Classification of Diabetic Retinopathy Using ResNet50

La Ode Ansyarullah S. Sagala
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
23221066@std.stei.itb.ac.id

Agung Wahyu Setiawan
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
awsetiawan@itb.ac.id

***Abstract*—Deep learning has been proposed as an automated solution for classifying the severity levels of Diabetic Retinopathy (DR). In this study, we utilized ResNet50 architecture to classify DR using the APTOS2019 dataset. As an initial step, we initialized the model with pre-trained weights from ResNet50 on ImageNet and implemented augmentation and resampling during training. We adopted an ensemble approach combined with classifiers such as SVM, Random Forest, and Logistic Regression, resulting in a ResNet50-Ensemble (SVM+RF+LR), with outputs obtained using a Soft Voting Classifier. The model achieved an accuracy of 85%, with a precision of 0.72, recall of 0.71, and F1-score of 0.71. The AUC values for the normal, mild, moderate, severe, and proliferative classes were 1.00, 0.96, 0.95, 0.95, and 0.91, respectively, with a Macro-average AUC of 0.96. These findings indicate that the appropriate use of ensemble methods can significantly enhance DR classification performance with suitable optimization strategies.**

Keywords—APTOS2019, AUROC, diabetic retinopathy, ImageNet, ResNet50

I. INTRODUCTION

Diabetic retinopathy (DR) is a complication that affects the eyes because of diabetes mellitus (DM) and is one of the leading causes of blindness worldwide. This condition occurs due to high blood glucose levels, which can cause severe damage to the blood vessels in the retina [1]. According to the International Diabetes Federation (IDF), it is estimated that by 2045, there will be 463 million people with DM symptoms worldwide, and this number is expected to increase to 700 million.

The International Council of Ophthalmology (ICO) states that there are three stages of non-proliferative diabetic retinopathy (NPDR) based on severity: mild, moderate, and severe. In contrast, the proliferative stage presents more advanced symptoms, characterized by more pronounced signs of severe retinopathy. Therefore, early detection is crucial in helping prevent the progression of DR severity.

Computer-Aided Detection (CAD) can accelerate the diagnostic process in a more efficient manner compared to manual examination, which is time-consuming [2]. The development of CAD-based systems for DR diagnosis can assist in reducing the time required by ophthalmologists to diagnose

The symptoms of diabetic retinopathy (DR) can be detected through various lesions observed in fundus retinal images, such as microaneurysms, hemorrhages, soft exudates, and hard exudates [3]. Microaneurysms (MA) are an early sign of DR, appearing as small red dots with a round shape and a diameter of less than 125 μm. Hemorrhages (HM) are signs of DR characterized by red spots larger than 125 μm in size. Soft Exudates (cotton wool) appear as white spots on the retina, caused by swelling of nerve fibers that appear oval or round. Meanwhile, Hard Exudates (EX) are characterized by bright yellow spots with sharp edges in the outer layers of the retina, resulting from fluid leakage from the retina.

Various methods have been applied by researchers to classify the severity levels of diabetic retinopathy (DR) as an effort to address issues related to diabetes. Convolutional Neural Networks (CNNs) are commonly used in medical imaging [4], along with Deep Learning (DL) techniques [5] and Transfer Learning (TL) approaches [6]. One study applying these methods is by Taufiqurrahman et al. [7], who used the MobileNetV2 model, pre-trained on the ImageNet dataset. The classifier used was SVM, resulting in a hybrid model (MobileNetV2-SVM) with an accuracy of 85% and AUROC values of 1.00, 0.82, 0.94, 0.94, and 0.93 for the normal, mild, moderate, severe, and proliferative classes, respectively. Patel and Chaware [8] modified MobileNetV2 by adding a GlobalAveragePooling2D layer. Initially, the model was trained with frozen layers to avoid updating weights. Several layers were then fine-tuned to improve performance, successfully increasing training accuracy from 70% to 90% and validation accuracy from 50% to 81%. Another study by Mungloo-Dilmohammud et al. [9] used the private Blind Mauritian dataset with models VGG16, ResNet50, and DenseNet169 to identify the best model. The results showed that ResNet50 was the best model, achieving 82% accuracy in initial trials and 79% accuracy when classifying the Blind Mauritian dataset.

Based on various studies conducted, this research aims to classify the severity of DR using the ResNet50 model with a Transfer Learning (TL) approach. Furthermore, the analysis will be performed by utilizing ResNet50 as a feature extractor, replacing the classification layer with several different types of classifiers.

II. METHODS

The workflow of the proposed DR severity classification system is shown in Figure 1. For further clarification, the following is a detailed process for each stage of the workflow.

