanding efficacy on tasks like image classification, segmentation, a lond with detection. DL models can instantly detect and classify different diabetic retinopathy stages based on great extent of accuracy by working with vast sets of labeled retinal images. This will reduce the workload of medical

experts and provide a more scalable screening approach. In addition, it is currently possible to leverage pre- existing deep learning frameworks for specific purposes such as DR detection owing to advancement of transfer learning, fine-tuning existing models with domain-specific data by Kaili Sun[3]. The primary objectives of this research are enhancing diabetic retinopathy diagnosis using application for DL techniques. To evaluate the accuracy, strength, and effectiveness of seven pre-trained DL models used in diabetic retinopathy diagnosis, the primary objective involves making a comparative investigation. Some of the most widely recognized architectures for computer vision are these models: ResNet50, VGG16, InceptionV3, DenseNet121, Xception, MobileNetV2, and EfficientNetB0. Applying these models to diagnosis of DR offers chance to investigate their potential for early disease identification, as each of them has demonstrated promise in a variety of image classification tasks. In addition to assisting in the identification of the best model for DR classification, the comparison analysis will illuminate benefits along with drawbacks for each model about how well it performs on retinal pictures shown by Jordi Piera-Jiménez[4]. Critical questions like: Which model achieves the highest diagnosis accuracy for diabetic retinopathy? are addressed by this objective. What is the performance of different designs on different DR data subsets? What aspects of regularization, fine-tuning, and data augmentation affect these models' resilience? This research's second goal is to create CNN model that optimized for the unique difficulties associated with diagnosing diabetic retinopathy. Although the pre-trained models are a good place to start, specific training approaches are needed to optimize model to work effectively with DR data. These techniques involve hyper-parameter tuning, implementing advanced data augmentation strategies, and utilizing regularization methods to prevent over fitting. For the purpose of classifying different stages of diabetic retinopathy, best model will attempt to achieve high level of accuracy, sensitivity, and specificity by Ravishankar Rao[5]. To further ensure model's robustness over various test sets as well as data distributions, its performance will be measured in terms of several parameters including recall, accuracy, precision, and F1-score. AUC-ROC. Stream-lit, a Python framework that simplifies the development of interactive web apps, will be combined with the optimized model after training. Stream-lit facilitates rapid prototyping and deployment for ML models and thus makes it an ideal candidate for creating a user-friendly interface that allows doctors to upload images of the retina and receive instant predictions on if a patient suffers from diabetic retinopathy, its severity, or other eye-related conditions. By this integration, physicians will possess a convenient screening tool for diabetic retinopathy, thereby enhancing the availability and effectiveness of the diagnostic process by Bing Lin[6]. Seven popular pre-trained models that have all demonstrated performance in general picture classification tasks are used in this study. This includes:

- ResNet50: A deep residual network that effectively trains very deep networks by including skip links to assist in alleviating the vanishing gradient issue.
- VGG16: A popular and straightforward architecture distinguished by its deep convolutional layers and consistent structure
- InceptionV3: A model with an inception module that enables network towards record characteristics at different levels of abstraction along with scales.
- DenseNet121: A network that improves gradient flow during training and facilitates greater feature reuse by using dense connections across layers.
- Xception: An architecture that creates a more effective model with fewer parameters by substituting depth- wise separable convolutions for the conventional convolutional layers.
- MobileNetV2: A computationally efficient, lightweight model for embedded and mobile devices.
- EfficientNetB0: A model that reaches cutting-edge accuracy with less specifications by optimizing the trade-off between model depth, width, and resolution via a compound scaling technique.

By contrast, we seek to determine which model best performs for the task of diabetic retinopathy detection and how each model approaches the complexity and variation of retinal images.

Apart from DL methodology optimization for DR

diagnosis, research efforts also seek to address accessibility of AI-based diagnostic tools. DL models have proven much potential but, in most cases, finding their way into clinical practice is dependent on technical expertise and advanced hardware. Stream lit breaks the mold by proposing a straightforward, interactive interface where healthcare providers can engage with the model without having to be experts in deep technicality. By combining the improved CNN model with a Streamline- based system, we allow medical professionals to upload retinal images and receive predictions in real time, which makes it simple to integrate AI-based diagnosis into regular clinical workflows. The Streamlined interface will show the results of the model's analysis in a simple, easy-to-use way, which enables clinicians to make informed patient care decisions by Shiyi Yin[7].

