IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 July 2023, accepted 1 August 2023, date of publication 3 August 2023, date of current version 10 August 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3301618*

RESEARCH ARTICLE

# Diabetic Retinopathy Detection and Grading: A Transfer Learning Approach Using Simultaneous Parameter Optimization and Feature-Weighted ECOC Ensemble

**W. K. WONG[1], FILBERT H. JUWONO[2], (Senior Member, IEEE), AND CATUR APRIONO[3], (Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Curtin University Malaysia, Miri 98009, Malaysia
[2]Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China
[3]Department of Electrical Engineering, Universitas Indonesia, Depok 16424, Indonesia

Corresponding author: Catur Apriono (catur@eng.ui.ac.id)

This work was supported by Universitas Indonesia's International Indexed Publication (PUTI) Q1 Grant, in 2022, under Contract NKB-500/UN2.RST/HKP.05.00/2022.

**ABSTRACT** Early detection of Diabetic Retinopathy (DR) is crucial as it may cause blindness. Manual diagnosis of DR severity by ophthalmologists is challenging and time consuming. Therefore, there has been a significant focus on developing an automated system for identifying DR using retinal fundus images. Recent research has revealed that utilizing pre-trained deep learning networks for diverse image classification tasks provides notable benefits in this context. In this paper, a Transfer Learning (TL) approach with optimized feature weights and parameters is proposed for DR detection and grading tasks. To obtain better generalization during training and to optimize classification, features are extracted from the average pooling layers and fed to an Error Correction Output Code (ECOC) ensemble configuration. Two pre-trained networks (ShuffleNet and ResNet-18) are considered as each pre-trained network offers a different "point of view" of the fundus images, thereby providing more opportunities for accurate "grade-wise" discrimination. A simultaneous feature selection and parameter tuning of the ensemble is applied to further enhance the overall DR detection and grading. Adaptive Differential Evolution (ADE) is chosen for this purpose because it automatically configures the parameters, eliminating the need for manual parameter selection. In this paper, we evaluate two public domain datasets: 1) APTOS and 2) combination of EyePac + Messidor-2. Simulation results show that our proposed method performs better that the conventional deep learning models and are on a par with the existing research work. In particular, the optimal configuration for APTOS 5-class DR grading achieves an accuracy rate of 82%, while for APTOS 2-class grading, it achieves a higher accuracy rate of 96%. Finally, the best configuration for EyePac + Messidor-2 3-class grading results in 75% accuracy.

**INDEX TERMS** Diabetic retinopathy, stochastic optimization, differential evolution, ECOC optimization.

## I. INTRODUCTION

Diabetes is a metabolic condition indicated by high level of blood sugar in the body. In particular, not enough insulin is produced by the body to transport glucose into the cells. Diabetes may be caused by the increase of standard of living in recent years due to advances in science and technology and economic growth, which has resulted in a decrease in regular exercise and an increase in the aging rate [1]. According to statistics, there are currently about 347 million diabetes patients in the world and it is estimated that global diabetes patients will reach 4.4% of the world's population in 2030 [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

The condition known as Diabetic Retinopathy (DR), which is a damage to the retina, can occur as a diabetes complication. Severe DR may cause blindness if not well treated [3]. When a person has had diabetes for at least ten years, DR symptoms may start to manifest. Since there are not any apparent indications or symptoms of diabetes in its early stages, detecting the DR symptoms is a challenging task [4]. Therefore, early detection of retinopathy is very important. It can prevent retinopathy at an early stage. Even if it happens, it can be treated as early as possible, which is helpful in inhibiting retinopathy from worsening [5].

With the increasing number of diabetes patients, ophthalmologists are burdened with the task of diagnosing patients with DR. It is noteworthy that this task is tedious. The manual diagnosis of DR by ophthalmologists can now be aided by using automated system, thanks to the advancements in signal processing, machine learning algorithms, and computer vision technology. Although detecting and classifying DR are not easy, researchers have paid close attention to it because of the enormous implications for retinal health monitoring. Researchers commonly use retinal fundus images, optical coherence tomography (OCT), and OCT angiography (OCTA) to diagnose DR [6]. However, in this paper, we focus on the use of fundus images.

Currently, deep learning (DL) has been commonly used for image classification in computer vision field. In particular, the most effective model in the field of computer vision is the Convolutional Neural Network (CNN) through Transfer Learning (TL). A CNN is a multi-layered cognition inspired by a biological neural network. Typically, a simple CNN consists of a convolutional layer, a pooling layer, and a full connection layer [7]. When designing the convolution layer, we need to consider parameters such as multiple filters, the size of the convolution kernel, and the step of the sliding window. To minimize the network parameters, a pooling layer is added between the convolution layers to simplify the model. Finally, the full connection layer acts as a classifier in the overall CNN. The feature space receives the original data from the convolution layer and the pooling layer and transfers the learned features through full connection layers into the label space of the sample [8]. Subsequently, the features are used for specific image processing tasks such as detection [9], segmentation [10] and recognition [11].

On the other hand, the success of Extreme Learning Machine (ELM) has somewhat shown that rank-independent random weights at convolution layers can work for certain cases [12], [13]. As a result, the classification layer seems to play an important role in the case of TL. Moreover, ensemble learning may further increase optimization to the entire classification in in comparison to implementation of a single multi-class shallow classifiers.

With the apparent success of DL in general and TL in particular, the following research questions are raised and answered in this paper: Can combining pre-trained features from more than one DL model contribute to better classification in the context of DR? How can the classification be further optimized? Can weighted features or parameter tuning be considered?

The rest of this paper is organized as follows. In Section II, relevant research work and contributions of this paper are elaborated. In Section III, the methodology of the experiments is presented. The results will subsequently be discussed in Section IV. Lastly, we conclude this paper with some remarks and future directions.

## II. LITERATURE REVIEW AND CONTRIBUTIONS

### A. RELATED WORK

In recent years, more and more researchers have used DL to classify DR [14], [15], [16]. However, medical datasets are expensive and, unfortunately, DL training needs a lot of labeled data samples. Note that this is because most lesion sites in medical diagnosis have distinguishable characteristics. Therefore, TL is effective to solve the lack of labeled data for neural networks. It should be noted that TL is a technique for training deep neural networks utilizing a lot of data samples, then changing the trained network to use it for classification task.

In [17], DR occurrence was classified using fundus images with GoogLeNet and ResNet. The testing accuracy for the two-class classification task for GoogleNet and ResNet, respectively, was 97.3% and 96.2%. In addition, the same researchers showed that color constancy preprocessing technique could improve the accuracy of the classification. Researchers in [18] applied Inception-V3 to classify DR into five levels. They applied Kaggle dataset which contained 4000 fundus retina images. The images were resized to 500 pixels to fit into existing pre-trained network architecture input. They obtained an accuracy of 48.8%. Authors in [19] compared the performance of two common CNN models, i.e., Inception-V3 and Xception to detect DR. The dataset contained 35,126 images. They reported an average accuracy of 87.12% for Inception-V3 and 74.49% for Xception. Apart from Retina, the implementation of Deep learning is seen as highly feasible approach in various other medical imaging research domain such malaria, diabetic retinopathy, brain tumor, and tuberculosis. Nevertheless, author also pointed out that despite several machine learning models being available for medical imaging applications, not many have been implemented in the real-world due to the uninterpretable nature of the decisions made by the network.Thus, this is a single most critical aspect that needs to be considered as a tradeoff to the highly efficient deep learning networks in various domains.