 How to cite: Sagala, L. O. A. S., Setiawan, A.W. (2025). Classification of Diabetic Retinopathy Using ResNet50. *JAREE (Journal on Advanced Research in Electrical Engineering), 9(2).*

Fig. 1. Research workflow stages.

A. Dataset

This study uses the APTOS2019 dataset, collected from the Aravind Eye Hospital in India. This public dataset contains 3,662 fundus retinal images, acquired through fundus photography techniques. The images are classified

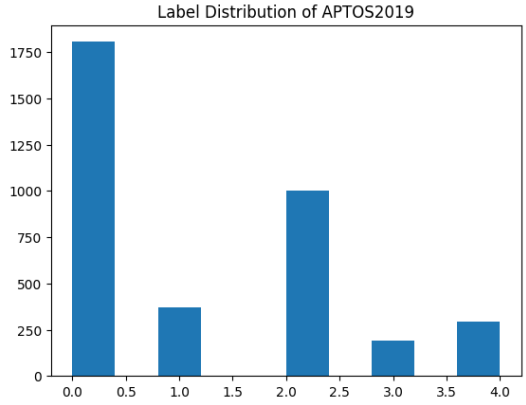

Fig. 2. Label distribution for APTOS2019 (N = 3662 images, DR severity labels are 0 to 4).

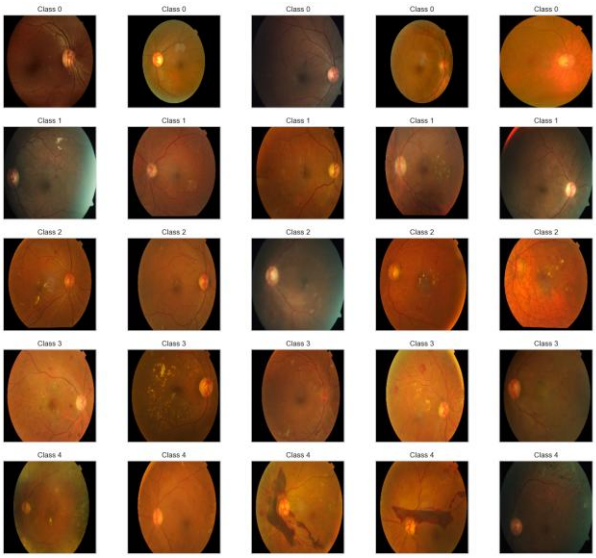

Fig. 3. Sample images from APTOS2019, the top to bottom rows corresponds to the gradual DR severity of class 0 (normal DR) to class 4 (proliferative DR)

into five categories representing the severity levels of diabetic retinopathy (DR): class 0 for normal DR, class 1 for mild DR, class 2 for moderate DR, class 3 for severe DR, and class 4 for proliferative DR. The resolution of the images in this dataset ranges from 474 x 358 pixels to 4288 x 2848 pixels.

Figure 2, shows the distribution of fundus retinal images, with a total of 1,805, 370, 999, 193, and 295 images for each class. Out of the total 3,662 fundus retinal images, the dataset is divided into 90% for training data and 10% for testing data during the training process.

B. Image Preprocessing

The APTOS2019 dataset contains images of varying sizes, with unnecessary black areas in each image that may interfere with the feature extraction process. However, the proportion of these black areas varies across images. Therefore, the first step was to resize each image to 224 x 224 pixels. Next, the proportion of the black areas was standardized using the auto-cropping method from Graham's [10], as shown in Figure 4.