II. RELATED WORK

The inaccuracy of training progress estimation in deep learning model training—especially when online data preparation is involved—is the primary problem tackled. The overhead of preprocessing is not accounted for by existing approaches, leading to inaccurate training progress and remaining time estimations. This may mislead users and complicate early halting decisions. The research introduces a novel method of progress estimation based on online data preprocessing in its calculations. The method, applied in Tensor-Flow, specifically considers preprocessing costs while regularly recalculating training progress and remaining time. It is designed for deep learning models that incorporate integrated preprocessing and end-to-end training, ensuring accurate monitoring and consistent estimates during training. The new method significantly improves the stability of training time estimations, based on tests run on several deep learning models. It offers a more reliable and practical solution to long-duration model training tasks by minimizing the error in estimating remaining training time by 16.0% on average compared to the previous approach. Great strides have been achieved in improving the accuracy for training time estimation of DL models by author G. Luo [8]. The primary issue is that traditional methods struggle to accurately diagnose and evaluate the performance of gearboxes with similar failure conditions, like worn and broken teeth. When fault states share equal or near equal membership values, the principle of maximum membership often fails and produces ambiguous or inaccurate diagnostics. The research proposes an online diagnosis model that integrates a fuzzy comprehensive evaluation technique with a Hidden Markov Model (HMM) to enhance diagnostic accuracy. This model offers a closeness evaluation method that incorporates semi-trapezoidal and intermediate-ridge membership functions and uses likelihood ratio of the HMM as similarity measure. Improved fault state discrimination is enabled by this method's more advanced evaluation, especially when membership values are closely correlated. The proposed model with the closeness evaluation method is superior to the traditional maximum membership concept, as demonstrated by B. Xu [9]. Based on experimental findings, it significantly enhances the accuracy and reliability of gearbox performance assessment by differentiating worn from broken tooth faults, particularly when membership values are nearly equal. Standardized objective measures to measure effectiveness of deep neural network (DNN) models are absent in vision-based autonomous driving. Existing methods often employ random measurements that do not entirely or perfectly represent driving performance, such as mean error on test sets. This lack of consistency complicates comparison between different models and architectures in a substantial manner. This paper introduces OPEMI, integrated statistic designed to objectively test autonomous driving models across numerous cases of driving. OPEMI provides a clear and whole picture of driving capabilities and integrates crucial performance indicators like control stability, trajectory stability, speed, success rate, route length, and travel distance. OPEMI tackles shortcomings for current measurements by examining the poor relationship between prediction accuracy along with actual driving performance. OPEMI metric was tested on two in-house models and four well-known DNN models in generic, urban, and racing settings, as shown by D. Kim[10].

The findings show that OPEMI provides an impartial and trustworthy method of assessment under a variety of driving circumstances by accurately capturing realistic driving performance. This statistic offers a workable way to standardize the evaluation of models for autonomous driving. The redundancy and possible confusion in the methods used to calculate a multi-label confusion matrix have been recognized as the issue. The method used in the publication "MLCM: Multi-Label Confusion Matrix" to assess multi-label classification using a confusion matrix is said to be unique. Nevertheless, a similar approach was presented in a previous article by D. Krstinić et al. [11], which raised concerns regarding originality and variations in the contributions of error calculation between the two approaches.

III. METHODOLOGY

1. Convolutional Neural Network (CNN)

A systematic solution to a difficult classification problem with five classes—Mild, Moderate, No_ DR , Proliferative_ DR, and Severe is guaranteed through the diabetic retinopathy classification project with steps that involve data preparation, model development, training, and performance measurement. The dataset is separated into three subsets: training, validation, and testing. Each subset contains labelled photographs of each of the five classes. To standardize the input to the neural network, images are shrunk to 128x128 pixels. Image Data Generator is using for normalize pixel values by re-scaling them between 0 and 1. This ensures consistent data distributions, which speeds up model convergence. To account for variability, training, and validation generators mix the data during training; however, to guarantee accurate assessment metrics, shuffling is turned off for the test set. Although this particular implementation concentrates on baseline performance without augmentation, data augmentation techniques can be used if necessary increasing variability in training data. CNN constructed with TensorFlow/Keras's Sequential API is used for the classification challenge. To down sample feature maps and minimize dimensionality of space. While maintaining key characteristics, the network begins with three layers of convolution that progressively add more filters (32, 64, and 118). These are ReLU activation and max-pooling layers come next. Convolutional layer output that has been flattened is passed through a dense layer with 128 units to reduce over- fitting. That is completely linked by Charles C [12]. After that, Regularization is done using a dropout layer with a rate of 0.5. The final output layer forecasts class probabilities for each of the five categories using a soft max activation function. A batch size of 32 is used to train the CNN across 20 epochs. Batches of photos from training set are fed into model during training phase, along with its performance on validation set is tracked. To monitor development and identify problems such as over- fitting or under-fitting, accuracy and loss curves are shown. With a validation accuracy of 75% at the end of training, the model exhibits reasonable generalization and shows consistent growth during training Performance is evaluated in depth on the test set, with a 71% test accuracy rate. A confusion matrix is generated to display the true vs predicted labels for every class and determine areas of misclassification and strength. Classification report provides a detailed assessment of each class's accuracy, recall, and F1 score, giving insight into model's performance by Noyan Aendikov[13]. While still room for improvement, model's effectiveness in identifying this critical stage is shown by the Proliferative_ DR class, a severe condition, which achieves a recall of 68%. Random test, validation, and training set samples are expected to qualitatively confirm the model. The true and predicted labels for all images are displayed, where the predictions of the model can be visually confirmed. This phase gives a helpful level of testing by showing how well the model performs on given samples, even under challenging conditions. The formula for the last output layer of the CNN for classifying diabetic retinopathy can be written as:

$$\hat{y}i = \frac{\exp(zi)}{\sum_{j}^{5}\exp(zj)}, i = 1,2....,5$$

Where:

- $\hat{y}i$ is predicted probability for class i
- zi is raw output (logit) from last Dense layer for class i
- The denominator normalizes logits to ensure probabilities sum to 1 across all 5 classes.

- K Total number of classes (here, K=5).

2. RESNET50

This study uses a ResNet50-based CNN along with methodical approach improving diagnosis of diabetic retinopathy. Retinal pictures are first classified into five classifications via data processioning: mild, moderate, no_ DR, proliferate_ DR, as well as severe. Towards comply with ResNet50's input specifications, the photos are scaled to 224 × 224 pixels and normalized to fall between [0, 1] [0, 1]. To lessen over-fitting and enhance model generalization, data augmentation methods such as rotation, zooming, width/height shifts, and horizontal flips are used in the training set. To preserve features acquired from the ImageNet datasets, the model architecture makes use of a pre-trained ResNet50 with frozen convolutional layers. Worldwide Average Dense layers with 1024 neuron's with ReLU activation add non-linearity, the pooling layer shrinks spatial dimensions, and the final Dense layer with 5 neurons and softmax activation produces probabilistic predictions for every class. On top of this base model, these custom layers are created. 0.0001 learning rate and a categorical cross-entropy loss function are aspects of the model, is developed using Adam optimizer. In order to prevent over fitting and maintain the best-performing model, training is done using a batch size of 32 over 20 epochs. Model check-pointing and early halting are employed. Eric Adjei Lawer tracks and displays accuracy and loss of training and validation in order to evaluate the model's learning progress [14]. When the model is evaluated on various dataset, To evaluate its efficacy, measures including accuracy, precision, recall, and F1 score are computed. An analysis of categorization and a matrix of confusion offer comprehensive information on how well the model can categorize each category, highlighting areas for improvement. The trained model is also tested on individual images, where it preprocesses the input, makes predictions using the softmax function, and assigns the class label with the highest probability. This methodology integrates visualization of performance metrics, insights from the confusion matrix, and deployment for inference, making it a robust solution for diabetic retinopathy diagnosis.

$$p(yi \mid x) = \frac{\exp(zi)}{\sum_{j=1}^{k} \exp(zj)}$$

- $p(yi \mid x)$ Probability of the image belonging to class iii.
- zi Logit for class iii.

3. INCEPTIONV3

The InceptionV3 DL model is used in this study towards categorize retinal pictures to diagnose diabetic retinopathy. The process is organized into several phases, starting with the preparation and augmentation for data. The five types of retinal images—mild, moderate, no DR, proliferate DR, and severe— are divided into sets for testing, validation, and training. Training images are subjected to data augmentation techniques such as rotation, shifting, shearing, zooming, and horizontal flipping, towards enhance model generalization and decrease over-fitting. To comply with the input criteria of InceptionV3, images are downsized to 299 × 299 pixels, and all sets are re- scaled to a [0, 1] range. To enable bespoke categorization, the InceptionV3 model is then set up using ImageNet pre-trained weights, with the exception of the top layers. Michael Balas adds custom layers to produce probabilities for each class, such as three layers: a dense layer with 1024 neuron's activated by ReLU, a Global Average Pooling layer, and a final dense layer activated by SoftMax [15]. Important qualities are preserved by freezing the layers of underlying model. In this setup, robust InceptionV3 architecture is altered for the classification of diabetic retinopathy. We construct the model employing Adam optimizer combined with categorical cross-entropy loss
(learning rate 0.001). It is then trained for 20 epochs, with learning progress checked at the end of each epoch using validation. The accuracy and loss plots of the model are presented to measure performance and identify any indication of over- or under-fitting. A variety of classification measures, including accuracy, precision, recall, and F1 score, are used to assess the model's performance on the test set. Classification reports also confusion matrices indicate how accurately the model separates the five

classes. This methodology guarantees a reliable and effective InceptionV3-based solution to diabetic retinopathy detection. Data preparation, model configuration, training, evaluation, and metrics calculation are all encompassed.

