Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

A lightweight transfer learning based ensemble approach for diabetic retinopathy detection

S JAHANGEER SIDIQ *, T BENIL

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

ARTICLE INFO

Keywords:
Classification
Deep learning
Diabetic retinopathy (DR)
Ensemble
Transfer learning

ABSTRACT

Diabetic retinopathy (DR) is a fatal and irreversible eye disease that affects millions of people worldwide. It occurs due to high blood sugar level in the body of a diabetic patient, so it requires immediate attention which goes beyond the clinical solutions. With the advancements in deep learning and computer vision there are maximum possibilities of predicting this disease at early stages. Based on the severity of disease, different labels have been assigned to different classes of this disease as follows: 4 for proliferative DR, 3 for severe DR, 2 for moderate DR, 1 for mild DR and 0 for No DR. In this paper we proposed a deep learning-based ensemble approach using pre-trained and customized bi-class (CNN) base-learners like MobileNet, InceptionV3and DenseNet121 which were identified during initial investigation. These deep learning models were used as the base learner because of their promising performance in ensembles compared to the other deep learning base learners. All the work in the literature has studied this as a single complex multi-class problem or a bi-class problem where earlier stages are grouped together (0 to 3) and treated as one class and 4 as separate another class. Our work breaks this multi-class problem into multiple simpler two class problems using OVO(One-Versus-One) approach. Several benchmark data sets such as APTOS 2019, IDRiD, Messidor-2 and DDR which are multi-class data sets were used for training and testing our models. Data augmentation techniques were also utilized. Performance metrics such as precision, recall, f1-score, and accuracy were used for evaluation. Our ensemble models showed a remarkable performance with precision, recall, f1-score, and accuracy for most of the datasets used in this study. In addition to this our ensemble models have minimum number of trainable parameters which makes them an ultimate choice.

1. Introduction

Diabetic retinopathy (DR) is a condition that can cause vision loss and blindness in diabetic patients. It may affect patient's blood vessels in the retina. Diabetic retinopathy may not have any symptoms at earlier stages so it's very important for diabetic patients to get their eyes examined at least once a year. Detecting it early can help you take the necessary steps to save your vision. DR is classified in to two major categories: Proliferative Diabetic Retinopathy (PDR) and Non-Proliferative Diabetic Retinopathy (NPDR) (Memon et al, 2017). The DR in final stage is called Proliferative Diabetic Retinopathy (PDR) while as DR in earlier stages is referred to as Non-Proliferative Diabetic Retinopathy (NPDR). And this (NDPR) is further classified into Severe, Moderate and Mild stages. It's difficult to differentiate between Normal and Mild as they resemble each other closely. The sample of each stage is shown below in Fig. 1.

According to a report by WHO diabetes is increasing at a faster rate and it is expected to reach 700 million by 2045.Therefore keeping these facts and figures in mind the diagnosis and early screening of DR are of vital importance. Health professionals diagnose DR using fundus images which is a time-consuming process and requires expertise. Due to the shortage of experienced health professionals and the increasing number of DR patient's misdiagnosis is a common issue. Computerized diagnosis can greatly overcome this problem by reducing the workload of health professionals and thereby increasing the accuracy of diagnosis. In recent years deep learning has contributed a lot in the field of computerized diagnostics. Convolutional Neural Networks (CNN) is the most widely used and effective in the field of computer vision because of its performance in image classification. Multiple CNN algorithms like ResNet (He et al, 2016), GoogleNet (Szegedy et al, 2015), EfficientNet (Tan et al, 2019) and VGGNet (Simoyan et al, 2014) are of great significance. The revolutionized progress in the medical fields such as cancer

* Corresponding author.
E-mail address: jahangeersidiq.s@vit.ac.in (S.J. SIDIQ).

https://doi.org/10.1016/j.jjimei.2025.100372

Available online 26 August 2025

classification (Hashimoto et al, 2020), glaucoma screening (Raghavendra et al, 2018) and retinal vascular segmentation (Wang et al, 2015) is all because of the development of CNN.

Convolutional Neural Networks (CNNs) are a type of supervised machine learning algorithm used in deep learning that require large datasets for effective training. One of their most significant advantages is the ability to automatically learn and extract features from input data. With recent advancements, CNNs have been widely employed in medical image analysis, resulting in considerable research efforts. CNNs have proven to be highly effective in various tasks, including image classification, detection, segmentation, and registration. Their use in the medical field has grown due to their high accuracy and efficiency. CNNs are typically executed on powerful workstations with graphical processing units (GPUs) because the convolution operations they perform are computationally demanding, requiring significant processing power and memory. Advances in GPU technology have led to improvements in CNN performance, further boosting their adoption in healthcare. However, many CNN architectures are complex and involve many parameters, which can make the models slow to process and unsuitable for all platforms, despite GPU support. To address this, research has focused on lightweight CNN architectures, such as MobileNet, MobileNetV2 and others which are more suitable for low-resource devices and offer efficient performance in tasks like Diabetic Retinopathy (DR) multiclass classification. The aim of this study is to explore the use of ensembles created using these lightweight CNNs for DR classification. These CNN architectures are well suited for medical dataset classification but there are insufficient medical data sets that make the model training challenging. The solution to this problem is a technique called Transfer learning (Pan and Yang, 2009) which ensures that the network is well trained as it transfers information from the huge dataset which was previously used for training. It is the best technique for making the network stable and efficient based on insufficient data. The motivation of this work is that most of the studies present in literature achieved a good classification accuracy for binary classification tasks only, that is to detect DR or No-DR. The drawback of these studies is that they are of limited use since they are not able to make an in-depth classification of specific stage of disease. It is therefore mandatory to study the grade classification of DR.

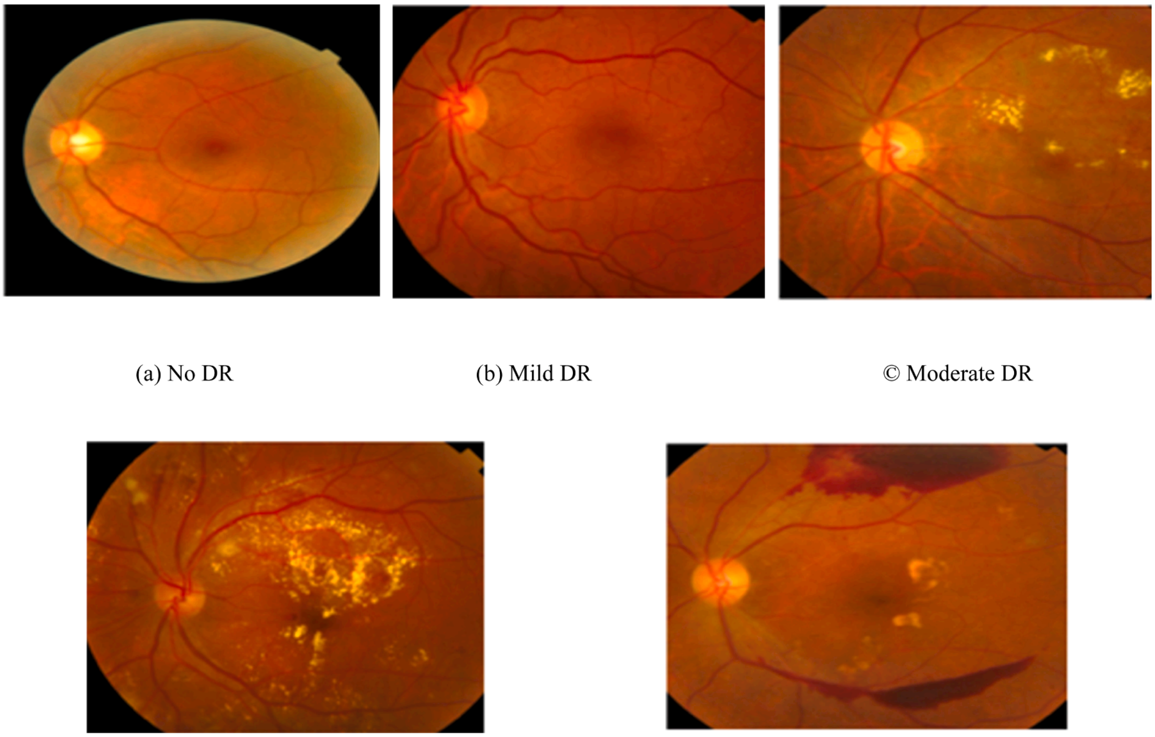
The major contribution of this study is:

1) Investigating and identifying the best base learner i.e., fine-tuned, and modified binary (CNN) classifier for all the different data sets used in this study.
2) Deep learning and transfer learning-based lightweight ensembles were proposed using the best identified and customized bi-class (CNN) as a base learner.
3) The original multi-class problem was broken down into simpler subproblems and each subproblem was solved using modified and fine-tuned binary (CNN) because of its improved performance as compared to other base learners.
4) The proposed ensembles outperformed almost all the existing models on data set by detecting each of the stages of the disease with greater accuracy.