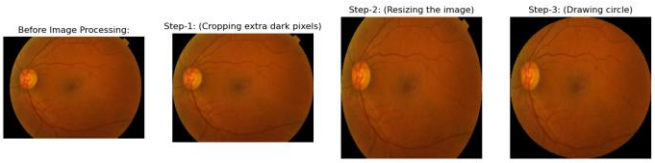

Fig. 4. Auto cropping technique with resize image

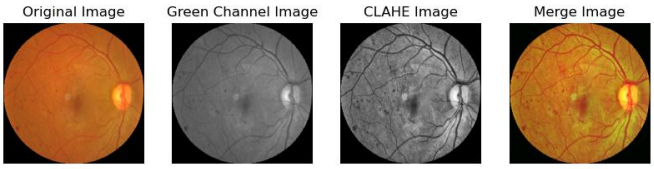

Fig. 5. Image Preprocessing

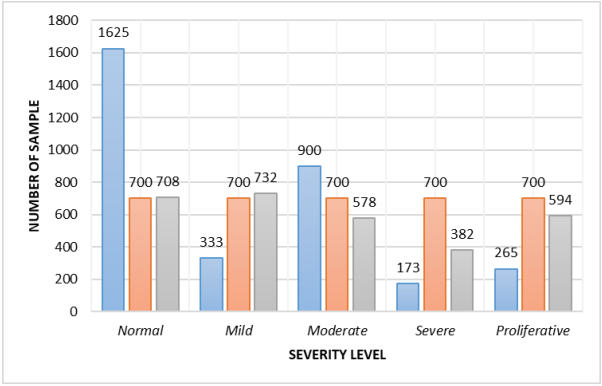

Fig. 6. Original and resampled training data distribution

To enhance the contrast in the APTOS2019 dataset, previous studies used in [11] [12] have shown significant results. Based on this, the present study will adopt a preprocessing technique by utilizing the green channel, which is processed using CLAHE, and then combining it to produce optimal images, as shown in Figure 5.

C. Augmentation and Resampling

Figure 2 shows that the data distribution in the APTOS2019 dataset is highly imbalanced. Therefore, this study applies augmentation and resampling techniques to balance the data distribution in the training set. Augmentation is performed by enlarging the images (90%), flipping the images (both horizontally and vertically), and applying random rotations to the images (0-45 degrees).

In the 10-fold cross-validation scheme, this study will use two resampling approaches. The first approach balances the number of samples in each class by equalizing the counts, resulting in 700 samples per class or a total of 3,500 training samples. The second approach involves randomly splitting the data in each class, yielding varying numbers of training samples per class, with a total of 2,994 training

samples, as shown in Figure 6, which displays the training data distribution for both the original data and the two resampling schemes.

D. Training

This study uses the ResNet50 architecture, which has been pre-trained on ImageNet.

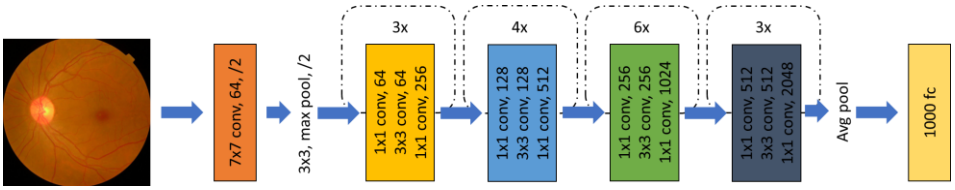

Fig. 7. Illustration of ResNet50 architecture

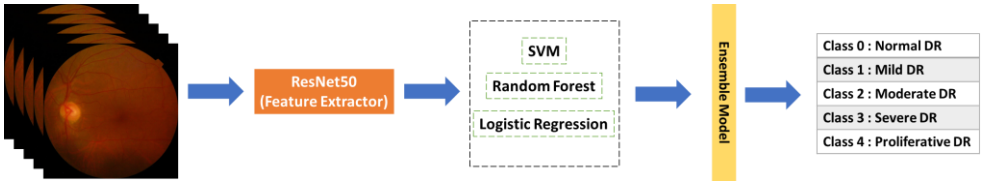

Fig. 8. Workflow Ensemble Learning

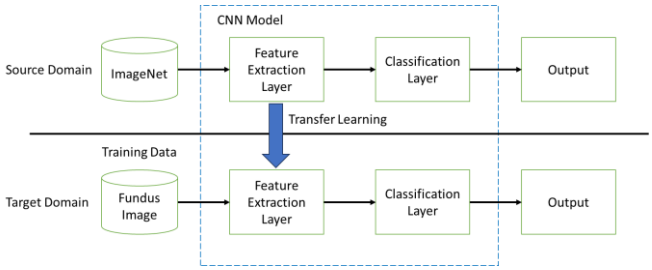

Fig. 9. Transfer Learning process

1) ResNet50

ResNet50 is a CNN model with 50 layers, achieving an error rate of 3.6% [13] and a total of 25.6 million parameters. This model consists of 48 convolutional layers, along with 1 MaxPool layer and 1 AveragePool layer. ResNet50 is known for its "skip connection" approach, which helps address the gradient vanishing problem during training, as shown in Figure 7. The illustration of the ResNet50 model architecture.

This study will use the ResNet50 model by applying Transfer Learning (TL), as implemented by He et al. [13]. The model's pre-trained weights on the ImageNet dataset will be fine-tuned using the APTOS2019 dataset for 100 epochs with a batch size of 32. The optimizer used is Adam with a learning rate of 0.0001, while the loss function employed is categorical cross-entropy to measure how well the model predicts the correct class in a multi-class classification task.

2) ResNet50 as Feature Extractor

In Transfer Learning (TL), the weights in the pre-trained convolutional layers are typically kept fixed, while the fully connected layers are retrained using the new dataset. This approach assumes that the convolutional layers, trained on a large-scale dataset, serve as effective feature extractors. Subsequently, an analysis will be conducted on the use of the convolutional layers of the ResNet50 model as a fixed feature extractor, with the classification layer replaced by various types of classifiers, such as Support Vector Machine (SVM). The TL process flow is shown in Figure 8.