$$p(yi \mid x) = \frac{\exp(zi)}{\sum_{j=1}^{k} \exp(zj)}$$

- $p(yi \mid x)$ Probability of the image belonging to class iii.
- zi Logit (model output before softmax) for class iii.

  K Total number of classes, which in this case is 5.

4. DENSENET 121

The classification of retinal images into five classes—Mild, Moderate, No_DR, Proliferate_DR, and Severe—diabetic retinopathy classification project method follows a DenseNet121-based deep learning structure alongside preprocessing and evaluation methods. Training, validation, and testing are three subsets for preparing data set at beginning of procedure. In order to ensure model compatibility, images within the datasets are reduced to a fixed size of 128×128128 \\times 128128×128 pixels. Pixel values scaling to range of 0−10−10−1 using data augmentation and normalization enhances model performance and generalizability. The DenseNet121 model is used as the base for feature extraction, which is already trained on the ImageNet data-set. Its architecture and pre-trained weights serve as a good starting point, minimizing extensive calculations and initial training. To facilitate personalization for the specific classification problem, the surface layer of the model is taken out by Roomasa Channa[16]. In order to reduce over fitting and to acquire top-level spatial features from the images, dense layers of the original architecture are replaced with a global average pooling layer. Additional layers are added to the model to adapt it to the DR dataset . A dropout layer with a 0.5 rate is used to prevent overfitting after a dense layer of 128 neurones and ReLU activation to find non-linear correlations in the data. A soft-max-activated dense layer of five neurons subsequently outputs the probability for the five classes. Utilising the accuracy metric, categorical cross-entropy loss function, and Adam optimiser, the model is constructed. Though minimizing difference between predicted as well as true class probabilities, these parameters ensure successful learning. In 20 epochs, model is trained by providing training images and labels in batches of 32 Precisio, recall, and F1-score are computed per class using a classification report, and predictions for the test images are generated. For better illustration, a heatmap is employed in order to visually represent the contrast between actual labels and predicted ones in a confusion matrix. Lastly, random images are selected for visualization from the validation, test, and training sets. To provide qualitative insights into model's performance, predictions are displayed alongside true labels. To be utilized and deployed in real-world scenarios, the model is saved in HDF5 format by Ayushi Chahal[17].

$$y = soft\max(W_2.RELU(W_1.GAP(DENSEnET121(X)) + b1) + b2)$$

Where:

- X: Input image
- DENSEnET121(X): Feature extraction using DenseNet121 (pre-trained model)
- GAP: Global Average Pooling layer
- W1,b1W2,b2: Weights and biases of the fully connected layers
- RELU: Activation function max(0,x)
- Softmax (z) = $\frac{ez^i}{\sum_j ez^j}$ Outputs probabilities for each class
- $\hat{y}$ Predicted class probabilities.

5. VGG 19

The data set is made up of images that belong to the five classes of diabetic retinopathy and are organized into three folders: train, validation, and test. Images are reduced to 128×128128 \\times 128128×128 pixels to meet the VGG19 model input requirements. TensorFlow's ImageData Generator is employed for image augmentation, which normalizes the pixel values by scaling them to the range of 0 to 1. The VGG19 model, pre-trained on ImageNet weights except for fully connected layers, is used as the feature extractor. To preserve learned weights, the basic model layers are frozen. A custom classification head is added, which consists of 128 neurons in a dense layer with ReLU activation, a flattened layer to turn the feature maps and