The rest of the paper is organized as follows: Section 2 contains the latest literature review to detect DR. Section 3 contains the proposed methodology. Section 4 contains materials and methods used. Section 5 containing Results and Discussion and the last section contains a conclusion.

2. Related Work

A lot of work has been done in the literature, but we are presenting some notable and the more relevant and recent. (Jian et al., 2023) proposed a triple-cascade network for diabetic retinopathy detection. It subdivides the task of classification of DR and greatly improves the



(a) No DR

(b) Mild DR

© Moderate DR

(d) Severe DR

(e)Proliferative DR

Fig. 1. Diabetic retinopathy stages (Kassani et al, 2019).

grading performance. This Triple-DRNet on dataset achieves the accuracy of 92.08 %, which is the proof that this devised model is good as compared to the other models. Islam et al. (2022) introduced supervised contrastive learning (SCL). Two stage training method using contrastive loss function was implemented for DR detection and its severity stage detection using APTOS 2019 dataset. The accuracy and AUC of 84.364 %, 93.819 % was obtained on this five-class problem. Sugeno et al. (2021) used transfer learning and convolutional neural network namely EfficientNet-B3 and a publicly available Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 dataset. After pre-processing the image data set the model achieved the sensitivity and specificity of more than 0.98 for DR detection. The classification accuracy achieved for severity grading for first three labels was 0.84,0.95 and 0.98.Applied data pre-processing on color fundus images, extracted features using PCA and used pretrained ResNet50, ResNet152 and SqueezeNet1 there by achieving the accuracy of 93.67 %, 91.94 % and 94.40 %. Qummar et al. (2019) proposed a novel ensemble by using five Convolution Neural Network (CNN) models on publicly available retinopathy dataset and improved the classification accuracy of different stages of DR and provided the extension of the same work by reducing the trainable parameters of the ensemble thereby making it more optimal. Alyoubi et al. (2021) proposed a system which classifies DR images into five stages namely proliferative DR, severe, moderate, mild and No-DR. In PDR they also localized the affected lesions on the surface of retina. This system consists of two deep learning-based models CNN512 and the other model adopted YOLOv3 to detect as well as localize the lesions. Finally, the two models were fused together which obtained an accuracy and sensitivity of 89 %. Bodapati et al. (2020) extracted features from multiple pre-trained ConvNet models for deriving the optimal representation of retinal images to improve the DR detection. Since each Convnet extracted different features and fusing them using Cross polling and 1D polling led to far better representation than using single ConvNet. This proposed model achieved an accuracy of 97.41 % for DR identification and an accuracy of 81.7 % for severity level prediction. Kassani et al. (2019) presented a new feature extraction method by using a customized Xception architecture. The proposed method is based on deep layer aggregation which combines features from different layers of the Xception model. These extracted features are subsequently fed into a multi-layer perceptron that is used for DR severity classification. The four feature extractors used were ResNet50, MobileNet, InceptionV3 and Xception. Additionally hyper-parameter tuning, and transfer learning strategy was used to improve overall accuracy. The final model obtained an accuracy of 83.09 %, sensitivity of 88.24 % and specificity of 87.00 %. Liu et al. (2020) proposed an improved loss function and the three hybrid models namely Hybrid-a, Hybrid-c and Hybrid-f to improve the classification accuracy of DR. Xception, NASNetLarge, InceptionResNetV2, EfficientNetB4 and EfficientNetB5 were used as the base models. The cross-entropy loss and enhanced cross entropy loss were utilized for training the basic models. The output from these basic models were used for training the hybrid model. The proposed model obtained an accuracy of 86.34 %, sensitivity of 98.77 %, specificity of 74.76 %, precision of 91.37 % and f1-score of 93.9 %. Dondeti et al. (2020) collected, preprocessed, and extracted the features and fed them to a deep CNN namely Neural Architecture Search Network (NASNet) which searches for the best convolutional layer in NASNet. The severity level of disease was obtained by giving the representations of the retinal images in deep space to the classification model. The study revealed that ν-Support Vector Machine (ν-SVM) improved the accuracy of the model in comparison to the other machine learning models. It was observed that t-distributed stochastic neighbor embedding (t-SNE) deep feature transformation scheme yields more discriminative representations thereby achieving an accuracy of 77.90 %. Gargeya and Leng (2017) used a decision tree classifier and ResNet to classify fundus images and obtained AUC as 0.94 for the Messidor dataset. Rao et al. (2020) used a pre-trained ResNet50 for an APTOS 2019 dataset and obtained an accuracy of 96.59 % for the binary classification, that is to detect DR or No-DR. The above two classification models performed very well; however, they are binary classification models, meaning they are unable to make an in-depth classification of DR disease by stage. The grade classification of DR is very important. Shanthi et al. (2019) designed a neural network using the Messidor dataset to classify DR into stage 1, stage 2 and stage 3 with an accuracy of 96 % for each stage. In recent years five classifications of DR have grabbed a lot of attention because they are able to reflect the severity level of DR. Dondeti et al. (2020), combined the pre-training model (NASNET) with T-SNE space for extracting deep features thereby achieving an accuracy rate of 77.90 %. Bodapati et al. (2021) used Xception,VGG16 with gated mechanism to compose a deep neural network on APTOS 2019 dataset and achieved the accuracy of 82.54 %. Majumder and Kehtarnavaz (2021) combined the regression model with Xception model to classify five different stages of DR and achieved an accuracy of 86 %, 82 % on APTOS 2019 and EyePACS datasets respectively. Patel and Chaware (2020) obtained an accuracy of 91 % by using pre-trained MobileNetV2 on APTOS 2019 dataset. The authors Usman et al. (2024) propose Ens5B-UNet, an ensemble model combining five modified U-Net networks for improved microaneurysm (MA) segmentation in Diabetic Retinopathy (DR). Trained on the IDRiD dataset and tested on EOphta_MA, it outperforms existing models, achieving a 3.17 % higher AUPR than the top model in the IEEE ISBI 2018 competition. Ens5B-UNet shows significant improvements in segmentation metrics, including a 19.67 % boost in IOU. These results demonstrate its potential for accurate and efficient clinical use in eye care. The paper presents GAPS-U-NET, a modified U-Net for accurate Optic Disc segmentation, vital for detecting Glaucoma and Diabetic Retinopathy (DR). It improves segmentation by using a pretrained CNN backbone, gating attention for focus on key features, and pixel shuffling for better resolution. Trained on the IDRID dataset, it achieved 92.83 % IOU and 99.71 % accuracy, making it suitable for Glaucoma and DR screening, and adaptable for other medical image tasks. The review paper analyzes Machine Learning models for predicting Diabetic Retinopathy (DR) progression, highlighting a lack of longitudinal studies. It examines 13 relevant studies, identifies research gaps, and suggests strategies to improve AI-based prediction models. The paper discusses key challenges and recommends future directions for enhancing DR progression prediction.

The accuracy of DR classification has improved a lot however the DR severity classification still deserves improvement. Most of the work mentioned above uses APTOS 2019 dataset for DR classification and performs binary classification on the dataset by classifying a case as DR or No-DR. The problem with this approach is that it does not focus on predicting the stage or severeness if a patient has DR. This problem can be solved by considering DR classification as multi-class problem rather than a bi-class problem. In multi-class classification we classify DR into five different classes as discussed in the introduction of this paper. It is very important to predict the appropriate stage of this disease so that a patient can be saved from entering another severe stage. To the best of our knowledge no paper has ever classified different stages of this disease using ensemble of customized binary CNN,s. These customised binary base classifiers solve the sub parts of this complex multi-class problem and these sub solutions are combined using voting strategy to generate the solution for this multiclass problem.

3. Proposed method

This section contains the data sources, data set description, preprocessing, architecture of one of the base learners, experimental framework, and the proposed algorithm in detail.