In line with the nature of the project, it is worthwhile to investigate some application of deep learning in particular transfer learning [20], [21], [22]. The authors in [23] compared three CNN architectures, i.e., AlexNet, VGG-16, and Inception-V3 to perform five-class DR classification using a total of 166 images from Kaggle. The images were resized into 227 × 227, 224 × 224, and 299 × 299 for AlexNet, VGG-16, and Inception-v3, respectively.

The highest accuracy values were 63.2%, 50.03%, and 37.43% for Inception-V3, VGG-16, and AlexNet, respectively. Using a more shallow architecture, authors in [19] used AlexNet to classify the DR into two classes. A total number of 580 images from Messidor dataset were resized into 227 × 227 to evaluate the network. The reported accuracy was 88.3%.

Related research work has been reported in [24] where the authors used DR images from Kaggle dataset and extracted the patches to obtain important features of the images to train and validate five CNN models (AlexNet, VGG-16, GoogLeNet, ResNet, and Inception-v3). A subset of 243 images from Kaggle were labeled to generate 1,324 image patches containing hemorrhages, microaneurysms, exudates, retinal neovascularization, or normal. On the other hand, images from eOphta dataset were used to test the models. The result showed that the best CNN architecture was Inception-v3 with accuracy of 96% for the five-class classification task and 98% for binary classification.

In [25], the performance of four CNN architectures (AlexNet, VGGNet, GoogLeNet and ResNet) with parameter tuning to classify fundus images for DR detection was compared. The paper reported that GoogLeNet achieved the highest accuracy of 86.35%. Paper [26] used Inception-v3 to detect DR from fundus images. A total of 2500 images from Kaggle were resized to 300 × 300 to be fed into Inception-V3 input. This paper classifies the images into two classes. The reported accuracy of Inception-V3 was 90.9%. In a separate research, [27] classified the fundus images into five classes by using Inception-V3 architecture using a public dataset from Kaggle containing 7,023 images. An image size of 229 × 229 was used in the network. The reported accuracy was 80%. In [28], a blended multimodal deep convolution network was applied for two tasks (detecting DR and predicting the severity level of DR). The authors stated that classifying the severity of DR was a more challenging task. Accuracies between 76% and 80% were acquired on the evaluation of severity level prediction.

In [29], CNN has been used as transfer learning to classify DR using DRIVE dataset. ResNet18 was discovered to have the best performance. In [30], a source-free transfer learning was proposed to handle the problem of number of labelled data. When evaluated for DR classification task, the source-free transfer learning showed a competitive result with an accuracy of 91.2%, a sensitivity of 0.951, and a specificity of 0.858. In [31], transfer learning (Xception) model were used to obtain the deep features of the DR lesions. With the optimized hyperparameters, the classification tasks using KNN and neural network were performed. Numerous studies have presented in-depth reviews of the usage of various machine learning and deep learning algorithms for identifying and classifying DR [32], [33], [34]. The comparisons somehow indicate that there are no superior deep learning models, which is in line with the computational ''no free lunch theorem''. Therefore, the performance may be more relevant to the dataset. As such, improving deep leaning features may be of interest, i.e., how the existing networks can be better optimized.

It can be observed that DL continues to be a desirable option with a variety of factors. Given the present TL methods, the majority of studies have concentrated on the implementation. It makes the applicability of the retrieved features is now better understood, but the TL technique as a whole still has to be optimized. The current research papers have not fully addressed this issue.

### B. MOTIVATION AND CONTRIBUTIONS

Based on the literature review in the preceeding subsection, there are two related tasks: DR detection (binary classification) and DR severity grading (multi-class classification). The second task is more challenging. On the opposite, classifying DR from non-DR has provided more accurate test results for application purposes, with the majority of researchers attaining more than 90% accuracy.

When training is carried out using TL, overfitting frequently occurs. Combining features from a few pre-trained networks has demonstrated effectiveness in addressing this issue [35]. As a result, in this paper, we propose to implement multiple frozen network layers as input to the classifier. In particular, we select four networks in which the frozen layers are used as inputs to the Support Vector Machine (SVM) with Error Correction Output Code (ECOC) ensemble. A stochastic approach is proposed to further optimize the TL and ensemble configuration. Due to the apparent nature of the TL, it can be assumed that not all features extracted will be useful. Thus, weighting the numerical features from the extracted layers can be advantageous.

From the above, the main objective of this research is to investigate the implementation of optimized ECOC ensemble and hybrid features from various deep learning network. This can be beneficial to researchers who wish to hybrid these features for classification for improvement in the future. It can be inferred that the extracted features are highly dependent on the images. However, most of the images are from different sources. Hence, some applications utilize a normalization technique called histogram normalization approach. As an additional investigation, we also consider the constrast limited adaptive histogram equalisation (CLAHE), a type of adaptive histogram equalization that restricts the noise amplification [36].

In summary, the contributions of this paper are listed as follows:

1) Multiple frozen network layers from a few TL networks are used as input features to the ECOC-SVM classifier. ECOC is an efficient ensemble configuration that has been proven to work well in optimizing a multi-class classifier by breaking it down to individual binary classification.
2) As not all features are useful, the features are weighted using a stochastic approach in the form of Adaptive Differential Evolution (ADE). ADE is chosen due to its auto-tuning capability where it can adjust the

**TABLE 1.** Number of samples of APTOS and EyePac + Messidor-2 datasets.

| Dataset | Severity Level (Classes) | | | | | |
|---|---|---|---|---|---|---|
| | Class 0 | Class1 | Class 2 | Class 3 | Class 4 | Total |
| APTOS | 1805 | 370 | 999 | 193 | 295 | 3662 |
| EyePac + Messidor-2 | 360 | 360 | 360 | 360 | 360 | 1800 |

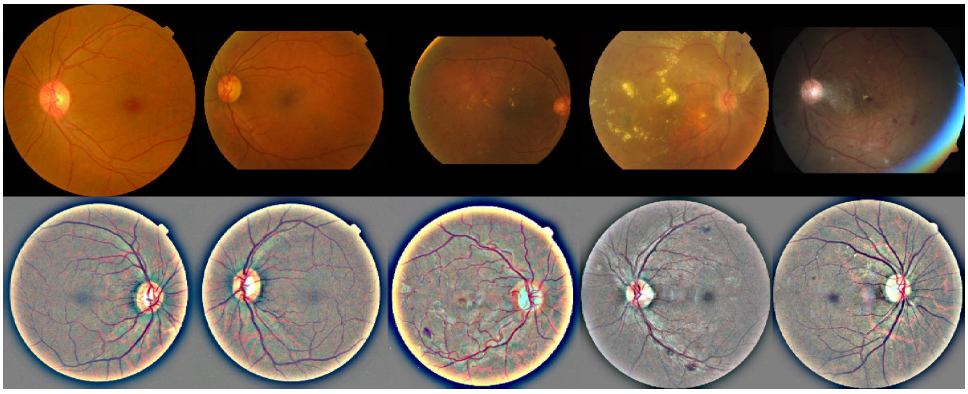


**FIGURE 1.** Sample fundus images (Top: APTOS, bottom: EyePac + Messidor-2).

spectrum between exploration and exploitation when it converges. This research work presents a solution to the long-standing research gap in which feature extracted from TL are not fully optimized. In addition, both features and classifier ensemble are optimised simultaneously.