a dropout layer with a rate of 0.5 to prevent over-fitting into a 1D vector. Last layer produces probabilities for each of the five classes and is layer with a softmax activation function that is dense. Using categorical cross-entropy as the loss function, the model is trained using the Adam optimiser, and accuracy as the metric of evaluation. Training as well as validation sets are employed for training for 20 epochs. The test set consisting of unseen pictures during training and validation is used for model evaluation after training. Model performance is assessed through accuracy, loss, classification reports and confusion matrix generated by Curtis P. Langlotz [18]. For a more thorough assessment, the classification report provides measurements for accuracy, recall, and F1-score, and the confusion matrix shows each class's true positives, false positives, true negatives, and false negatives. For observing patterns also identifying issues such as over fitting, training and validation accuracies and losses are displayed. Heat-map visually represents performance of classification by the confusion matrix. F1-score, recall, average precision, and accuracy are also calculated as a part of the evaluation process. The class of new images is predicted based on the training model. An input image is preprocessed by an inference function, which then infers using the model and displays the result along with the image. This aspect ensures that the model can be utilized to diagnose diabetic retinopathy in actual practice. Below is an equation that depicts loss function utilized in model:

$$L = -\sum_{i=1} yi \log pi$$

Where:

- L is loss function (categorical cross-entropy).
- C is number of classes (in your case, 5 classes).
- yi is true label of the i-th class (binary value: 1 if instance belongs to class i, else 0).
- pi is predicted probability of instance belonging to i class.

6. VGG16

There are several important components in the training and assessment process for the VGG16-based deep learning model for diabetic retinopathy classification, from data pre-processing to model evaluation. Based on the degree of diabetic retinopathy, the data-set's images are initially divided into five categories: mild, moderate, no_DR, proliferate_DR, and severe. The photos are initially pre-processed using ImageData Generator for data augmentation in order to improve model generalization. The pictures are reduced to a consistent 128x128 pixel size after Pixel values have been rescaled to fall between 0 and 1. by dividing by 255. To evaluate the end results, the test set is left unsorted while pictures from the validation and training sets are intermingled. Data is loaded into three generators for testing, validation, and training. VGG16 network that has already been trained is used as a feature extractor for the model architecture, except for the top fully connected layers (because we are fine-tuning it for a new purpose). Towards detect diabetic retinopathy, the VGG16 network can be used because it has previously acquired valuable features from a large-scale dataset (ImageNet). Model just trains the custom layers and keeps the learned features through the pre-trained layers' weights being frozen. 3D output of VGG16 becomes a 1D vector. by adding flattened layer, completely linked layer with 128 neurons as well as ReLU activation comes next to create sequential model. During training, dropout layer is incorporated to randomly deactivate neurones in order to avoid overfitting. Model, It is put together using the categorical cross-entropy loss function and is suitable for multi-class classification applications and the Adam optimizerModel is trained over 20 epochs, and its performance is monitored by monitoring its accuracy and loss during training and validation.In order to make its prediction on training and validation data more precise, model learns to reduce the loss while adjusting its weights over the course of training by Lisa M. Hamm[19].

$$\hat{y} = Soft\max(W.x + b)$$

Where:

- $\hat{y}$ is the predicted probability distribution over classes.
- W is weight matrix of final layer.

- x is feature vector extracted by the VGG16 network.
- b is bias term for final layer.

Softmax function converts logits (raw outputs) into probabilities that sum to 1 for each cl.

7. Inception ResNet V2

Dataset comprising five classes—Mild, Moderate, No DR, Proliferative DR (Proliferate_DR), and Severe—was used to train and assess the Inception-ResNetV2 architecture model for diabetic retinopathy classification. Using The model was trained over 20 epochs with a batch size of 32 using the Adam optimiser and a categorical cross-entropy loss function. The data collection was preprocessed using Image Data Generator, normalizing the pixel values for the test, validation, and training sets between 0 and 1. There were 2000 photos in training dataset's, 400 in validation datasets, and 500 in the test set. When Tests were employed to assess the model's functionality set following training, it produced a test accuracy of 60.8% along with test loss of 0.949. There were variances in the rather modest precision**.** Similar trends were seen in the validation accuracy, which peaked in epoch 14 at 75% accuracy. To further assess the model's performance, a confusion matrix was developed. Distribution of predictions among five classes was displayed in the confusion matrix. Additional information on the model's accuracy in classifying each category was revealed by the classification report. With accuracy of 0.92, a recall of 0.87, along with F1 score of 0.89, class 'No_DR' (No Diabetic Retinopathy) performed the best, showing that the model successfully detected cases that were not retinopathy by Xiumei Ma[20]. The overall One may calculate the F1 score. using formula:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}$$

In these formulas:

- TP is number of true positives (correct predictions of a class),
- FP is number of false positives (incorrect predictions of a class.