In this study, each retinal image is represented as a feature value derived from the extraction process at the last fully connected layer. The SVM classifier will use these feature values from the training data with default parameters and then evaluate its performance on the test data. To enhance the model's results, the researchers assess the model's ability to classify the severity of DR by calculating the Area Under the Receiver Operating Characteristic (AUROC) for each class.

E. Ensemble Learning

Ensemble Learning (EL) is a technique that combines multiple models to enhance predictive performance and help reduce the risk of overfitting. This approach has been applied by Mondal et al. [14], who demonstrated that model fusion can yield better predictive performance compared to a single model. The workflow for using the EL method in this study is shown in Figure 8.

This approach will implement several types of classifiers, such as SVM, Random Forest, and Logistic Regression, utilizing the ResNet50 model as a feature extractor. The results will be obtained using the Soft Voting Classifier.

F. Performance Evaluation

To analyze the classification results in this study, a confusion matrix is used to measure the model's performance, as shown in Figure 10.

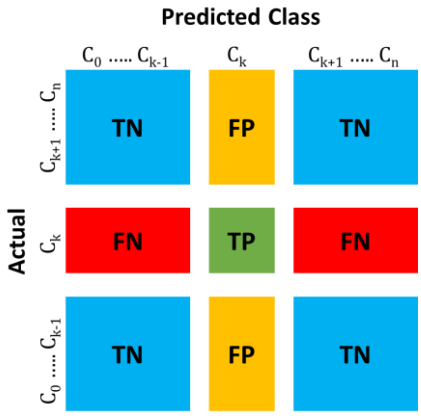

Fig. 10. Confusion Matrix

True Positive (TP) occurs when the model predicts positive, and the outcome is indeed positive. True Negative (TN) occurs when the model predicts negative, and the outcome is truly negative. False Positive (FP) occurs when the model predicts positive, but the outcome is negative. False Negative (FN) occurs when the model predicts negative, but the outcome is actually positive.

Common metrics used in multi-class classification include accuracy, precision, recall, and F1-score with a macro-average (M) approach, which the equations shown in Equation (1), (2), (3), and (4). The macro-average is chosen because the metrics are calculated independently for each class, and then the average is taken, ensuring that each class has equal weight. This metric is preferred in classification tasks with imbalanced datasets, as it can highlight greater errors when the model performs poorly on minority classes (mild and severe) [15].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

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

The final metric used in this study is the ROC or AUROC. The ROC curve visualizes the effectiveness of the model across various thresholds, showing the relative trade-off between the true positive rate (TPR) and the false positive rate (FPR), which the equations shown in Equation (5) and (6).

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{TN}{TN + FP} \quad (6)$$

AUROC indicates how well the model assigns higher probabilities to positive examples compared to negative ones. The macro-average AUROC value is used as a score to evaluate the model's performance. This value is applied to label predictions with probabilities, so in this study, the metric is measured only on the SVM output and not on the ResNet50 classification layer output. This metric has previously been used to display the classification performance of DR for existing models [16] [17].

III. Results and Discussion

A. Performance of ResNet50

Table 1 presents the performance of the ResNet50 model on the test data using a 10-fold cross-validation scheme. The model achieved an average accuracy, precision, recall, and F1-score of 79%, 0.64, 0.65, and 0.64, respectively. The confusion matrix associated with the fold that demonstrated the best model performance (Fold 8, as shown in Table 1) is presented in Table 2. The results indicate that the accuracy, precision, recall, and F1-score for this fold reached 83%, 0.69, 0.71, and 0.69, respectively.

Table 2 shows that the overall performance of the ResNet50 model was best in classifying the normal and severe classes. Conversely, the model experienced misclassifications in the mild, moderate, and proliferative classes. Specifically, the mild class was often misclassified as moderate, the moderate class as severe, and the proliferative class was frequently misclassified as severe. These misclassifications tend to occur among classes that are adjacent in severity levels.