3.1. Preprocessing

The preprocessing phase includes dividing the datasets. The data sets contain the images belonging to five different classes based on the severity of disease. The APTOS 2019 dataset contains a total of 5590 fundus images each of size (224*224). The source and training dataset distribution for all the datasets used in this study is shown in Table (1) to Table (5) below in detail.

Table 1, 2, 3, 4, 5

All the above datasets containing five classes is divided into ten binary class datasets using OVO binarization approach and then augmentation, rotation, normalization, horizontal and vertical flipping is applied on each of the ten subsets. And then these subsets are fed to the best identified base learners containing a single node in the output layer for generating the base learners (voters) for our ensemble model as depicted in Figure 3 below.

3.2. *Proposed ensemble model*

The ensemble model combines several machine learning classifiers together to improve the classification accuracy. Different types of ensembles can be created depending on whether the individual classifiers present in the ensemble are of the same type or different. If the individual classifiers are not of the same type, it is called stacking. In this work we have used the individual classifiers of the same type, but we have fed them with different bi-class subsets of data generated in the pre-processing phase and we got very promising results.

Algorithm 1 explains the proposed ensemble approach in detail. The proposed approach combines the ten deep learning-based base-learners which are frozen, and a single node output layer is appended at their end prior to their training to use the pretrained weights and make them fit for binary classification. This is because each of the individual datasets obtained from pre-processing contains only two classes in it. The individual base learners of the ensemble are fine-tuned by setting batch size to 32, epochs to 50, threshold to 0.60, learning rate to Adam 1e-3, activation function to sigmoid and polling to average. The input to the ensemble will be image dataset (X, Y); where X is the set of N images, with the size of each image as (224*224) and $Y = \{y|y\epsilon\{PDR, Severe, Moderate, Mild, Normal\}\}$.While as the base learner value for X will be subset of images from X containing only two distinct classes in them. While testing we use a majority voting scheme to combine the predictions of different base Learners and generate the result. This binary classifier-based ensemble combines the strengths of individual models which ultimately leads to better performance. The proposed binary classifier-based ensemble is shown in Fig. 3.

4. Materials and methods

In this section OVO binarization and the performance metrices used for evaluating the proposed ensemble are discussed. The experiments were performed using Kaggle P100 GPU and software libraries like Kera's, TensorFlow and Ski-learn.

4.1. *Performance metric*

The performance metrics used for evaluation were accuracy, precision, recall F1-score and AUC.

$$Accuracy = \frac{TruePositive + TrueNegative}{(TruePositive + TrueNegative + FalsePositive + FalseNegative)} \quad (1)$$

Table 1
Data sources.

Data Set	Source
APTOS 2019	https://www.kaggle.com/competitions/aptos2019-blindness-detection/data
IDRiD	https://www.kaggle.com/datasets/mariaherrerot/idrid-dataset
Messidor-2	https://www.kaggle.com/datasets/mariaherrerot/messidor2preprocess
DDR	https://www.kaggle.com/datasets/mariaherrerot/ddrdataset

Table 2
APTOS dataset description.

APTOS Class	Class Representation	Class Distribution
NO DR	0	1805
MILD DR	1	370
MODERATE DR	2	999
SEVERE DR	3	193
PROLIFERATIVE DR	4	295

Table 3
IDRiD dataset description.

IDRiD Class	Class Representation	Class Distribution
NO DR	0	129
MILD DR	1	22
MODERATE DR	2	156
SEVERE DR	3	84
PROLIFERATIVE DR	4	64

Table 4
Messidor-2 dataset description.

Messidor-2 Class	Class Representation	Class Distribution
NO DR	0	1017
MILD DR	1	270
MODERATE DR	2	347
SEVERE DR	3	75
PROLIFERATIVE DR	4	35

Table 5
DDR dataset description.

DDR Class	Class Representation	Class Distribution
NO DR	0	6266
MILD DR	1	630
MODERATE DR	2	4477
SEVERE DR	3	236
PROLIFERATIVE DR	4	913

$$precision = \frac{TruePositive}{(TruePositive + FalsePositive)} \quad (2)$$

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)} \quad (3)$$

$$F1 - Score = 2 \times \frac{precision \times recall}{(precision + recall)} \quad (4)$$

AUC = It represents the degree of separability of different classes. A higher score for AUC is an indication that the model is better.

4.2. *One-versus-one approach*

The OVO binarization approach (Hastie and Tibshirani, 1998) trains a binary classifier for each pair of classes, by ignoring the samples that do not belong to these two classes. During the testing phase, a query is given to all binary models created during the training phase, and the resulting predictions of these binary models are combined into the result (Hüllermeier and Brinker, 2008; Hüllermeier and Vanderlooy, 2010). This technique is shown in Fig.. 3. The procedure for generating the class label is to represent the predicted output of each binary model in the form of matrix M (Allwein et al, 2000):

$$M(i,j) = \{1 if output = i, 0\} \quad (5)$$

And when

Algorithm 1
Proposed ensemble algorithm.

```
Input: Fundus Images $(X, Y); whereY = \{y|y\epsilon\{PDR, Severe, Moderate, Mild, Normal\}\}$
Output: Trained ensemble model that classifies the image $x\epsilon X$.
Begin
Step 1: Divide the input multi-class dataset using OVO approach into $\frac{n(n-1)}{2}$ bi-class datasets.
For each of the bi-class subset generated in Step 1
  Step 2: Pre-process the training dataset.
  Perform augmentation.
  Horizontal, Vertical flipping and rotation.
  Normalize each image having size $(224 \times 224)$.
    Step 3: Import a pretrained model and append a single node at the end of the output layer originally containing multiple output nodes.
Step 4: Compile/train the model from bi-class data pre-processed in Step 2 using the model obtained in Step 3.
  Step 5: Save the binary class model.
  Step 6: If all the bi-class data subsets exhausted.
  Then generate results for the test data set using all bi-class models saved in step 5 using equation $= \text{arg}max_{i=1\dots C}\left\{\sum_{j=1}^{C} Mij\right\}$
End.
```

And when $M(i,j) = 1$ then $M(j,i) = 0$ and vice versa. The resulting class is assigned by maximum voting using the following equation:

$$Class = \text{arg}\underset{i=1\dots C}{max}\left\{\sum_{j=1}^{C} Mij\right\} \tag{6}$$

5. Results and discussions

Table 6 contains the hyper-parameter settings in greater detail and Fig. 2 explains the architecture of the customizes bi-class MobileNet base learner for APTOS 2019 dataset.

The performance of various customized bi-class models on the APTOS 2019 Diabetic Retinopathy dataset is shown below in Table 7. MobileNet achieved the highest Accuracy, F1-score, Precision and Recall of 92.00 % making it the top performer for this dataset. Xception ranked second with an Accuracy, F1-score, Precision and Recall of 90.63 %. DenseNet121 follows closely, with Accuracy, F1-score, Precision and Recall of 90.18 %. InceptionV3 and InceptionResNetV2 show almost similar performance, with F1-scores of 89.60 % and 88.78 %, respectively, and moderate Precision and Recall.MobileNetV2 and VGG19 performed the least effectively, with F1-scores of 87.59 % and 86.35 %, respectively, indicating lower precision and recall compared to the top models. Overall, MobileNet and Xception are the top classifiers, with MobileNet being particularly well-suited for efficient deployment in resource-constrained environments.

The performance of various customized bi-class models on the IDRiD Dataset is shown below in Table 8. InceptionV3 emerges as the best-performing model on the IDRiD dataset, with an Accuracy, F1-score, Precision and Recall of 85.19 %, demonstrating a strong performance. Xception and MobileNet follows closely behind, achieving an Accuracy, F1-score, Precision and Recall of 84.57 %. InceptionResNetV2 performs similarly, with an Accuracy, F1-score, Precision and Recall of 84.26 %. DenseNet121 ranks fourth with an F1-score of 79.63 % and Accuracy of 79.63 %, showing slightly lower performance compared to the top three models. MobileNetV2, DenseNet121 and VGG19 show relatively lower performance.

Table 6
Hyper parameters.