IV. MODEL ARCHITECTURE

Several deep learning architectures, including DenseNet121, InceptionV3, ResNet50, ResNet101, VGG16, VGG19, and Inception-ResNetV3, are utilized during training to detect features indicative of diabetic retinopathy. Model performance is assessed through accuracy, loss graphs, a confusion matrix, and classification results. Once training is complete, the model is saved as **model.h5** and tested on unseen data to generate predictions. The final step involves evaluating key performance metrics such as precision, recall, and F1-score to assess the model's effectiveness in diagnosing diabetic retinopathy. This structured and modular approach ensures the development of a reliable deep-learning model for medical diagnosis.

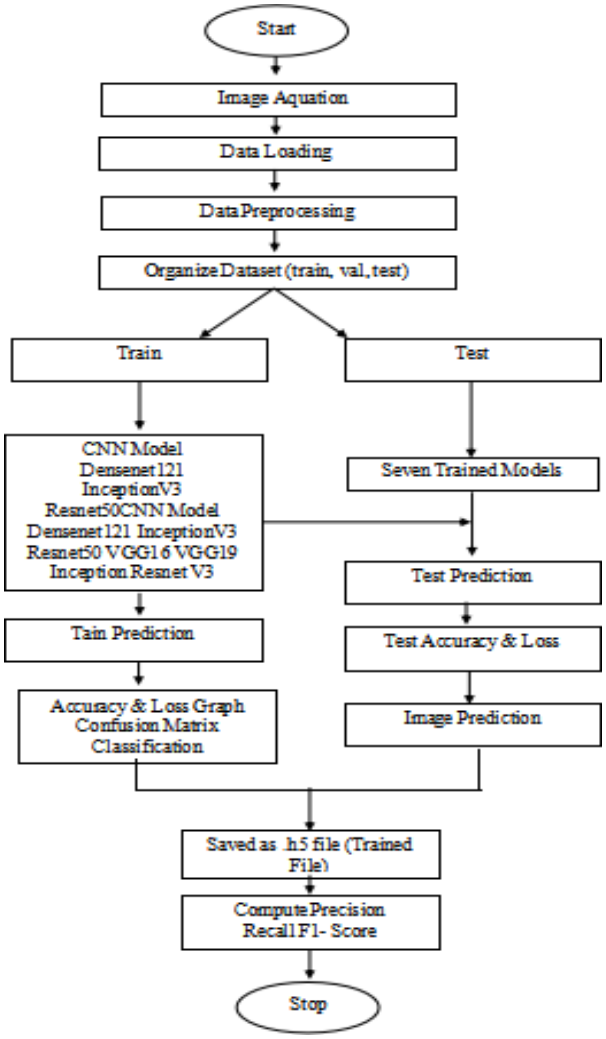

Figure1. Diabetic Retinopathy Detection Model Training and Evaluation Workflow.

Result:

CNN model is clearly most effective model for diagnosing diabetic retinopathy, according to the table that is supplied. Along with a steady F1 Score of 0.71, Recall of 0.71, and Precision of 0.71, it attains the greatest accuracy of 71.00%. According to these criteria, the CNN model finds a good balance between identifying positive and negative cases without giving any preference to one over the other. In the field of medicine, this is essential since both false positives (inaccurate diagnoses) and false negatives (missing diagnoses) can have detrimental effects. Both in detecting diabetic retinopathy and minimizing false positives, the CNN model is performing well, as per F1 Score, important statistic that balances precision with recall. How accurately the model classifies retinal images into the right classes is evidenced by its F1 Score of 0.71, which offers reliable predictions to assist doctors in making more accurate diagnoses. To ensure all potential cases are flagged for further testing, the high Recall value of the model (0.71) would mean that the model is particularly good at identifying true cases of diabetic retinopathy. Compare this to the other models (DenseNet121 and Inception ResNetV2) with relatively lesser values in F1, Recall, and Precision, and their lower accuracy measures (62.60% and 60.80%, respectively). These models are less accurate and comprehensive than the CNN model, yet still respectable in performance. The VGG16 and VGG19 models, with the lowest accuracy of 47.80%, are relatively poor, which suggests that their architecture is not best suited to the complexity of diabetic retinopathy.