## III. METHODOLOGY AND PROPOSED APPROACH

### A. DATASET

There are various publicly available fundus datasets that can be used to detect DR. The EyePac, APTOS, Messidor-1, and Messidor-2 are the common datasets used for such work. EyePac dataset provides the highest number of image samples. However, 25% of the images are not gradable by experts. The APTOS dataset, on the other hand, continues to be the one that is most frequently used for benchmarking. However, due to the imbalance class problem, this dataset creates a challenge for multi-class classification applications. For the Messidor-1 and Messidor-2 dataset, the same is true.

Several authors have made an effort to merge these datasets to solve the imbalance class issue. One such attempt is available in Kaggle public domain,[1] where EyePac and Messidor-2 datasets are combined to create a balanced multi-class dataset (EyePac + Messidor-2). In this paper, we consider APTOS and EyePac + Messidor-2 datasets. While APTOS dataset is widely used, the EyePac + Messidor-2 presents a more balanced dataset in each class. As a result, it would be a better indicator to ''balanced'' dataset scenario. It is important to highlight that integrating these two datasets might raise questions because the severity grading definitions can vary widely. The severity of unbalanced class can be visualised from Table 1. Hence, for EyePac + Messidor-2 dataset we divide into three classes: No-DR, mild/moderate DR, and Proliferate DR.

For APTOS dataset, the DR severity can be classified into five classes as listed in Table 2. It is noteworthy that only dataset with similar class number and grading approaches can be combined. In addition, classes are continuous and the grading from one class to another may be different from one dataset to another. Therefore, this need to be taken into account when considering the results. Since, the dataset is an amalgamation of two known or widely used datasets, we believe that there should not be much difficulties in the issue pertaining the validity of the classes grading. Therefore, a safer and more conservative approach is deployed for EyePac + Messidor-2 dataset where only three classes are considered, i.e., Non-Referable DR (class 0), Referable DR (classes 1-3), and Proliferative DR (class 4). For APTOS dataset, we apply both binary and multi-class classification tasks. The binary classification consists of Non-Referable DR (class 0) and Referable DR (classes 1 - 4). Meanwhile, the multi-class classfication task uses all the five classes shown in Table 2. Several sample fundus images from the two datasets are shown in Fig. 1. The grades (classes) in Table 2 can be explained as follows [37]. Class 0 is referred to be the class free of any DR features. The patient has mild Non-Proliferative DR when there is at least one microaneurysm, regardless of whether there are retinal haemorrhages, hard exudates, cotton wool patches, or venous loops. Patient with moderate Non-Proliferative DR has a few cotton wool areas of venous beading, multiple microaneurysms, and retinal haemorrhages. Meanwhile, numerous haemorrhages and microaneurysms in 4 retinal quadrants, venous beading in 2 or

[1] https://www.kaggle.com/datasets/mohammadasimbluemoon/diabeticretinopathy- messidor-eyepac-preprocessed

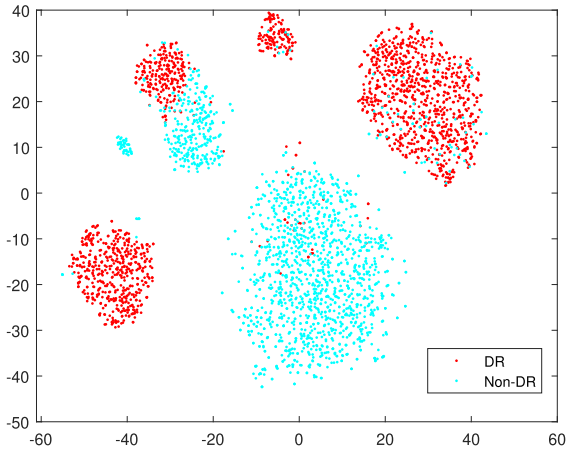



**FIGURE 2.** t-sne plot for APTOS dataset (2 classes).

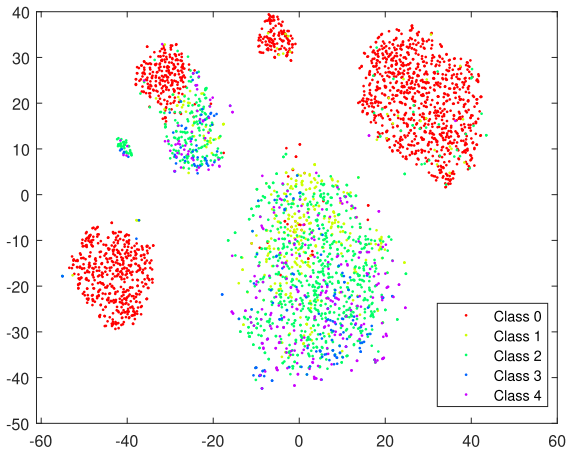



**FIGURE 3.** t-sne plot for APTOS dataset (5 classes).

more quadrants, and intraretinal microvascular irregularities in at least 1 quadrant are all signs of severe Non-Proliferative DR. The last stage of DR, Class 4, exhibits the weak and brittle blood vessel walls that can lead to blood leakage, which can result in blindness and visual loss.

In order to observe the separability of the data, t-sne plot of the features is applied to the APTOS dataset. Figs. 2 and 3 shows the separability of the two-class and five-class separation, respectively. It can be seen that the separability is moderate for the binary classification task but it is more difficult for the multi-class classification task. This justifies the need for applying weighted approach to the raw numerical input.

### B. ADAPTIVE DIFFERENTIAL EVOLUTION (ADE)

The concept of Differential Evolution (DE) has been discussed in many literature [38], [39], [40]. DE algorithm can be listed as follows:

1) For each $i$-th solution vector $\overrightarrow{X}_i^G = [x_{i,1}^G, \cdots, x_{i,j}^G, \cdots, x_{i,N}^G]^T$, select three $N$-dimensional auxiliary vectors $\{\overrightarrow{X}_{r_1}^G, \overrightarrow{X}_{r_2}^G, \overrightarrow{X}_{r_3}^G\}$, where $i = 1, \cdots, P$, $r_1, r_2, r_3 \in \{1, \cdots, P\}$, $i \neq r_1 \neq r_2 \neq r_3$, $P$ is the population size, and $G = 1, \cdots, G_{max}$ is the maximum generation.
2) Form the mutated vector $\overrightarrow{V}_i^G = [v_{i,1}^G, \cdots, v_{i,j}^G, \cdots, v_{i,N}^G]^T$ using
$$\overrightarrow{V}_i^G = \overrightarrow{X}_{r_1}^G + F(\overrightarrow{X}_{r_2}^G - \overrightarrow{X}_{r_3}^G), \tag{1}$$
where $F$ is the differential weight.
3) Generate a trial vector using the mutated vector and the principal parent using
$$u_{i,j}^G = \begin{cases} v_{i,j}^G, & \text{if } rand_{i,j}[0,1] \leq Cr, \\ x_{i,j}^G, & \text{otherwise,} \end{cases} \tag{2}$$
where $Cr$ is the crossover rate and $rand_{i,j}[0,1]$ is a random number drawn from standard uniform distribution for each $j$-th component of the $i$-th vector.
4) Calculate the next generation using
$$\overrightarrow{X}_i^{G+1} = \begin{cases} \overrightarrow{U}_i^G, & \text{if } \mathcal{F}(\overrightarrow{X}_i^G) \geq \mathcal{F}(\overrightarrow{U}_i^G), \\ \overrightarrow{X}_i^G, & \text{if } \mathcal{F}(\overrightarrow{X}_i^G) < \mathcal{F}(\overrightarrow{U}_i^G), \end{cases} \tag{3}$$
where $\mathcal{F}(\cdot)$ is the fitness function to be optimized.