Parameter	Value
Image Size	224*224*3
Batch Size	32
Training Split	.80
Validation Split	.20
Polling	Average
Learning rate	Adam 1e-3
Weights	ImageNet
Epochs	50
Threshold	.60
Rescale	1./255
Shear_range	0.2
Zoom_range	0.2
Rotation_range	45
Horizontal_flip	True
Vertical_flip	True

The performance of various customized bi-class models on the Messidor-2 Dataset is shown below in Table 9. DenseNet121 is the top-performing model with an Accuracy, F1-score, Precision and Recall of 85.42 making it the best for diabetic retinopathy detection on the Messidor-2 dataset. Xception follows closely, achieving an Accuracy, F1-score, Precision and Recall of 84.05 %, showcasing strong performance. MobileNetV2 delivers solid results with an Accuracy, F1-score, Precision and Recall of 82.97 %, making it a good choice for efficient deployment. InceptionResNetV2 and InceptionV3 both perform moderately well with F1-scores around 81 % but fall behind DenseNet121 and Xception in terms of overall accuracy and balance. VGG19 shows lower performance with an F1-score of 79.96 % and Accuracy of 79.96 %, while MobileNet ranks the lowest with an F1-score of 72.92 % and Accuracy of 73.00 %. DenseNet121 and Xception are the top classifiers, with DenseNet121 offering the best overall performance for the Messidor-2 dataset.

The performance of various customized bi-class models on the DDR Dataset is shown below in Table 10. MobileNet leads the performance on the DDR dataset with the highest Accuracy, F1-score, Precision and Recall of 87.89 %. Xception follows closely with an Accuracy, F1-score, Precision and Recall of 87.80 %. InceptionResNetV2 performs well with an Accuracy, F1-score, Precision and Recall of 87.50 %. DenseNet121 achieves an Accuracy, F1-score, Precision and Recall of 86.60 %, demonstrating solid performance, but not quite matching the top three models. InceptionV3 and VGG19 show lower results, with F1-scores of 84.65 % respectively, but still perform reasonably well on the DDR dataset. MobileNet and Xception outperform other models on the DDR dataset, with both achieving nearly identical performance. InceptionResNetV2 also performs well, while other models like InceptionV3 and VGG19 lag slightly behind.

The performance of the individual binary base models of the ensemble created using customised MobileNet on APTOS 2019 dataset is shown in Table 11. Table 11 depicts that in terms of precision, recall and f1-score the best performing binary model is (No DR-Severe DR) with the value for precision, recall and f1-score as 0.99, 0.99 and 0.99.The second best performing binary model is (No DR-Proliferative DR) with the value for precision, recall and f1-score as 0.98, 0.98 and 0.98.The third best performing binary model is (Moderate DR- No DR) with the value for precision, recall and f1-score as 0.97, 0.97 and 0.97. The fourth best performing binary model is (Mild DR-No DR) with the value for precision, recall and f1-score as 0.94, 0.94 and 0.94. The fifth best performing binary model is (Mild DR-Severe DR) with the value for precision, recall and f1-score as 0.88, 0.88 and 0.88 followed by the (Moderate DR-

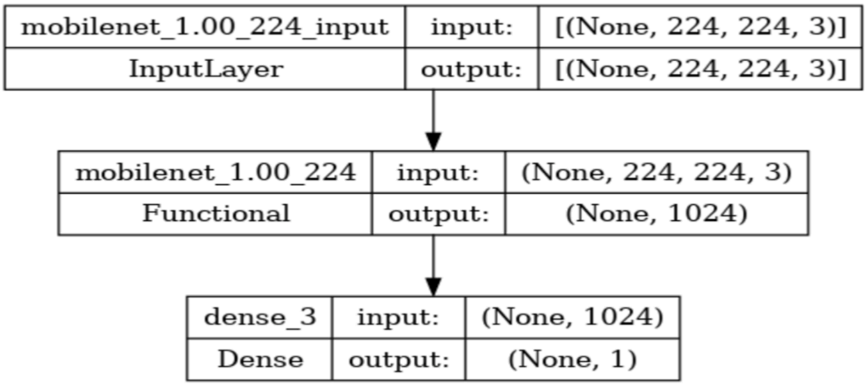


Fig. 2. Customized bi-class MobileNet.

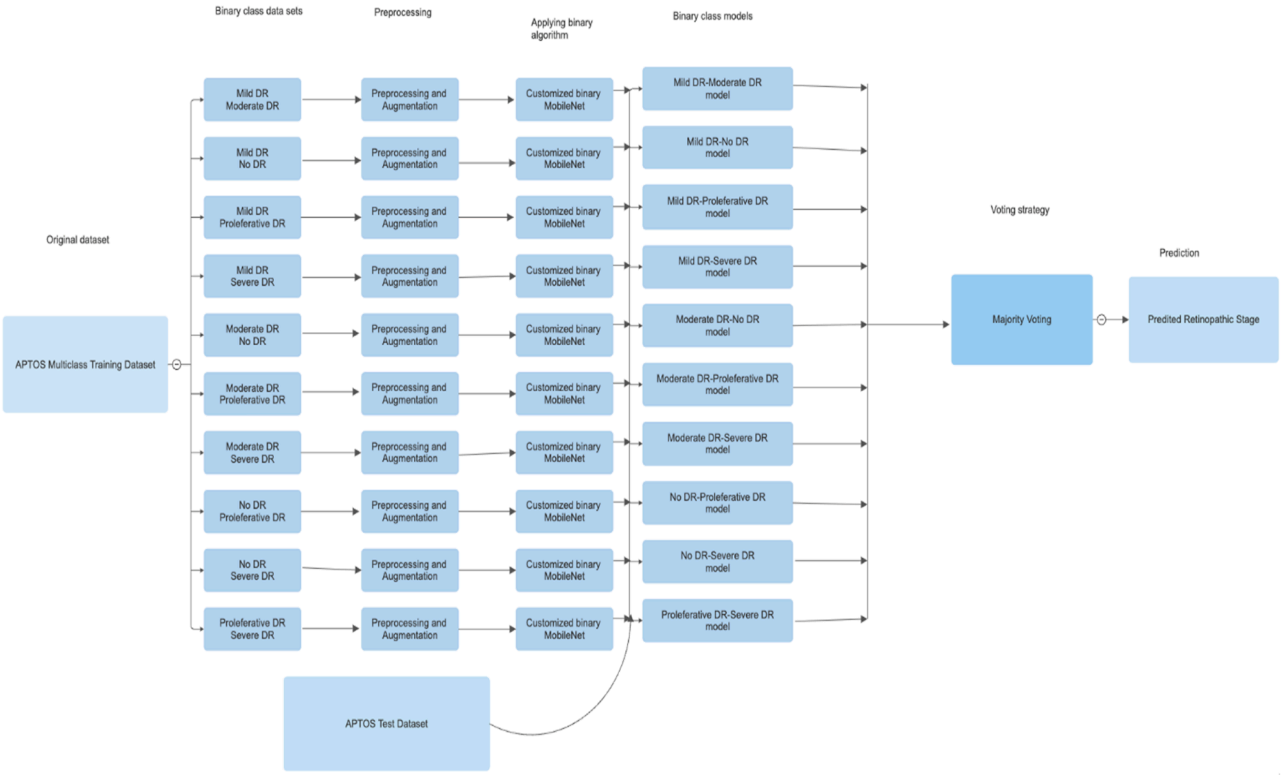

Binary class data sets
Preprocessing
Applying binary algorithm
Binary class models
Original dataset
APTOS Multiclass Training Dataset
Mild DR Moderate DR
Mild DR No DR
Mild DR Proleferative DR
Mild DR Severe DR
Moderate DR No DR
Moderate DR Proleferative DR
Moderate DR Severe DR
No DR Proleferative DR
No DR Severe DR
Proleferative DR Severe DR
Preprocessing and Augmentation
Customized binary MobileNet
Mild DR-Moderate DR model
Mild DR-No DR model
Mild DR-Proleferative DR model
Mild DR-Severe DR model
Moderate DR-No DR model
Moderate DR-Proleferative DR model
Moderate DR-Severe DR model
No DR-Proleferative DR model
No DR-Severe DR model
Proleferative DR-Severe DR model
APTOS Test Dataset
Voting strategy
Majority Voting
Prediction
Predited Retinopathic Stage

Fig. 3. Experimental framework.

Table 7
Performance of different customized bi-class base leaners on APTOS dataset.

APTOS 2019	DenseNet121	InceptionResnetV2	InceptionV3	MobileNet	MobileNetV2	VGG119	Xception
F1 Score	90.18 %	88.78 %	89.60 %	**92.00 %**	87.59 %	86.35 %	90.63 %
Accuracy	90.18 %	88.78 %	89.60 %	**92.00 %**	87.59 %	86.35 %	90.63 %
Precision	90.18 %	88.78 %	89.60 %	**92.00 %**	87.59 %	86.35 %	90.63 %
Recall	90.18 %	88.78 %	89.60 %	**92.00 %**	87.59 %	86.35 %	90.63 %

Table 8
Performance of different customized bi-class base leaners on IDRiD dataset.