Table below summarizes the comparison of model performance:

Model Name	Accuracy	F1 Score	Recall	Precision
CNN	71.00%	0.71	0.71	0.71
DENSENET121	62.60%	0.61	0.63	0.62
VGG19	54.20%	0.53	0.53	0.55
VGG16	47.80%	0.44	0.48	0.53
Inception ResnetV2	60.80%	0.61	0.61	0.61
ResNet V2	0.38	0.30	0.38	0.35
Inception V3	20.00%	0.0667	0.2000	0.0400

Table 1. Analysis of Comparative Table

CONCLUSION

Notable insights into efficacy for a number of DL models for diabetic retinopathy classification are provided through confusion matrix comparison and performance metrics. EfficientNet B0 holds up as the most accurate model out of those evaluated, achieving the highest accuracy and performing outstandingly in discriminating between the five levels of illness severity. This is a result of its equally well-rounded architecture, which effectively maintains computational efficiency while managing to capture detailed patterns in retinal images successfully. Models such as VGG16, VGG19, ResNet152, and InceptionV3, however, demonstrate varying levels of complexity, particularly in regards to identifying correct intermediate classes such as "Moderate" or "Mild," reflecting inefficiencies in catching minute illness progression. The impact of depth in design and architecture on performance is better exemplified by DenseNet121 and CNN, which perform well but are poor at dealing with intricate classifications. The results show the importance of employing advanced architectures, stable data augmentation techniques, and hyperparameter tuning in a bid to enhance classification accuracy. EfficientNet B0 is better overall, but more testing and refinement need to be done for all models to enhance their ability to properly classify diabetic retinopathy at all stages.

REFERENCES

[1] Abdur Rahman Onik, Trevor T. Spinosa, Abdulla M. Asad, Ibrahim Baggili,Hit and run: Forensic vehicle event reconstruction through driver-based cloud data from Progressive's snapshot application,Forensic Science International: Digital Investigation,Volume 49, Supplement,2024,301762,ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2024.301762.

[2] Jordi Piera-Jiménez, Toni Dedeu, Claudia Pagliari, Tatjana Trupec,Strengthening primary health care in Europe with digital solutions,Atención Primaria,Volume 56, Issue 10,2024,102904,ISSN 0212-6567, https://doi.org/10.1016/j.aprim.2024.102904.

[3] Kaili Sun, Xucheng Wu, Jiening Yu, Weiran Chen, Xueqing Jia, Dongwen Li, Hao Lyu, Wenhai Huang, Hongqing Tian, Weijing Gao, Bin Zhu,

Xinwei Lyu, Li Wang, Jing Shao, Xiaoting Liu, Zuyun Liu,Development and validation of a screening tool for sarcopenia in Chinese nursing home residents,iScience,Volume 27, Issue 11,2024,111221,ISSN 2589-0042, https://doi.org/10.1016/j.isci.2024.111221.

[4] Jordi Piera-Jiménez, Toni Dedeu, Claudia Pagliari, Tatjana Trupec,Strengthening primary health care in Europe with digital solutions,Atención Primaria,Volume 56, Issue 10,2024,102904,ISSN 0212-6567, https://doi.org/10.1016/j.aprim.2024.102904.

[5] Ravishankar Rao, Raunak Jain, Mrityunjai Singh, Rahul Garg, Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data, Healthcare Analytics,Volume 6,2024,100351,ISSN 2772-4425 https://doi.org/10.1016/j.health.2024.100351.

[6] Bing Lin, Shihan Feng, Jiaxiu Liu, Kangjie Li, Guiqian Shi, Xiaoni Zhong,Using an interactive web application to identify pre-exposure prophylaxis adherence among men who have sex with men,International Journal of Clinical and Health Psychology,Volume 24, Issue 3,2024,100490,ISSN 1697-2600, https://doi.org/10.1016/j.ijchp.2024.100490.

[7] Shiyi Yin, Jinyuan Wang, Jingyuan Zhu, Ximeng Feng, Haihan Zhang, Haowen Li, Jingying Xiu, Chuanqing Zhou, Qiushi Ren, Wenbin Wei,A multimodal imaging approach to investigate retinal oxygen and vascular dynamics, and neural dysfunction in bietti crystalline dystrophy,Microvascular Research,2024,104762,ISSN00262862, https://doi.org/10.1016/j.mvr.2024.104762. (https://www.sciencedirect.com/science/article/pii/S0026286224001110)

[8] Q. Dong and G. Luo, "Progress Estimation for End-to-End Training of Deep Learning Models With Online Data Preprocessing," in *IEEE Access*, vol. 12, pp. 18658-18684, 2024, doi: 10.1109/ACCESS.2024.3359996.

[9] Y. -K. Gu, B. Xu, H. Huang and G. Qiu, "A Fuzzy Performance Evaluation Model for a Gearbox System Using Hidden Markov Model," in *IEEE Access*, vol. 8, pp. 30400-30409, 2020, doi: 10.1109/ACCESS.2020.2972810.

[10] D. Kim, A. Khalil, H. Nam and J. Kwon, "OPEMI: Online Performance Evaluation Metrics Index for Deep Learning-Based Autonomous Vehicles," in *IEEE Access*, vol. 11, pp. 16951-16963, 2023, doi: 10.1109/ACCESS.2023.3246104.