TABLE 1. Performance of ResNet50 on 10-fold cross validation scheme

Fold	**Metrics**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
1	76%	0.61	0.63	0.60
2	81%	0.66	0.66	0.66
3	81%	0.66	0.67	0.67
4	81%	0.65	0.62	0.63
5	81%	0.69	0.70	0.68
6	81%	0.68	0.68	0.66
7	76%	0.59	0.61	0.59
8	83%	0.69	0.71	0.69
9	79%	0.62	0.61	0.61
10	74%	0.57	0.58	0.57
Average	79%	0.64	0.65	0.64

TABLE 2. Confusion matrix of ResNet50 classification results on fold-8 test data

		Predicted Class				
		Normal	*Mild*	*Moderate*	*Severe*	*Proliferative*
Actual	*Normal*	174	5	1	0	0
	Mild	1	23	11	0	2
	Moderate	1	5	77	12	5
	Severe	0	1	2	13	3
	Proliferative	1	0	6	7	15

B. Performance of ResNet50-SVM

The results shown in Table 2 demonstrate improved performance when using classifiers compared to the traditional classification layer of ResNet50. By leveraging convolutional layers as feature extractors, each retinal image was subsequently trained using an SVM classifier with a 10-fold cross-validation scheme.

Table 3 presents the performance of the model using a 10-fold cross-validation scheme. The results show that the model achieved an overall average accuracy, precision, recall, and F1-score of 82%, 0.67, 0.66, and 0.66,

respectively, which are higher than the performance metrics of the ResNet50 model for each metric used. The confusion matrix corresponding to the fold with the best model performance (Fold 8 in Table 3) is shown in Table 4. These results indicate that the accuracy, precision, recall, and F1-score for this fold reached 84%, 0.71, 0.71, and 0.71, respectively.

TABLE 3. Performance of ResNet50-SVM on 10-fold cross validation scheme

Fold	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-Score*
1	82%	0.66	0.66	0.66
2	81%	0.66	0.68	0.67
3	81%	0.66	0.68	0.66
4	83%	0.68	0.67	0.67
5	84%	0.71	0.70	0.70
6	83%	0.70	0.70	0.70
7	79%	0.61	0.61	0.61
8	84%	0.71	0.71	0.71
9	79%	0.60	0.59	0.59
10	82%	0.67	0.64	0.65
Average	82%	0.67	0.66	0.66

The confusion matrix for the best-performing ResNet50-SVM model is shown in Table 4. Compared to the confusion matrix for the best-performing ResNet50 model in Table 2, the hybrid ResNet50-SVM model demonstrated better performance in classifying majority classes (normal and moderate) compared to minority classes. However, it exhibited strong performance in distinguishing adjacent severity levels, particularly between the mild and proliferative classes.

The ROC curve of the best-performing ResNet50-SVM model is shown in Figure 11, where the model demonstrated excellent performance in classifying the severity levels of DR. This is evident from the high AUC values across all classes, with a Macro-Average AUC of 0.95. The normal class achieved the best performance with an AUC of 1.00, indicating that the model is highly accurate in classifying the normal class.

For the performance in other classes, the mild class achieved an AUC of 0.92, the moderate class an AUC of 0.95, the severe class an AUC of 0.94, and the proliferative class an AUC of 0.91, indicating strong results. With all AUC values exceeding 0.90, the model demonstrates a robust ability to classify the severity levels of DR effectively.

TABLE 4. Confusion matrix of ResNet50-SVM classification results on fold-8 test data

		Predicted Class				
		Normal	*Mild*	*Moderate*	*Severe*	*Proliferative*
Actual	*Normal*	175	5	0	0	0
	Mild	2	24	10	0	1
	Moderate	1	6	81	7	5
	Severe	0	1	3	11	4
	Proliferative	1	0	7	5	16

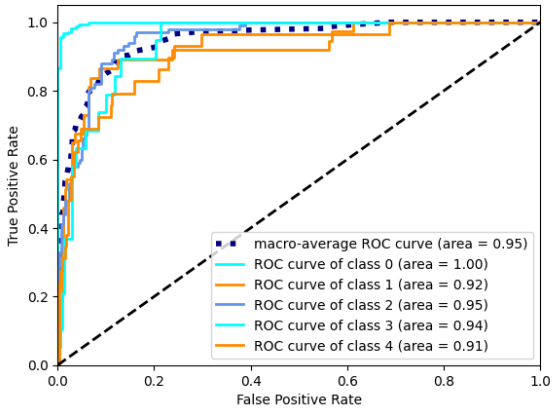

Fig. 11. ROC curves for each class and Macro-Average of the ResNet50-SVM on fold-8 test data.

C. *Performance Ensemble*

The ensemble method was tested with four ensemble approach scenarios: ResNet50-Ensemble (SVM+RF+LR), ResNet50-Ensemble (SVM+RF), ResNet50-Ensemble (SVM+LR), and ResNet50-Ensemble (RF+LR), using a 10-fold cross-validation scheme. By combining these models, it is expected to enhance the performance of the model in classifying the severity levels of DR.