Furthermore, DE is highly sensitive to $Cr$ and $F$. Proper settings of $Cr$ and $F$ are challenging. To solve this problem, an adaptive (i.e., automated) tuning approach was proposed in [41]. The adaptive $F$ and $Cr$ are respectively given by
$$F = \begin{cases} \hat{\alpha} + (1-\hat{\alpha}) \times \sin\left(\frac{\pi \hat{t}}{maxiter} - \frac{\pi}{2}\right), & \text{if } \hat{t} \leq \frac{maxiter}{2}, \\ \hat{\alpha} - (1-\hat{\alpha}) \times \cos\left(\frac{\pi}{2} - \frac{\pi \hat{t}}{maxiter}\right), & \text{otherwise,} \end{cases} \tag{4}$$
$$Cr = \begin{cases} \hat{\beta} + (1-\hat{\beta}) \times \sin\left(\frac{\pi \hat{t}}{maxiter} - \frac{\pi}{2}\right), & \text{if } \hat{t} \leq \frac{maxiter}{2}, \\ \hat{\beta} - (1-\hat{\beta}) \times \cos\left(\frac{\pi}{2} - \frac{\pi \hat{t}}{maxiter}\right), & \text{otherwise,} \end{cases} \tag{5}$$
where $\hat{\alpha}$ and $\hat{\beta}$ are constants, $\hat{t}$ is the generation of iteration, and $maxiter$ is the maximum number of iterations.

**TABLE 2.** Grade and details.

| Grade | Details |
|---|---|
| 0 | No DR |
| 1 | Mild Non-Proliferative DR |
| 2 | Moderate Non-Proliferative DR |
| 3 | Severe Non-Proliferative DR |
| 4 | Proliferative DR |

### C. ECOC-SVM CLASSIFICATION LAYER

ECOC is the state-of-the-art classifier ensemble configuration which is commonly employed for high-dimensional and extremely non-linear data. The concept calculates distance metric for a binary string using the classifier's output. A class is assigned to the data sample that is closest to that class. The Hamming distance and the Euclidean distance are frequently

used for this purpose. In general, any binary classifier can be used in the ECOC configuration. In this paper, we choose SVM as the binary classifier as it is frequently used as the configuration's basic classifier. Furthermore, SVM offers some advantages, such as efficiency compared to Neural Network, better performance in large dataset, and simple training process [42].

To separate two target classes, the individual binary SVM employs a kernel projection and plane. Let us consider $(g_i, y_i)$, where $g_i = \{g_{i,n}\}$ is the $i$-th data instance with $n$ features, $y_i \in [-1, +1]$ is the respective label values, and $i = 1, 2, \cdots, \mathcal{N}$. The hyperplane that separates the data vectors $g_i$ into the label of -1 or +1 is given by $f(g) = w^T g + b$, where $w$ is the weight and $b$ is the bias. Essentially, the goal of plane optimization is to reduce $w^T w$ s.t. $y_i f(g) \geq 1$ to a minimum.

We note that the optimization will not yield much practicality without considering a "soft margin" in optimization during training. The soft margin mechanism considers slack variables during optimization as follows

$$\begin{aligned} \min_{w,b,\{\zeta_i\}} \quad & ||\langle w, w\rangle|| + c\sum_i \zeta_i \\ \text{s.t.} \quad & y_i(f(g_i)) \geq 1 - \zeta_i, \quad \forall i, \\ & \zeta_i \geq 0, \quad \forall i, \end{aligned} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\zeta_i$ are the slack variables, and $c$ is the constraint parameter determines the weightage during optimization of boundary to reduce slack variables. The aspect of SVM that highly determines the effectiveness of hyperplane setting lies in $\lambda$ value by incorporating a kernel that projects the original data into a higher dimension of $\mathcal{N}+1$. This way, we can further optimize non-linear separation among classes. Various transforms can be used to gain higher dimensions.

In particular, $\lambda$ is related to the scalling of the gram matrix as follows

$$G_{j,k} = \lambda \cdot K(x_j, x_k), \tag{7}$$

where $K(x_j, x_k) = x_j' x_k$ is a linear kernel. The linear kernel SVM is a type of SVM that uses a linear function as its kernel. In SVM, the kernel is responsible for transforming the input data into a higher-dimensional feature space, where it becomes easier to separate different classes. The decision boundary in a linear kernel SVM is a hyperplane, which is a flat, multidimensional surface that separates the data points of different classes. The objective of the SVM algorithm is to find the optimal hyperplane that maximally separates the classes while minimizing the classification error.

Let $|m_{k,j}|$ be the absolute value of the $(k, j)$-th element of the coding matrix which is used to calculate the distance from an assigned class. Let $\mathcal{C}$ be the set of class labels, where $\mathcal{C} = \{0, 1, 2, 3, 4\}$, and $|\mathcal{C}|$ be the cardinality of $\mathcal{C}$. The assigned class, $\hat{k}$, therefore, can be expressed as

$$\hat{k} = \min_k \frac{\sum_{j=1}^{L} |m_{k,j}| \mathcal{G}(m_{k,j}, s_j)}{|m_{k,j}|}, \tag{8}$$

where $k = 1, \cdots, |\mathcal{C}|$ is the index of the class, $L$ is the length of the code, which is $\lfloor 10 \log_2 K \rfloor$. In (8), $\mathcal{G}(\cdot, \cdot)$ is the binary loss function given by

$$\mathcal{G}(u, v) = \frac{\max(0, 1 - uv)}{2}, \tag{9}$$

where $\max(0, a)$ returns $a$ when $a \geq 0$ and 0 otherwise.

### D. FITNESS FUNCTIONS

As there is also an issue of imbalanced dataset, we implement a weighted balanced scoring approach. Let

$$\begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,|\mathcal{C}|} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,|\mathcal{C}|} \\ \vdots & \vdots & \vdots & \vdots \\ q_{|\mathcal{C}|,1} & q_{2,2} & \cdots & q_{|\mathcal{C}|,|\mathcal{C}|} \end{bmatrix} \tag{10}$$

be the confusion matrix the system and $q_{i,i}$, where $i = \{1, 2, \cdots, |\mathcal{C}|\}$, be the correct classified samples in $i$-th class. As such, the optimization process will try to optimize the classes equally using the following fitness

$$fitness = \sum_{i=1}^{|\mathcal{C}|} \frac{q_{i,i}}{\sum q_{:,i}} \tag{11}$$

In order to observe the effects on mitigating against unbalanced class, this will be compared against the results acquired using the fitness based on the overall accuracy (conventional) as expressed by

$$fitness = \frac{\sum_{i=1}^{|\mathcal{C}|} q_{i,i}}{\sum_{i=1}^{|\mathcal{C}|} q_{:,:}} \tag{12}$$

where $|\mathcal{C}| \geq 2$.