IDRiD	DenseNet121	InceptionResnetV2	InceptionV3	MobileNet	MobileNetV2	VGG119	Xception
F1 Score	79.63 %	84.26 %	**85.19 %**	84.57 %	78.09 %	76.23 %	84.57 %
Accuracy	79.63 %	84.26 %	**85.19 %**	84.57 %	78.09 %	76.23 %	84.57 %
Precision	79.63 %	0.842593	**85.19 %**	84.57 %	78.09 %	76.23 %	84.57 %
Recall	79.63 %	0.842593	**85.19 %**	84.57 %	78.09 %	76.23 %	84.57 %

Table 9
Performance of different customized bi-class base leaners on messidor-2 dataset.

Messidor-2	DenseNet121	InceptionResnetV2	InceptionV3	MobileNet	MobileNetV2	VGG19	Xception
F1 Score	**85.42 %**	81.32 %	81.03 %	72.92 %	82.97 %	79.96 %	84.05 %
Accuracy	**85.42 %**	81.32 %	81.03 %	72.92 %	82.97 %	79.96 %	84.05 %
Precision	**85.42 %**	81.32 %	81.03 %	72.92 %	82.97 %	79.96 %	84.05 %
Recall	**85.42 %**	81.32 %	81.03 %	72.92 %	82.97 %	79.96 %	84.05 %

Table 10
Performance of different customized bi-class base leaners on DDR dataset.

DDR	DenseNet121	InceptionResnetV2	InceptionV3	MobileNet	MobileNetV2	VGG19	Xception
F1 Score	86.60 %	87.50 %	84.65 %	**87.89 %**	85.77 %	84.65 %	87.80 %
Accuracy	86.60 %	87.50 %	85.00 %	**87.89 %**	85.77 %	84.65 %	87.80 %
Precision	86.60 %	87.50 %	84.65 %	**87.89 %**	85.77 %	84.65 %	87.80 %
Recall	86.60 %	87.50 %	84.65 %	**87.89 %**	85.77 %	84.65 %	87.80 %

Table 11
APTOS 2019 Individual bi-class model performance.

Bi-class model Name	Precision	Recall	F1-Score	AUC
Mild DR-Moderate DR	0.79	0.79	0.79	0.73
Mild DR-No DR	0.94	0.94	0.94	0.87
Mild DR-Proliferative DR	0.81	0.80	0.80	0.79
Mild DR-Severe DR	0.88	0.88	0.88	0.85
Moderate DR- No DR	0.97	0.97	0.97	0.96
Moderate DR- Proliferative DR	0.82	0.83	0.82	0.70
Moderate DR-Severe DR	0.85	0.86	0.85	0.69
No DR-Proliferative DR	0.98	0.98	0.98	0.94
No DR-Severe DR	0.99	0.99	0.99	0.98
Proliferative DR-Severe DR	0.79	0.77	0.78	0.78

Severe DR) model, (Moderate DR- Proliferative DR) model, (Mild DR-Proliferative DR) model, (Mild DR-Moderate DR) model, (Proliferative DR-Severe DR) model. The performance ranking of these ten binary models from best to worst is depicted here in this order (No DR-Severe DR) > (No DR-Proliferative DR) > (Moderate DR- No DR) > (Mild DR-No DR) > (Mild DR-Severe DR) > (Moderate DR-Severe DR) > (Moderate DR- Proliferative DR) > (Mild DR-Proliferative DR) > (Mild DR-Moderate DR) > (Proliferative DR-Severe DR) which means that (No DR-Severe DR) base model is best at differentiating between its two classes while as (Proliferative DR-Severe DR) model is worst performing base model in ensemble but still with a precision, recall and f1-score of 0.79,0.77 and 0.78. The overall performance of the proposed ensemble after voting by these ten individual classifiers is shown in Table 15. The table shows the values of precision, recall and f1-score for the 5 different classes. (No DR) represented by 0 class has highest value for precision, recall and f1-score as 0.97, 0.99 and 0.98. (Moderate DR) represented by 2 class has second highest value for precision, recall and f1-score as 0.90, 0.92 and 0.91. (Mild DR) represented by 1 class has third highest value for precision, recall and f1-score as 0.79, 0.80 and 0.80. (Proliferative DR) represented by 4 class has fourth highest value for precision, recall and f1-score as 0.85, 0.70 and 0.77. (Severe DR) represented by 3 class has the least high value for precision, recall and f1-score as 0.77, 0.75 and 0.76.The performance ranking of these five different classes from best to worst is depicted here in this order (No DR) > (Moderate DR) > (Mild DR) > (Proliferative DR) > (Severe DR) which means identification of (No DR) class by our ensemble is easiest and (Severe DR) is toughest but still with a good precision, recall and f1-score of .77,.75 and .76. Fig. 4. displays the confusion matrix for APTOS 2019 dataset. Overall performance comparison of our proposed ensemble with other state-of-the-art models is shown in Table 19 in detail. The values for precision, recall and accuracy are 92.00, 92.00 and 92.00 which is highest among all. Table 20 shows the lightweight nature of our ensemble by displaying the total and trainable parameters of our base model and the proposed ensemble. Table 21 compares these parameters with the most closely related work (Patel et al., 2020). Consequently, the proposed ensemble model utilizes minimum computational resources than other works without compromising on the classification performance.

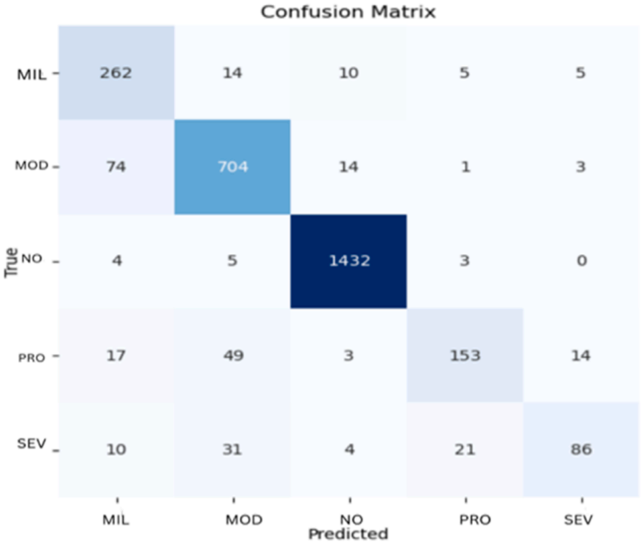


Fig. 4. APTOS 2019 dataset confusion matrix.

The performance of the individual binary base models of the ensemble created using IDRiD dataset using customised InceptionV3 is shown below in Table 12. Table 12 depicts that in terms of precision, recall, and f1-score, the best performing binary model is (Mild DR - Proliferative DR) with the values for precision, recall, and f1-score as 0.92, 0.92, and 0.92, respectively. The second best performing binary model is (No DR - Severe DR) with the values for precision, recall, and f1-score as 0.90, 0.90, and 0.90. The third best performing binary model

Table 12
IDRiD individual bi-class model performance.

Bi-class model Name	Precision	Recall	F1-Score	AUC
Mild DR-Moderate DR	0.90	0.90	0.90	0.62
Mild DR-No DR	0.90	0.90	0.90	0.73
Mild DR-Proliferative DR	0.92	0.92	0.92	0.87
Mild DR-Severe DR	0.88	0.88	0.88	0.75
Moderate DR- No DR	0.81	0.81	0.81	0.81
Moderate DR- Proliferative DR	0.86	0.86	0.86	0.72
Moderate DR-Severe DR	0.73	0.73	0.73	0.60
No DR-Proliferative DR	0.88	0.88	0.88	0.88
No DR-Severe DR	0.90	0.90	0.90	0.89
Proliferative DR-Severe DR	0.78	0.78	0.78	0.74