Table 5 shows the average results of the performance evaluation comparison of the ensemble method across the four scenarios. The results indicate that using the ensemble method provides better model performance in terms of accuracy and stability. Specifically, the ResNet50-Ensemble (SVM+RF+LR) and ResNet50-Ensemble (SVM+LR) scenarios demonstrated the best model performance in terms of accuracy, recall, and F1-score. However, the ResNet50-Ensemble (SVM+RF+LR) scenario outperformed the others in precision. Therefore, the ensemble method in the ResNet50-Ensemble (SVM+RF+LR) scenario is considered the better approach.

TABLE 5. Comparison of model performance with the ensemble method using a 10-fold cross-validation scheme.

Classifier	**Average Performance**			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Output ResNet50	79%	0.64	0.65	0.64
ResNet50-SVM	82%	0.67	0.66	0.66
ResNet50-Ensemble (SVM+RF+LR)	83%	0.69	0.66	0.67
ResNet50-Ensemble (SVM+RF)	82%	0.67	0.65	0.66
ResNet50-Ensemble (SVM+ LR)	83%	0.68	0.66	0.67
ResNet50-Ensemble (RF+LR)	82%	0.68	0.65	0.66

Table 6 shows the performance of the ResNet50-Ensemble (SVM+RF+LR) model using a 10-fold cross-validation scheme. The model achieved an average accuracy, precision, recall, and F1-score of 83%, 0.69, 0.66, and 0.67, respectively, which are higher than the performance of the ResNet50-SVM model across all metrics. The confusion matrix corresponding to the fold that demonstrated the best model performance (Fold 8 in Table

6) is shown in Table 7. These results indicate that the accuracy, precision, recall, and F1-score for this fold reached 85%, 0.72, 0.71, and 0.71, respectively.

TABLE 6. Performance of Ensemble on 10-fold cross validation scheme

Fold	**Metrics**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
1	83%	0.66	0.65	0.65
2	83%	0.68	0.68	0.68
3	82%	0.67	0.68	0.67
4	84%	0.71	0.68	0.69
5	84%	0.73	0.69	0.70
6	83%	0.70	0.68	0.68
7	80%	0.64	0.63	0.63
8	85%	0.72	0.71	0.71
9	81%	0.63	0.60	0.61
10	82%	0.69	0.63	0.65
Average	83%	0.69	0.66	0.67

The confusion matrix for the best-performing ResNet50-Ensemble (SVM+RF+LR) model is shown in Table 7. Compared to the confusion matrix for the best-performing ResNet50-SVM model in Table 4, the ResNet50-Ensemble (SVM+RF+LR) model demonstrates strong overall performance, especially in recognizing the normal, moderate, and proliferative classes. However, there are some misclassifications, such as mild being classified as moderate, severe as proliferative, and proliferative as moderate. These misclassifications may be due to the similarity of features between these classes, such as mild with moderate, and moderate with severe and proliferative.

The ROC curve for the best-performing ResNet50-Ensemble (SVM+RF+LR) model is shown in Figure 12. It demonstrates superior performance in classifying DR severity levels compared to the performance of the ResNet50-SVM model displayed in Figure 11. This is indicated by the high Macro-Average AUC value of 0.96, which signifies that the model has excellent classification ability across all classes.

The AUC values obtained for each class are as follows: the normal class with an AUC of 1.00, the mild class with an AUC of 0.96, the moderate class with an AUC of 0.95, the severe class with an AUC of 0.95, and the proliferative class with an AUC of 0.91. With all AUC values above 0.90, these results indicate that the model has excellent ability in classifying DR severity levels.

TABLE 7. Confusion matrix of Ensemble classification results on fold-8 test data

		Predicted Class				
		Normal	*Mild*	*Moderate*	*Severe*	*Proliferative*
Actual	*Normal*	175	5	0	0	0
	Mild	2	24	10	0	1
	Moderate	1	4	83	7	5
	Severe	0	0	4	9	6
	Proliferative	1	0	7	3	18

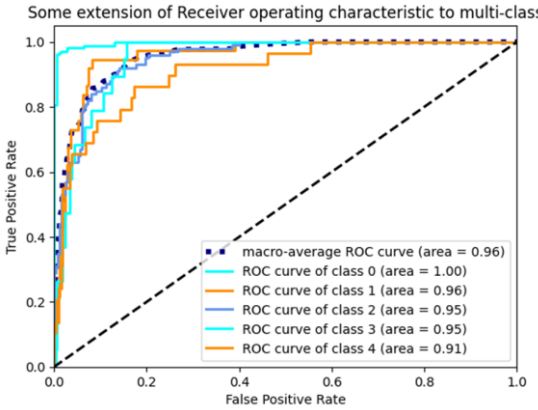

Fig. 12. ROC curves for each class and Macro-Average of the Ensemble on fold-8 test data.