### E. PROPOSED APPROACH

The overall proposed approach is depicted in Fig. 4. The conventional deep learning block shows the typical approach in which a classification layer receives input from the fully connected layer. In our approach, the average pooling layers are extracted and treated as numerical features for training and evaluation using ECOC-SVM ensemble. The features can be obtained by concatenating more than one frozen average pooling layers. In this paper, we use two pre-trained deep learning networks, i.e., ShuffleNet and ResNet-18. The reason of choosing ShuffleNet and ResNet-18 is because of the computation issues. In particular, the number of feature samples from the concatenated frozen layers of the ShuffleNet and ResNet-18 is $\approx$ 1000 , which is considered as an optimization problem with average complexity. Please keep in mind that architectures other than ShuffleNet and ResNet-18 may be employed subject to the suitability of the parameters (e.g., same input size, etc.).

The entire ECOC-SVM ensemble block is inclusive of the SVM parameters $\lambda$ and $c$. Moreover, the One-vs-One (OVO) configuration is applied in this paper. In this configuration, the entire multi class is separated into binary SVM in

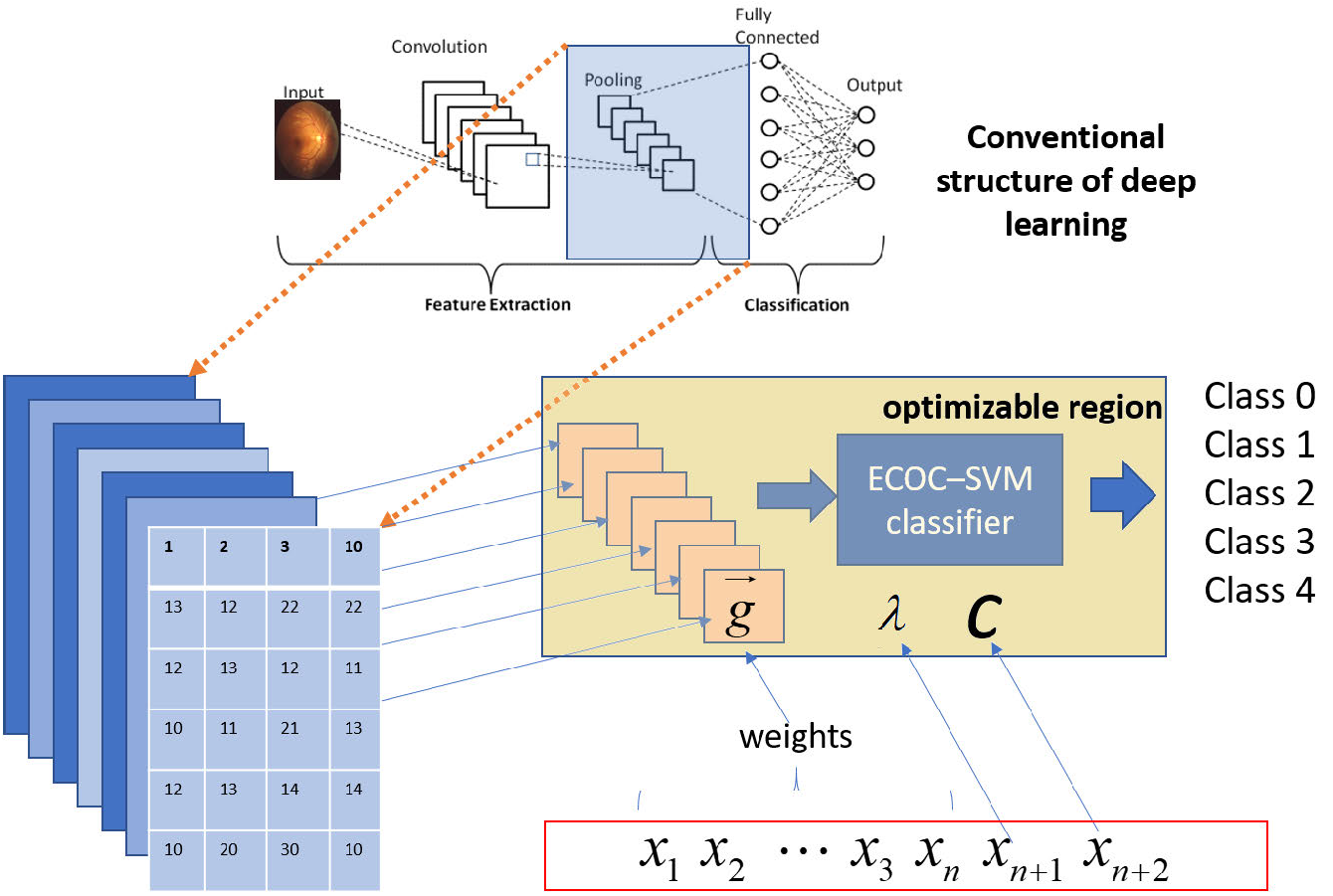



**FIGURE 4.** Proposed approach.

which each distinguishes between one class against a pairing class. Furthermore, the feature weights are optimized using a stochastic optimization (ADE).

By using ADE, the optimal feature weights and SVM parameter values are obtained. The solution of ADE can be expressed by

$$\vec{x} = (x_1, x_2, \ldots, x_n, x_{n+1}, x_{n+2}), \tag{13}$$

where

$$x_{n+1} = \lambda, \tag{14}$$

and

$$x_{n+2} = c \tag{15}$$

Note that $x_1, \cdots, x_n$ refer to the feature weights. The following expression is used to obtain the optimal parameter

$$\mathcal{F} = \min_{\vec{x}}(-fitness), \tag{16}$$

with either (11) or (12) as the *fitness*.

Let $\hat{g}_i$ be the output of the global average pooling from the selected network. Note that $\hat{g}$ refers to concatenated features from the selected networks. The weighted features, $\hat{g}^w$, are acquired based on the product of feature weights and $\hat{g}$. For each instance $i$, the weighted features can be expressed by

$$\hat{g}_i^w = \{x_1\hat{g}_{i,1}, x_2\hat{g}_{i,2}, \cdots, x_n\hat{g}_{i,n}\}. \tag{17}$$

Note that as both $\lambda$ and $c$ are the SVM binary classifier configuration, $\vec{x}$ represents the entire ECOC-SVM configuration inclusive the feature weights of the input from the average pooling layers.

### F. EVALUATION METRICS

Five evaluation metrics, accuracy ($Acc$), precision ($Pr$), recall ($Re$), specificity ($Sp$), and F1-score ($F1$), are used to evaluate the performance of our proposed classification method. The five metrics are defined as follows

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{18}$$

$$Pr = \frac{TP}{TP + FP}, \tag{19}$$

$$Re = \frac{TP}{TP + FN}, \tag{20}$$

$$Sp = \frac{TN}{TN + FP}, \tag{21}$$

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re}, \tag{22}$$

where $TP$, $TN$, $FP$, and $FN$ are the number of true positive, true negative, false positive, and false negative samples, respectively. Note that these definitions are in the case of binary class.