is (Mild DR - No DR) with the values for precision, recall, and f1-score as 0.90, 0.90, and 0.90. The fourth best performing binary model is (Mild DR - Severe DR) with the values for precision, recall, and f1-score as 0.88, 0.88, and 0.88. The fifth best performing binary model is (No DR - Proliferative DR) with the values for precision, recall, and f1-score as 0.88, 0.88, and 0.88, followed by the (Moderate DR - No DR) model, (Moderate DR - Proliferative DR) model, (Proliferative DR - Severe DR) model, and (Moderate DR - Severe DR) model. The performance ranking of these binary models from best to worst is depicted in this order:(Mild DR - Proliferative DR) > (No DR - Severe DR) > (Mild DR - No DR) > (Mild DR - Severe DR) > (No DR - Proliferative DR) > (Moderate DR - No DR) > (Moderate DR - Proliferative DR) > (Proliferative DR - Severe DR) > (Moderate DR - Severe DR). This means that the (Mild DR - Proliferative DR) base model is the best at differentiating between its two classes, while the (Moderate DR - Severe DR) model is the worst performing base model in the ensemble, with precision, recall, and f1-score values of 0.73, 0.73, and 0.73, respectively. The overall performance of the proposed ensemble after voting by these individual classifiers is shown in Table 16. The table shows the values of precision, recall, and f1-score for the 5 different classes: Class 0 (No DR) has the values for precision, recall, and f1-score as 0.89, 0.50, and 0.64, respectively. Class 1 (Mild DR) has the values for precision, recall, and f1-score as 0.80, 0.98, and 0.88. Class 2 (Moderate DR) has the highest values for precision, recall, and f1-score as 0.93, 0.87, and 0.90. Class 3 (Proliferative DR) has the values for precision, recall, and f1-score as 0.84, 0.72, and 0.78. Class 4 (Severe DR) has the values for precision, recall, and f1-score as 0.84, 0.75, and 0.79. The average precision, recall, and f1-score for the ensemble model are 86.00, 85.00, and 85.00, respectively. Overall performance comparison of our proposed ensemble with other state-of-the-art models is shown in Table 19 in detail. Fig. 5 displays the confusion matrix for the ensemble model, highlighting the performance across the five classes in the dataset.

Table 13
Messidor-2 individual bi-class model performance.

Bi-class model Name	Precision	Recall	F1-Score	AUC
Mild DR-Moderate DR	0.70	0.70	0.70	0.69
Mild DR-No DR	0.74	0.79	0.72	0.52
Mild DR-Proliferative DR	0.92	0.92	0.90	0.64
Mild DR-Severe DR	0.93	0.93	0.92	0.83
Moderate DR- No DR	0.75	0.77	0.75	0.63
Moderate DR- Proliferative DR	0.93	0.92	0.89	0.57
Moderate DR-Severe DR	0.81	0.83	0.78	0.55
No DR-Proliferative DR	0.98	0.98	0.97	0.64
No DR-Severe DR	0.97	0.97	0.97	0.80
Proliferative DR-Severe DR	0.45	0.64	0.53	0.46

The performance of the individual binary base models of the ensemble created on Messidor-2 dataset using customised DenseNet121 is shown in Table 13. Table 13 depicts that in terms of precision, recall and f1-score the best performing binary model is (No DR- Proliferative DR) with the value for precision, recall and f1-score as 0.98, 0.98 and 0.97.The second best performing binary model is (No DR-Severe DR) with the value for precision, recall and f1-score as 0.97, 0.97 and 0.97. The third best performing binary model is (Mild DR- Severe DR) with the value for precision, recall and f1-score as 0.93, 0.93 and 0.92. The fourth best performing binary model is (Moderate DR-Proliferative DR) with the value for precision, recall and f1-score as 0.93, 0.92 and 0.89. The fifth best performing binary model is (Mild DR-Proliferative DR) with the value for precision, recall and f1-score as 0.92, 0.92 and 0.90 followed by the (Moderate DR-Severe DR) model, (Moderate DR- No DR) model, (Mild DR-No DR) model, (Mild DR-Moderate DR) model, (Proliferative DR-Severe DR) model. The performance ranking of these ten binary models from best to worst is depicted here in this order (No DR- Proliferative DR) > (No DR-Severe DR) > (Mild DR- Severe DR) > (Moderate DR-Proliferative DR)> (Mild DR-Proliferative DR) > (Moderate DR-Severe DR)> (Moderate DR- No DR) > (Mild DR-No DR) > (Mild DR-Moderate DR) > (Proliferative DR-Severe DR) which means that (No DR- Proliferative DR) base model is best at differentiating between its two classes while as (Proliferative DR-Severe DR) model is worst performing base model in ensemble with a precision, recall and f1-score of 0.45,0.64 and 0.53. The overall performance of the proposed ensemble after voting by these ten individual classifiers is shown in Table 17. The table shows the values of precision, recall and f1-score for the 5 different classes. (No DR) represented by 0 class with the value for precision, recall and f1-score as 0.82, 0.68 and 0.98. (Moderate DR) represented by 2 class has highest value for precision, recall and f1-score as 0.88, 0.97 and 0.91. (Mild DR) represented by 1 class has the value for precision, recall and f1-score as 0.78, 0.78 and 0.80. (Proliferative DR) represented by 4 class has a value for precision, recall and f1-score as 0.81, 0.58 and 0.77. (Severe DR) represented by 3 class has the least high value for precision, recall and f1-score as 0.83, 0.18 and 0.76. Overall performance comparison of our proposed ensemble with other state-of-the-art models is shown in Table 19 in detail. Figure 6. displays the confusion matrix for Messidor-2 dataset.

The performance of the individual binary base models of the ensemble created on the DDR dataset using customized MobileNet is shown in Table 14. Table 14 depicts that in terms of precision, recall, F1-score, and AUC, the best performing binary model is (No DR - Severe DR), with the values for precision, recall, F1-score, and AUC as 0.98, 0.98, 0.98, and 0.93, respectively. The second best performing binary model is (No DR - Proliferative DR), with the values for precision, recall, F1-score, and AUC as 0.98, 0.98, 0.98, and 0.77, respectively. The third best performing binary model is (Moderate DR - Proliferative DR), with the values for precision, recall, F1-score, and AUC as 0.95, 0.95, 0.95,

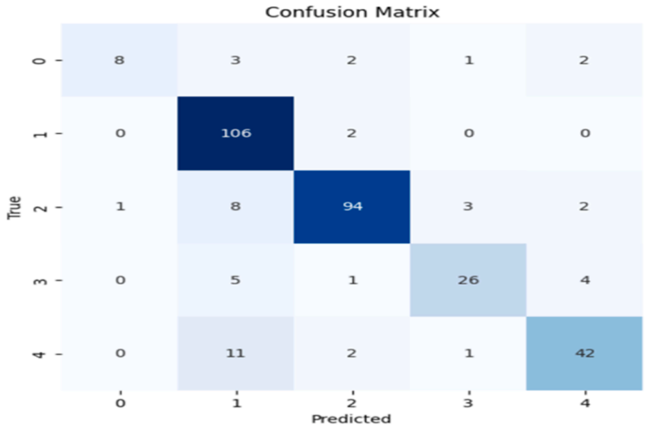


Fig. 5. IDRiD dataset confusion matrix.

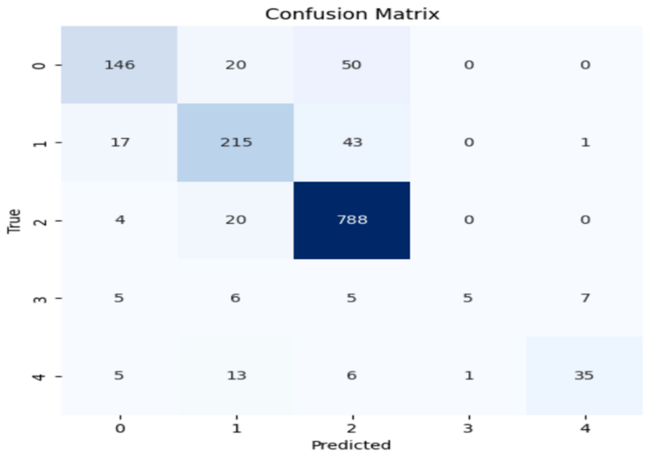


Fig. 6. Messidor-2 dataset confusion matrix.

Table 14
DDR individual bi-class model performance.

Bi-class model Name	Precision	Recall	F1-Score	AUC
Mild DR-Moderate DR	0.87	0.87	0.87	0.50
Mild DR-No DR	0.87	0.87	0.87	0.55
Mild DR-Proliferative DR	0.87	0.87	0.87	0.88
Mild DR-Severe DR	0.86	0.86	0.86	0.83
Moderate DR- No DR	0.72	0.72	0.72	0.72
Moderate DR- Proliferative DR	0.95	0.95	0.95	0.62
Moderate DR-Severe DR	0.95	0.95	0.95	0.55
No DR-Proliferative DR	0.98	0.98	0.98	0.77
No DR-Severe DR	0.98	0.98	0.98	0.93
Proliferative DR-Severe DR	0.84	0.84	0.84	0.84

Table 15
APTOS class wise and overall ensemble performance.