[11] D. Krstinić, L. Šerić and I. Slapničar, "Comments on “MLCM: Multi-Label Confusion Matrix”," in *IEEE Access*, vol. 11, pp. 40692-40697, 2023, doi: 10.1109/ACCESS.2023.3267672.

[12] Charles C. Wykoff, Diana V. Do, W. Lloyd Clark, David S. Boyer, Dilsher S. Dhoot, Dennis M. Marcus, Robert Vitti, Alyson J. Berliner, Kimberly Reed, Yenchieh Cheng, Hadi Moini, David M. Brown, Longitudinal Changes in Diabetic Retinopathy Severity: Learnings from PANORAMA (78/80 characters including spaces),Ophthalmology Retina,2024,,ISSN 2468-6530, https://doi.org/10.1016/j.oret.2024.10.007.

[13] Noyan Aendikov, Aeila Azayeva,Integration of GIS and machine learning analytics into Streamlit application,Procedia Computer Science,Volume 231,2024,Pages 691-696,ISSN 1877-0509 https://doi.org/10.1016/j.procs.2023.12.160.

[14] Eric Adjei Lawer,An evaluation of single and multi-date Landsat image classifications using random forest algorithm in a semi-arid savanna of Ghana, West Africa, Scientific African, Volume 26,2024,e02434,ISSN 2468-2276, https://doi.org/10.1016/j.sciaf.2024.e02434.

[15] Michael Balas, Jonathan A. Micieli, Jovi C.Y. Wong,Integrating AI with tele-ophthalmology in Canada: a review,Canadian Journal of Ophthalmology,2024,ISSN 0008-4182, https://doi.org/10.1016/j.jcjo.2024.08.013.

[16] Roomasa Channa, Risa M. Wolf, Rafael Simo, Mitchell Brigell, Patrice Fort, Christine Curcio, Stephanie Lynch, Frank Verbraak, Michael D. Abramoff, Michael D. Abramoff, Roomasa Channa, Risa M. Wolf, Rafael Simo, Mitch Brigell, Patrice Fort, Christine Curcio, Stephanie Lynch, Frank Verbraak, Thomas W. Gardner,A New Approach to Staging Diabetic Eye Disease: Staging of Diabetic Retinal Neurodegeneration and Diabetic Macular Edema,Ophthalmology Science,Volume 4, Issue 3,2024,100420,ISSN 2666-9145, https://doi.org/10.1016/j.xops.2023.100420.

[17] Ayushi Chahal, Preeti Gulia, Nasib Singh Gill, Mohammad Yahya, Mohd Anul Haq, Mohammed

Aleisa, Abdullah Alenizi, Arfat Ahmad Khan, Piyush Kumar Shukla,Predictive Analytics Technique based on hybrid Sampling to manage Unbalanced Data in Smart Cities,Heliyon,2024,e39275,ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e39275 .

[18] Curtis P. Langlotz, Johanna Kim, Nigam Shah, Matthew P. Lungren, David B. Larson, Somalee Datta, Fei Fei Li, Ruth O'Hara, Thomas J. Montine, Robert A. Harrington, Garry E. Gold, Developing a Research Center for Artificial Intelligence in Medicine, Mayo Clinic Proceedings: Digital Health,2024,ISSN 29497612, https://doi.org/10.1016/j.mcpdig.2024.07.005

[19] Lisa M. Hamm, Iris Wainiqolo, Nayana Pant, Subash Bhatta, Danielle Petrie-Deely, Pushkar Silwal, Benjamin Zuvani, Ana Patricia Marques, Nimisha Chabba, Laite Tuiloma, Valeria Lopez, Osea Masilaca, Jacqueline Ramke, Research about eye health and eye health services in Pacific Island Countries and Territories: a scoping review, The Lancet Regional Health - Western Pacific, Volume 50,2024,101152,ISSN 2666-6065, https://doi.org/10.1016/j.lanwpc.2024.101152

[20] Xiumei Ma, Jinlin Wang, Kefa Zhou, Wenqiang Zhang, Zhixin Zhang, Shuguang Zhou, Yong Bai, Philippe De Maeyer, Tim Van de Voorde, Quantitative evaluation of the impact of band optimization methods on the accuracy of the hyperspectral metal element inversion models, International Journal of Applied Earth Observation and Geo information, Volume 132,2024,104011,ISSN 1569-8432, https://doi.org/10.1016/j.jag.2024.104011.