D. Comparison with Other Results

Several studies have successfully classified DR severity using Convolutional Neural Network (CNN) models effectively through fundus image analysis, as demonstrated by Zhang et al. [11], Eko et al. [18], and Liu et al. [17]. Additionally, there are studies that use features extracted by CNNs as inputs for traditional Machine Learning (ML) models for the final classification process. This approach aims to maximize the CNN's ability to detect complex visual patterns in fundus images by leveraging the strengths of models, as done by Yaqoob et al. [19] and Mohanty et al. [20]. These studies applied ML techniques such as Random Forest or XGBoost to achieve more accurate and robust classifications.

CNN models are effective in classifying DR severity through fundus image analysis. Several studies also integrate features extracted by CNNs with traditional Machine Learning (ML) models, such as Random Forest or XGBoost, to improve classification accuracy. This approach optimizes the CNN's ability to capture complex visual patterns in fundus images, while leveraging the strength of ML models to achieve more accurate and robust classifications. Additionally, there are studies that apply ensemble models, where pre-trained models are combined, as demonstrated by Qummar et al. [16], that uses the ensemble method with a Stacking Classifier in their model training.

Table 7 presents a comparison with previous studies on the classification of DR severity. The results show that the proposed ensemble model performs exceptionally well in terms of accuracy, precision, recall, f1-score, and AUC for each class, making it superior to the other models and optimal for classification.

TABLE 8. Comparison of model performance in classification of DR with another research

Method	Performance
(Qummar, et al. [16]) **Model Ensemble**	Accuracy : 80.8%, Precision : 63.85%, Recall : 51.5%, Specificity : 86.72%, F1-Score : 53.74% AUC Macro-Average : 0.87 AUC Normal : 0.85, AUC Mild : 0.71, AUC Moderate : 0.85, AUC Severe : 0.96, AUC Proliferative : 0.97
(Yaqoob, et al. [19]) **ResNet50-Random Forest**	Accuracy: 75.09%
(Zhang, et al. [11]) **ResNet50**	Accuracy: 83.7%
(Eko, et al. [18]) **EfficientNet-B7**	Accuracy: 84.36%
(Mohanty, et al. [20]) **Model *Hybrid*** (VGG16-XGBoost)	Accuracy: 79.50%
(Liu, et al. [17]) **ResNet50** (*Transfer Learning-Fine Tuning)*	Accuracy: 81.97% AUC = 0.9531
(Liu, et al. [17]) **InceptionV3** (*Transfer Learning-Fine Tuning*)	Accuracy: 83.61% AUC = 0.9256
ResNet50-SVM (This paper)	Accuracy: 84%, Precision; Recall; F1-Score : 0.71 AUC Macro-Average : 0.95 AUC Normal : 1.00, AUC Mild : 0.92, AUC Moderate : 0.95, AUC Severe : 0.94, AUC Proliferative : 0.91
ResNet50-Ensemble (SVM+RF+LR) (This Paper)	Accuracy: 85%, Precision : 0.72, Recall; F1-Score : 0.71 AUC Macro-Average : 0.96 AUC Normal : 1.00, AUC Mild : 0.96, AUC Moderate : 0.95, AUC Severe : 0.95, AUC Proliferative : 0.91

IV. CONCLUSION

In this study, we propose the use of an ensemble learning model, the ResNet50-Ensemble (SVM+RF+LR) model, to classify DR severity levels. This model achieved accuracy, precision, recall, and F1-score of 85%, 0.72, 0.71, and 0.71, respectively, as well as AUC values of 1.00, 0.96, 0.95, 0.95, and 0.91 for the normal, mild, moderate, severe, and proliferative classes, with a macro-average AUC of 0.96. These results demonstrate that the model has a good balance between accuracy and AUC across all classes. This indicates that with the appropriate optimization strategy and proper parameter use, the model's performance in DR classification can be enhanced, providing a significant contribution to early detection and more accurate diagnosis.

REFERENCES

[1] R. Yasashvini, V. Raja Sarobin M, R. Panjanathan, S. Graceline Jasmine, dan L. Jani Anbarasi, "Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks," *Symmetry (Basel).*, vol. 14, no. 9, 2022, doi: 10.3390/sym14091932.

[2] M. Saini dan S. Susan, "Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets," *Comput. Biol. Med.*, vol. 149, no. August, hal. 105989, 2022, doi: 10.1016/j.compbiomed.2022.105989.

[3] A. Mustapha, L. Mohamed, H. Hamid, dan K. Ali, "Diabetic Retinopathy Classification Using ResNet50 and VGG-16 Pretrained Networks," *Int. J. Comput. Eng. Data Sci.*, vol. 1, no. 1, hal. 2737–8543, 2021, [Daring]. Tersedia pada: www.ijceds.com.