Class	Precision	Recall	F1-Score
0	0.97	0.99	0.98
1	0.79	0.80	0.80
2	0.90	0.92	0.91
3	0.77	0.75	0.76
4	0.85	0.70	0.77
Avg	**92.00**	**92.00**	**92.00**

Table 16
IDRiD class wise and overall ensemble performance.

Class	Precision	Recall	F1-Score
0	0.89	0.50	0.64
1	0.80	0.98	0.88
2	0.93	0.87	0.90
3	0.84	0.72	0.78
4	0.84	0.75	0.79
Avg	**86.00**	**85.00**	**85.00**

Table 17
Messidor-2 class wise and overall ensemble performance.

Class	Precision	Recall	F1-Score
0	0.82	0.68	0.98
1	0.78	0.78	0.80
2	0.88	0.97	0.91
3	0.83	0.18	0.76
4	0.81	0.58	0.77
Avg	**85.41**	**85.41**	**85.41**

and 0.62, respectively. The fourth best performing binary model is (Moderate DR - Severe DR), with the values for precision, recall, F1-score, and AUC as 0.95, 0.95, 0.95, and 0.55, respectively. The fifth best performing binary model is (Mild DR - Proliferative DR), with the values for precision, recall, F1-score, and AUC as 0.87, 0.87, 0.87, and 0.88, respectively. The sixth best performing binary model is (Mild DR - No DR), with the values for precision, recall, F1-score, and AUC as 0.87, 0.87, 0.87, and 0.55, respectively. The seventh best performing binary model is (Mild DR - Moderate DR), with the values for precision, recall, F1-score, and AUC as 0.87, 0.87, 0.87, and 0.50, respectively. The eighth best performing binary model is (Mild DR - Severe DR), with the values for precision, recall, F1-score, and AUC as 0.86, 0.86, 0.86, and 0.83, respectively. The ninth best performing binary model is (Proliferative DR - Severe DR), with the values for precision, recall, F1-score, and AUC as 0.84, 0.84, 0.84, and 0.84, respectively. The tenth and lowest ranked binary model is (Moderate DR - No DR), with the values for precision, recall, F1-score, and AUC as 0.72, 0.72, 0.72, and 0.72, respectively. The performance ranking of these ten binary models from best to worst is as follows: (No DR - Severe DR) > (No DR - Proliferative DR) > (Moderate DR - Proliferative DR) > (Moderate DR - Severe DR) > (Mild DR - Proliferative DR) > (Mild DR - No DR) > (Mild DR - Moderate DR) > (Mild DR - Severe DR) > (Proliferative DR - Severe DR) > (Moderate DR - No DR). This means that the (No DR - Severe DR) base model is the best at differentiating between its two classes, while the (Moderate DR - No DR) model is the worst performing base model in the ensemble with precision, recall, F1-score, and AUC values of 0.72, 0.72, 0.72, and 0.72, respectively. The overall performance of the proposed ensemble after voting by these ten individual classifiers is shown in Table 18. The table shows the values of precision, recall, and F1-score for the five different classes. Overall performance comparison of our proposed ensemble with other state-of-the-art models is shown in Table 19 in detail. Figure 7. displays the confusion matrix for DDR dataset.

6. Conclusion

This paper presents a comparative analysis of several customized bi-class deep learning models and their performance on multiple diabetic retinopathy (DR) datasets, aiming to identify effective models for automated DR detection. The results demonstrate that customized bi-class deep learning models like MobileNet, InceptionV3, and DenseNet121 perform optimally across various datasets, consistently achieving high accuracy, precision, recall, and F1-scores. Specifically, MobileNet emerged as the top performer on the APTOS 2019, DDR datasets, achieving an impressive accuracy of 92 % and 88 % respectively. On the IDRiD dataset, InceptionV3 was the most effective model, and on the Messidor-2 dataset, DenseNet121 outperformed other models.

The study further explores an ensemble approach that combines the above identified binary classification models, with a focus on improving the precision and recall capabilities of the individual classifiers. The "No DR - Severe DR" binary model demonstrated the best overall performance. The ensemble model, employing a voting mechanism, effectively balances high classification accuracy and efficiency, particularly on the APTOS 2019 dataset, where it achieved 92 % precision, recall, and F1-score. A key advantage of the proposed ensemble approach lies in its lightweight nature. By reducing the number of learnable parameters without compromising performance, the model offers significant computational efficiency, making it highly suitable for real-time deployment in resource-constrained environments. The reduced computational requirements were further validated through a comparison of parameters with related works, showcasing the efficiency of the ensemble approach. In terms of model architecture, this paper made a minor modification to the existing deep learning model by appending a single output node to tailor it for binary classification tasks, thus reducing the number of learnable parameters. This adjustment, coupled with the use of transfer learning, contributed to a simpler decision boundary, and minimized class overlap, ultimately enhancing the model's performance. Additionally, the binary classification approach simplifies the task, making it easier to distinguish between categories such as "No DR" and "Severe DR." In conclusion, this research demonstrates that MobileNet, InceptionV3 and DenseNet121 are highly effective models for DR classification tasks. The ensemble approach, utilizing binary classification models, further enhances performance, particularly for challenging categories such as "Severe DR." The proposed model not only achieves superior classification results but also provides an efficient

Table 18
DDR class wise and overall ensemble performance.

Class	Precision	Recall	F1-Score
0	0.69	0.49	0.58
1	0.86	0.92	0.89
2	0.91	0.91	0.91
3	0.91	0.72	0.81
4	0.83	0.64	0.72
Avg	**88.00**	**88.00**	**88.00**

Table 19
Proposed ensembles performance comparison with state-of-the-art.

Data Set	Ref.	Precision	Recall	Accuracy	AUC
APTOS 2019	(Alyoubi et al., 2021)	89.00	-	89.00	97.90
APTOS 2019	(Bodapati et al., 2021)	82.00	83.00	82.54	79.00
APTOS 2019	(Kassani et al., 2019)	87.00	88.24	83.09	91.80
APTOS 2019	(Liu et al., 2020)	91.37	98.77	86.34	-
APTOS 2019	(Bodapati et al., 2020)	80.00	81.00	81.70	-
APTOS 2019	(Dondeti et al., 2020)	76.00	77.00	77.90	-
APTOS 2019	(Majumder et al., 2021)	77.00	-	86.00	-
APTOS 2019	(Patel et al., 2020)	91.00	-	-	-
APTOS 2019	(Jian et al., 2023)	-	-	92.08	-
APTOS 2019	(AbdelMaksoud et al., 2022)			91.20	_
APTOS 2019	**Proposed Ensemble**	**92.00**	**92.00**	**92.00**	-
Messidor-2	(Mohammadi et al., 2024)	81.81	72.83	-	
Messidor-2	(Putra et al., 2020)			83.54	
Messidor-2	(Jaing et al., 2020)				
Messidor-2	**Proposed Ensemble**	**85.41**	**85.41**	**85.41**	
IDRiD	(Khan et al., 2023)			90.17	
IDRiD	(Beevi et al., 2023)			91.10	
IDRiD	(Uppamma et al., 2023)			94.20	
IDRiD	(Wang et al., 2023)			86.04	
IDRiD	(Elwin et al., 2022)			91.40	
IDRiD	**Proposed Ensemble**	**86.00**	**85.00**	**85.00**	
DDR	(Alyoubi et al., 2021)			88.6	
DDR	(Guo et al., 2023)	63.76	58.77	77.15	
DDR	(Alwakid et al., 2023)			79.67	
DDR	(Wang et al., 2021)	83.10		83.60	
DDR	**Proposed Ensemble**	**88.00**	**88.00**	**88.00**	

TABLE 20
Trainable parameters of our models.

Model	Total Parameters	Trainable Parameters
Customized MobileNet (Base model)	3,229,889	1,025
Proposed Ensemble	32,298,890	10,260

Table 21
Ensemble trainable parameters comparison.