In the case of multi-class classification, two types of metrics can be considered: macro average (weighted evaluation) and micro average metrics. The equivalent metrics for micro metrics are expressed as follows

$$Pr_{mic} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} TP_i + FP_i}, \tag{23}$$

$$Re_{mic} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} TP_i + FN_i}, \tag{24}$$

$$Sp_{mic} = \frac{\sum_{i=1}^{|\mathcal{C}|} TN_i}{\sum_{i=1}^{|\mathcal{C}|} TN_i + FP_i}, \tag{25}$$

$$F1_{mic} = \frac{2 \times Pr_{mic} \times Re_{mic}}{Pr_{mic} + Re_{mic}}, \tag{26}$$

On the other hand, the macro metrics are given by

$$Pr_{mac} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{TP_i}{TP_i + FP_i}, \tag{27}$$

$$Re_{mac} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{TP_i}{TP_i + FN_i}, \tag{28}$$

$$Sp_{mac} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{TN_i}{TN_i + FP_i}, \tag{29}$$

$$F1_{mac} = \frac{2 \times Pr_{mac} \times Re_{mac}}{Pr_{mac} + Re_{mac}}, \tag{30}$$

Note that the accuracy remains the same for both macro and micro statistics. Another definition for $Acc$ is ratio of the sum of all diagonal entries in the confusion matrix to the sum of all entries in confusion matrix. Macro metrics more accurately represent the class-wise statistics whereas the micro metrics represent the overall confusion matrix.

## IV. RESULTS AND DISCUSSION

Two training/test datasets are presented here. the first set involves the implementing extracted transfer learning features without any modification to the images. Then, ECOC-SVM ensemble with linear kernel is implemented and two parameters, $\lambda$ and $c$, are optimized. The second set involves implementing CLAHE pre-processing prior to feature extraction. Similar treatment is performed. In addition, the weighted and non-weighted optimization, as expressed in (11) and (12), respectively, are compared. Note that weighted optimization is used to address the imbalanced dataset issue since simple augmentation approach (random rotation, translation, noise addition) results in severe overfitting.

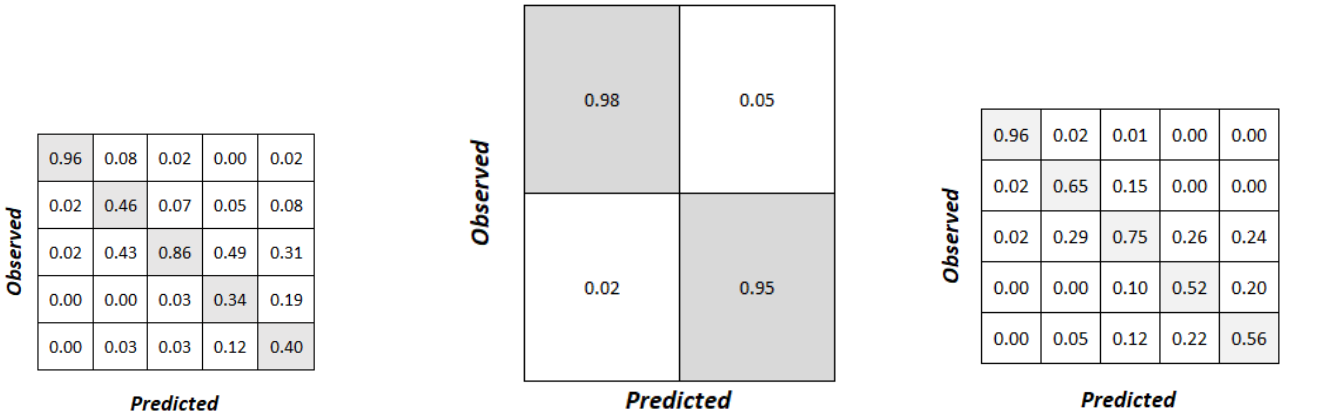
Tables 3 and 4 show the test results evaluated on independent test data with and without CLAHE pre-processing. Note that 3-class evaluation is also added and the reason will be discussed in the subsequent paragraphs. From the generated feature samples, 80% of the features are applied for training/validation while the remaining 20% are reserved for independent testing. The results are based on 10 trial runs. Macro and micro statistics indicate the class-wise and overall F1-score, specificity, sensitivity, precision. The accuracy for both metrics are calculated similarly and there is no distinction between the two metrics. The median and mean indicate the consistency of the results. It is worth mentioning that the optimization process is optimal but it has stochastic nature. As the process is stochastic, various solutions are generated. Nevertheless, the performance of the mean and median indicates that the variations converge to similar performance range. This is commonly known as ''multi-modality'' where there are many optimal solutions, especially when multiple parameters are simultaneously optimized.

The corresponding confusion matrices are depicted in Figs. Figs. 5 and 6. Fig. 5a shows a sample of confusion matrix for 5-class DR classification. It can be seen that multi-class classification task still experiences misclassification in certain classes. This is a common problem attributed to unbalanced class. On the other hand, good separation is shown for binary classification (i.e., DR detection) as depicted in Fig. 5b. Fig. 5c shows the confusion matrix when CLAHE is employed. Some improvements, although not significant, can be noticed.

The CLAHE enhancements are implemented with the weighted approach expressed in (11). From the confusion matrix and table, we can see that the difference is inconclusive from the two configurations (with and without CLAHE). Moreover, the t-test indicates that the p-value between the results is $< 1\%$. Recall that in this implementation, frozen features are extracted for training. Therefore, features express the ''likeness'' towards the extracted objects and would not have a significant effect whether the contrast of the features are adjusted or not.

Regarding the two fitness functions, it can be seen from Table 4, no significant improvement results is observed from the F1-score. The p-value from these trials is 0.2%, indicating that the observed mean difference is not statistically significant. Theoretically, the approach is designed as a means to balanced the accuracy from each class given that the dataset is unbalanced.

Furthermore, when evaluating the mean accuracy for each specific class, only class 3 demonstrates a slight improvement. It seems that the limitation may be related to how the classes are labelled. In particular, classes 1, 2, and 3 (non-proliferative stages) are determined subjectively by evaluating the presence of microaneurysms. As a result, these stages essentially rely on the same features but only differ in the amounts. Notice that misclassifications predominantly occur between adjacent classes. In fact, classes 1-3 are difficult to diagnose [43], even by medical professionals with training. To provide further analysis of the claim, APTOS dataset is restructured into three classes in which class 0 denotes no DR; classes 1, 2, and 3 are now combined




(a) Confusion Matrix for APTOS 5 classes, without CLAHE, fitness: (12)

(b) Confusion Matrix for APTOS 2 classes, without CLAHE

(c) Confusion Matrix for APTOS 5 classes, with CLAHE, fitness: (11)

**FIGURE 5.** Confusion matrix for APTOS dataset using independent test data).


(a) Confusion Matrix for APTOS 3 classes with CLAHE

(b) Confusion Matrix for EyePAC+ messidor-2 3 classes with CLAHE

**FIGURE 6.** Confusion matrix for APTOS and EyePAC + Messidor-2 dataset (using independent test data).

into a single class (non-proliferative DR); and class 4 is the proliferative class. Fig. 6a shows that non-adjacent misclassification rates are reduced.