[4] N. M. A. Tajudin *et al.*, "Deep learning in the grading of diabetic retinopathy: A review," *IET Comput. Vis.*, vol. 16, no. 8, hal. 667–682, 2022, doi: 10.1049/cvi2.12116.

[5] M. Y. T. Yip *et al.*, "Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy," *npj Digit. Med.*, vol. 3, no. 1, hal. 31–34, 2020, doi: 10.1038/s41746-020-0247-1.

[6] M. Z. Atwany, A. H. Sahyoun, dan M. Yaqub, "Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey," *IEEE Access*, vol. 10, hal. 28642–28655, 2022, doi: 10.1109/ACCESS.2022.3157632.

[7] S. Taufiqurrahman, A. Handayani, B. R. Hermanto, dan T. L. E. R. Mengko, "Diabetic Retinopathy Classification Using A Hybrid and Efficient MobileNetV2-SVM Model," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2020-Novem, hal. 235–240, 2020, doi: 10.1109/TENCON50793.2020.9293739.

[8] R. Patel dan A. Chaware, "Transfer learning with fine-tuned MobileNetV2 for diabetic retinopathy," *2020 Int. Conf. Emerg. Technol. INCET 2020*, hal. 6–9, 2020, doi: 10.1109/INCET49848.2020.9154014.

[9] Z. Mungloo-Dilmohamud, M. H. M. Khan, K. Jhumka, B. N. Beedassy, N. Z. Mungloo, dan C. Peña-Reyes, "Balancing Data through Data Augmentation Improves the Generality of Transfer Learning for Diabetic Retinopathy Classification," *Appl. Sci.*, vol. 12, no. 11, 2022, doi: 10.3390/app12115363.

[10] B. Graham, "Kaggle Diabetic Retinopathy Detection competition report," *Kaggle*, hal. 1–9, 2015.

[11] J. Zhang, B. Xie, X. Wu, R. Ram, dan D. Liang, "Classification of Diabetic Retinopathy Severity in Fundus Images with DenseNet121 and ResNet50," hal. 1–15, 2021, [Daring]. Tersedia pada: http://arxiv.org/abs/2108.08473.

[12] R. N. Lazuardi, N. Abiwinanda, T. H. Suryawan, M. Hanif, dan A. Handayani, "Automatic diabetic retinopathy classification with efficientnet," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2020-Novem, hal. 756–760, 2020, doi: 10.1109/TENCON50793.2020.9293941.

[13] K. He, X. Zhang, S. Ren, dan J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016, vol. 7, no. 1, hal. 770–778, doi: 10.1109/CVPR.2016.90.

[14] S. S. Mondal, N. Mandal, K. K. Singh, A. Singh, dan I. Izonin, "EDLDR: An Ensemble Deep Learning Technique for Detection and Classification of Diabetic Retinopathy," *Diagnostics*, vol. 13, no. 1, hal. 124, Des 2022, doi: 10.3390/diagnostics13010124.

[15] H. Narasimhan, W. Pan, P. Kar, P. Protopapas, dan H. G. Ramaswamy, "Optimizing the Multiclass F-Measure via Biconcave Programming," *2016 IEEE 16th Int. Conf. Data Min.*, hal. 1101–1106, 2017, doi: 10.1109/icdm.2016.0143.

[16] S. Qummar *et al.*, "A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection," *IEEE Access*, vol. 7, hal. 150530–150539, 2019, doi: 10.1109/ACCESS.2019.2947484.

[17] K. Liu, T. Si, C. Huang, Y. Wang, H. Feng, dan J. Si, "Diagnosis and detection of diabetic retinopathy based on transfer learning," *Multimed. Tools Appl.*, vol. 83, no. 35, hal. 82945–82961, 2024, doi: 10.1007/s11042-024-18792-x.

[18] A. Eko, M. Hazmi, C. Mandiri, Y. Azhar, dan F. Bimantoro, "Classification of Diabetic Retinopathy Disease Using Convolutional Neural Network," vol. 6, no. March, hal. 12–18, 2022.

[19] M. K. Yaqoob, S. F. Ali, M. Bilal, M. S. Hanif, dan U. M. Al-Saggaf, "Resnet based deep features and random forest classifier for diabetic retinopathy detection†," *Sensors*, vol. 21, no. 11, hal. 1–14, 2021, doi: 10.3390/s21113883.

[20] C. Mohanty *et al.*, "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," *Sensors*, vol. 23, no. 12, hal. 5726, Jun 2023, doi: 10.3390/s23125726.