Ref.	Total Parameters	Trainable Parameters
(Patel et al., 2020)	2,264,389	2,015,621
Proposed Ensemble	**32,298,890**	**10,260**

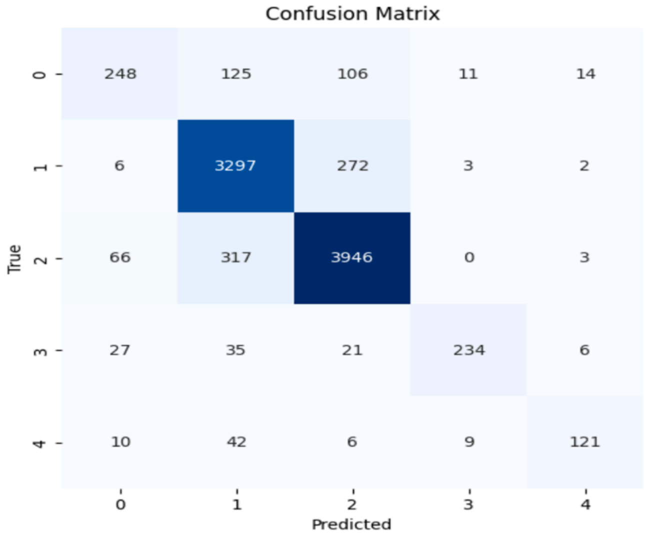


Fig. 7. DDR dataset confusion matrix.

solution for large-scale, real-time DR detection, paving the way for potential clinical applications in the early detection and management of diabetic retinopathy.

CRediT authorship contribution statement

S JAHANGEER SIDIQ: Validation, Methodology, Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Investigation, Data curation, Conceptualization. **T BENIL:** Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are thankful to the Vellore Institute of Technology, Chennai, India for providing necessary support to carry out this research work.

References

AbdelMaksoud, E., Barakat, S., & Elmogy, M. (2022). A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique. *Medical & biological engineering & computing, 60*(7), 2015–2038.

Alwakid, G., Gouda, W., Humayun, M., & Jhanjhi, N. Z. (2023). Deep learning-enhanced diabetic retinopathy image classification. *Digital Health, 9*, Article 20552076231194942.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2000). Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research, 1*, 113–141.

Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning. *Sensors, 21*(11), 3704. https://doi.org/10.3390/s21113704

Bodapati, J. D., Naralasetti, V., Shareef, S. N., Hakak, S., Bilal, M., Maddikunta, P. K. R., & Jo, O. (2020). Blended Multi-modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction. *Electronics, 9*(6), 914. https://doi.org/10.3390/electronics9060914

Bodapati, J. D., Shaik, N. S., & Naralasetti, V. (2021). Composite Deep Neural Network with Gated-Attention Mechanism for Diabetic Retinopathy Severity Classification. *Journal of Ambient Intelligence and Humanized Computing, 12*, 9825–9839. https://doi.org/10.1007/s12652-020-02749-3

Dondeti, V., Bodapati, J. D., Shareef, S. N., & Naralasetti, V. (2020). Deep Convolution Features in Non-linear Embedding Space for Fundus Image Classification. *Revue d'Intelligence Artificielle, 34*(3). https://doi.org/10.18280/ria.340301

Elwin, J. G. R., Mandala, J., Maram, B., & Kumar, R. R. (2022). AR-HGSO: Autoregressive-Henry Gas Sailfish Optimization Enabled Deep Learning Model for Diabetic Retinopathy Detection and Severity Level Classification. *Biomedical signal processing and control, 77*, Article 103712.

Gargeya, R., & Leng, T. (2017). Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology, 124*(7), 962–969. https://doi.org/10.1016/j.ophtha.2017.03.008

Guo, X., Li, X., Lin, Q., Li, G., Hu, X., & Che, S. (2023). Joint grading of diabetic retinopathy and diabetic macular edema using an adaptive attention block and semi-supervised learning. *Applied Intelligence, 53*(13), 16797–16812.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., & Takeuchi, I. (2020). Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3852–3861).

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics, 26*(2), 451–471.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hüllermeier, E., & Brinker, K. (2008). Learning valued preference structures for solving classification problems. *Fuzzy Sets and Systems, 159*(18), 2337–2352.

Hüllermeier, E., & Vanderlooy, S. (2010). Combining predictions in pairwise classification: an optimal adaptive voting strategy and its relation to weighted voting. *Pattern Recognition, 43*(1), 128–142.

... & Islam, M. R., Abdulrazak, L. F., Nahiduzzaman, M., Goni, M. O. F., Anower, M. S., Ahsan, M., & Kowalski, M. (2022). Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine, 146*, Article 105602.

Jian, M., Chen, H., Tao, C., Li, X., & Wang, G. (2023). Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images. *Computers in Biology and Medicine, 155*, Article 106631.

Kassani, S. H., Kassani, P. H., Khazaeinezhad, R., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2019). Diabetic Retinopathy Classification Using a Modified Xception Architecture. In *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISSPIT44525.2019.9008020.

Khan, I. U., Raiaan, M. A. K., Fatema, K., Azam, S., Rashid, R. U., Mukta, S. H., Jonkman, M., & De Boer, F. (2023). A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time. *Biomedicines, 11*(6), 1566.

Liu, H., Yue, K., Cheng, S., Pan, C., Sun, J., & Li, W. (2020). Hybrid Model Structure for Diabetic Retinopathy Classification. *Journal of Healthcare Engineering, 2020*. https://doi.org/10.1155/2020/5360705

Majumder, S., & Kehtarnavaz, N. (2021). *Multitasking Deep Learning Model for Detection of Five Stages of Diabetic Retinopathy*. arXiv:2103.04207.

Memon, W. R., Lal, B., & Sahto, A. A. (2017). Diabetic Retinopathy. The Prof. *Med. J., 24* (2), 234–238.

Pan, S. J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Patel, R., & Chaware, A. (2020). Transfer Learning with Fine-Tuned MobileNetV2 for Diabetic Retinopathy. In *Proceedings of the 2020 International Conference for Emerging Technology (INCET)* (pp. 1–4). Belgaum, India: IEEE.

Putra, R. E., Tjandrasa, H., Suciati, N., & Wicaksono, A. Y. (2020). Non-proliferative diabetic retinopathy classification based on hard exudates using combination of FRCNN, morphology, and ANFIS. In *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)* (pp. 1–6). IEEE.

Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., … Jadoon, W. (2019). A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection. *IEEE access : practical innovations, open solutions, 7*, 150530–150539. https://doi.org/10.1109/ACCESS.2019.2945196

Raghavendra, U., Fujita, H., Bhandary, S. V., Gudigar, A., Tan, J. H., & Acharya, U. R. (2018). Deep Convolution Neural Network for Accurate Diagnosis of Glaucoma Using Digital Fundus Images. *Information Sciences, 441*, 41–49. https://doi.org/10.1016/j.ins.2018.02.022

Rao, M., Zhu, M., & Wang, T. (2020). *Conversion and Implementation of State-of-the-Art Deep Learning Algorithms for the Classification of Diabetic Retinopathy*. arXiv:2010.11692.

Shanthi, T., & Sabeenian, R. S. (2019). Modified AlexNet Architecture for Classification of Diabetic Retinopathy Images. *Computers & Electrical Engineering, 76*, 56–64. https://doi.org/10.1016/j.compeleceng.2019.04.014

Sugeno, A., Ishikawa, Y., Ohshima, T., & Muramatsu, R. (2021). Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in Biology and Medicine, 137*, Article 104795.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning* (pp. 6105–6114). Long Beach, CA, USA: PMLR.

Uppamma, P., & Bhattacharya, S. (2023). Diabetic Retinopathy Detection: A Blockchain and African Vulture Optimization Algorithm-Based Deep Learning Framework. *Electronics, 12*(3), 742.

Usman, T. M., Saheed, Y. K., Ajibesin, A. A., & Nsang, A. S. (2024). Ens5B-UNet for Improved Microaneurysms Segmentation in Retinal Images. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1–6). IEEE.

Wang, S., Yin, Y., Cao, G., Wei, B., Zheng, Y., & Yang, G. (2015). Hierarchical Retinal Blood Vessel Segmentation Based on Feature and Ensemble Learning. *Neurocomputing,, 149*, 708–717. https://doi.org/10.1016/j.neucom.2014.08.114

... & Wang, X., Xu, M., Zhang, J., Jiang, L., Li, L., He, M., & Wang, Z. (2021). Joint learning of multi-level tasks for diabetic retinopathy grading on low-resolution fundus images. *IEEE Journal of Biomedical and Health Informatics, 26*(5), 2216–2227.

Wang, Z., Lu, H., Yan, H., Kan, H., & Jin, L. (2023). Vision Transformer Adapter-Based Hyperbolic Embeddings for Multi-Lesion Segmentation in Diabetic Retinopathy. *Scientific reports, 13*(1), Article 11178.