The EyePac + Messidor-2 dataset has more balanced data samples and has accuracies between 73% and 75%. It can be observed that the F1 score is closer to the accuracy when the data samples are more balanced. As each configuration is run 10 times, the differences between median and mean indicates the variation of the fitness solutions generated. Similar indication can be made by observing the standard deviation. As the research attempts to look into optimization for TF and the ECOC configuration, the iteration vs fitness plot would be a better visual indicator (see Fig. 7). We can observe that the improvement achieved over the iterations. Note that fitness is obtained using validation set. Without weighting the features and optimized the SVM parameters, the accuracy values from independent test data are 78.83% and 95.90% for 5-class and 2-class APTOS dataset, respectively. We can compare with the accuracy values using the best configuration obtained from the optimized ECOC-SVM, which are 82.10% and 96.10%, respectively. It needs to be highlighted that this is based on independent test data. Therefore, it can be stated that classification system has improved in "generalized" evaluation.

Table 5 shows a comparison with various approaches. Some preliminary results involving fine-tuned pre-trained networks are also presented. It can be seen that the proposed approach is relatively comparable with the state-of-the-art approaches. It can be seen that ECOC + TL model performs better than the fine-tuned approaches with single deep learning model. In addition, the implementation of stochastic optimization for optimizing the parameters (our proposed approach) gives further improvement. Finally, bench marking with other research results on similar datasets shows that our approach is on par with state-of-the-art approaches. It is interesting to see that, from Table 5, using the method proposed in [44] could give higher accuracy for the multi-class classification task. It is reasonable as that the work presented in [44] was based on sequential classification, i.e., DR detection followed by DR grading.

As concluding statement with regards to the results section, we would like to emphasize that optimizing both classifier parameters and feature weights has positive effect on the implementation of TL for DR detection and classification.

**TABLE 3.** Macro average metrics.

| Dataset | Stats | Precision | Recall | Specificity | Accuracy | F1-score |
|---|---|---|---|---|---|---|
| EyePac + Messidor-2 (3 classes) | Mean | 0.693 | 0.675 | 0.827 | 0.738 | 0.683 |
| fitness: (12) | Median | 0.695 | 0.678 | 0.828 | 0.740 | 0.686 |
| | Best | 0.707 | 0.692 | 0.833 | 0.750 | 0.696 |
| | Std. Dev. | 0.009 | 0.012 | 0.086 | 0.010 | 0.010 |
| APTOS (5 classes) | Mean | 0.597 | 0.604 | 0.945 | 0.784 | 0.598 |
| fitness: (11) | Median | 0.607 | 0.610 | 0.946 | 0.784 | 0.608 |
| | Best | 0.6267 | 0.646 | 0.952 | 0.806 | 0.633 |
| | Std. Dev. | 0.030 | 0.029 | 0.003 | 0.013 | 0.029 |
| APTOS (5 classes) | Mean | 0.608 | 0.673 | 0.953 | 0.808 | 0.628 |
| fitness: (12) | Median | 0.611 | 0.669 | 0.953 | 0.807 | 0.631 |
| | Best | 0.653 | 0.708 | 0.956 | 0.821 | 0.674 |
| | Std. Dev. | 0.028 | 0.027 | 0.003 | 0.011 | 0.028 |
| APTOS (2 classes) | Mean | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 |
| fitness: (12) | Median | 0.961 | 0.960 | 0.960 | 0.960 | 0.960 |
| | Best | 0.966 | 0.966 | 0.966 | 0.966 | 0.966 |
| | Std. Dev. | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| APTOS with CLAHE (3 classes) | Mean | 0.763 | 0.765 | 0.934 | 0.873 | 0.763 |
| fitness: (11) | Median | 0.765 | 0.765 | 0.934 | 0.870 | 0.765 |
| | Best | 0.781 | 0.791 | 0.942 | 0.886 | 0.781 |
| | Std. Dev. | 0.012 | 0.017 | 0.004 | 0.008 | 0.012 |
| APTOS with CLAHE (5 classes) | Mean | 0.583 | 0.597 | 0.942 | 0.770 | 0.587 |
| fitness: (11) | Median | 0.585 | 0.593 | 0.943 | 0.776 | 0.590 |
| | Best | 0.628 | 0.658 | 0.951 | 0.807 | 0.640 |
| | Std. dev | 0.023 | 0.037 | 0.007 | 0.027 | 0.027 |
| EyePac + Messidor-2 with CLAHE (3 classes) | Mean | 0.742 | 0.742 | 0.859 | 0.733 | 0.741 |
| fitness: (11) | Median | 0.745 | 0.745 | 0.861 | 0.736 | 0.744 |
| | Best | 0.768 | 0.759 | 0.870 | 0.754 | 0.763 |
| | Std dev | 0.017 | 0.014 | 0.008 | 0.015 | 0.015 |

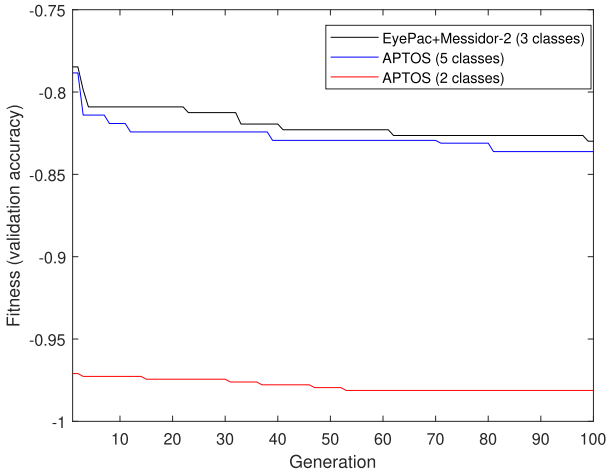



**FIGURE 7.** Fitness vs generation (best individual in the population).

This postulates that the proposed approach is effective for the application in DR detection and grading. The improvement, however, varies depending on the dataset as observed from the Fig. 7. The acquired is from the accuracy of the validation set, hence an approach to prevent overfitting. We highlight that when compared to the ECOC -SVM, there was significant difference as shown from Table. 5. The improvement from fine tuning with conventional Deep learning was significant. The t-test with the fine-tuned DL is $E^{-}10$(close to 0).

We reiterate that the purpose of the research is to assess a TL method optimization. Therefore, despite the slight class imbalance, APTOS dataset implementation is taken into consideration. For the sake of benchmarking, the majority of research studies have used the dataset exactly as it is, without any augmentations. However, a more balanced dataset that combines EyePac and Messidor-2 is believed to be the solution to the class imbalance issue.

To close the discussion, we should note that this work has a few limitations and assumptions that can be carried out as future work. First, we only consider three different fundus image datasets, i.e., APTOS, Messidor-2, and EyePac. Secondly, we only consider ADE as the optimizer for the feature weights. We do not compare ADE with any other algorithms as the goal is to show the effectiveness of TL optimization in DR detection and classification applications, not to compare the optimization algorithms. Thirdly, we only use frozen layers from two TL networks. More TL networks can be used with an increase in complexity.

**TABLE 4.** Micro average metrics.

| | Stats | Precision | Recall | Specificity | Accuracy | F1-score |
|---|---|---|---|---|---|---|
| EyePac + Messidor-2 (3 classes) | Mean | 0.738 | 0.738 | 0.869 | 0.738 | 0.738 |
| fitness: (12) | Median | 0.740 | 0.740 | 0.870 | 0.740 | 0.740 |
| | Best | 0.750 | 0.750 | 0.875 | 0.750 | 0.750 |
| | Std. Dev. | 0.010 | 0.010 | 0.052 | 0.010 | 0.010 |
| APTOS (5 classes ) | Mean | 0.784 | 0.784 | 0.946 | 0.784 | 0.784 |
| fitness : (11) | Median | 0.784 | 0.784 | 0.946 | 0.784 | 0.785 |
| | Best | 0.806 | 0.806 | 0.951 | 0.806 | 0.806 |
| | Std. Dev. | 0.013 | 0.013 | 0.003 | 0.013 | 0.013 |
| APTOS (5 classes ) | Mean | 0.808 | 0.808 | 0.952 | 0.808 | 0.808 |
| fitness: (12) | Median | 0.807 | 0.807 | 0.952 | 0.807 | 0.807 |
| | Best | 0.821 | 0.821 | 0.955 | 0.821 | 0.821 |
| | Std. Dev. | 0.011 | 0.011 | 0.003 | 0.011 | 0.011 |
| APTOS (2 classes) | Mean | 0.961 | 0.961 | 0.961 | 0.961 | 0.961 |
| fitness: (12) | Median | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 |
| | Best | 0.966 | 0.966 | 0.966 | 0.966 | 0.966 |
| | Std. Dev. | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| APTOS with CLAHE (3 classes) | Mean | 0.873 | 0.873 | 0.936 | 0.873 | 0.873 |
| fitness: (11) | Median | 0.870 | 0.870 | 0.935 | 0.870 | 0.870 |
| | Best | 0.886 | 0.886 | 0.943 | 0.886 | 0.886 |
| | Std. Dev. | 0.008 | 0.008 | 0.004 | 0.008 | 0.008 |
| APTOS with CLAHE (5 classes) | Mean | 0.770 | 0.770 | 0.942 | 0.770 | 0.770 |
| fitness: (11) | Median | 0.776 | 0.776 | 0.944 | 0.776 | 0.776 |
| | Best | 0.807 | 0.807 | 0.951 | 0.807 | 0.807 |
| | Std. Dev. | 0.027 | 0.027 | 0.006 | 0.027 | 0.027 |
| EyePac + Messidor-2 with CLAHE (3 classes) | Mean | 0.733 | 0.733 | 0.866 | 0.733 | 0.733 |
| fitness: (11) | Median | 0.736 | 0.736 | 0.868 | 0.736 | 0.736 |
| | Best | 0.754 | 0.754 | 0.877 | 0.754 | 0.754 |
| | Std. Dev. | 0.015 | 0.015 | 0.007 | 0.015 | 0.015 |

**TABLE 5.** Comparison of test results.

| Approach | Dataset | Accuracy (%) |
|---|---|---|
| Modified XCEPTION [45] | APTOS (5-class grading) | 83.09 |
| DenseNet/CBAM [46] | APTOS (5-class grading) | 82.00 |
| DenseNet/CBAM [46] | APTOS (2-class grading) | 97.00 |
| Blended Multi modal Deep ConvNet [28] | APTOS (5-class grading) | 81.79 |
| Combining features with color constancy [44] | APTOS (4-Class grading) | 96.30 |
| Transfer learning-SVM | APTOS (2-class detection) | 98.00 |
| Transfer learning-VGG16 model/Colour processing [47] | APTOS(2-class detection) | 91.23 |
| Fine-tuned AlexNet | APTOS (5-class grading) | 56.34* |
| Fine-tuned ShuffleNet | APTOS (5-class grading) | 66.56* |
| Fine-tuned DenseNet | APTOS (5-class grading) | 68.87* |
| Fine-tuned GoogLeNet | APTOS (5-class grading) | 67.97* |
| GoogLeNet [17] | APTOS (2-class grading) | $97.3^{+}$ |
| ResNet [17] | APTOS (2-class grading) | $96.2^{+}$ |
| ECOC + TL ($c = 1, \lambda = 1$) | APTOS (5-class grading) | 78.83* |
| ECOC + TL ($c = 1, \lambda = 1$) | APTOS (2-class grading) | 95.90* |
| Proposed approach, fitness: (11) | APTOS (5-class grading) | $80.60^{\aleph}$ |
| Proposed approach, fitness: (12) | APTOS (5-class grading) | $82.10^{\aleph}$ |
| Proposed approach | APTOS (2-class grading) | $96.10^{\aleph}$ |

Note: * mean of 10 trials, $^{+}$ no statistical analysis, $^{\aleph}$ best configuration among 10 trials

## V. CONCLUSION AND FUTURE WORK

We have proposed the ECOC-SVM model to classify the level of DR. The features are taken from average pooling layers of TL networks (ShuffleNet and ResNet-18) and weighted. The weights and SVM parameters are optimized using ADE algorithm. This approach has been proposed to avoid overfitting of the conventional fine-tuned TL networks. To evaluate the performance of our model, we use APTOS and EyePac + Messidor-2 datasets. According to the simulation results, our method significantly outperforms the tuned TL networks. Additionally, it has demonstrated comparable results when compared to similar work.

For future exploration, a more comprehensive dataset can be considered. However, we need to consider the unbalanced nature of the dataset. This is complex as the number of data, especially for the severe cases of DR, is limited. DL is, after all, fundamentally a ''black box'' method and they are not completely explainable, the obtained characteristics can only be quantified statistically. There are apparently exhaustive configurations of selected deep learning models that can be evaluated. This would required more exploration. However, it is likely that the results attained would suggest in this report would suggest an improvement in what ever transfer learning model adopted as the optimisation approach would have be able to eliminate non-performing features.

## REFERENCES


[1] T. Y. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, V. C. Lansingh, M. Maia, W. Mathenge, S. Moreker, M. M. K. Muqit, S. Resnikoff, J. Verdaguer, P. Zhao, F. Ferris, L. P. Aiello, and H. R. Taylor, ''Guidelines on diabetic eye care: The International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings,'' *Ophthalmology*, vol. 125, no. 10, pp. 1608–1622, 2018.

[2] M. Z. I. Chowdhury, F. Yeasmin, D. M. Rabi, P. E. Ronksley, and T. C. Turin, ''Predicting the risk of stroke among patients with type 2 diabetes: A systematic review and meta-analysis of C-statistics,'' *BMJ Open*, vol. 9, no. 8, Aug. 2019, Art. no. e025579.

[3] C. C. Wykoff, R. N. Khurana, Q. D. Nguyen, S. P. Kelly, F. Lum, R. Hall, I. M. Abbass, A. M. Abolian, I. Stoilov, T. M. To, and V. Garmo, ''Risk of blindness among patients with diabetes and newly diagnosed diabetic retinopathy,'' *Diabetes Care*, vol. 44, no. 3, pp. 748–756, Mar. 2021.

[4] D. S. Dhoot, K. Baker, N. Saroj, R. Vitti, A. J. Berliner, C. Metzig, D. Thompson, and R. P. Singh, ''Baseline factors affecting changes in diabetic retinopathy severity scale score after intravitreal aflibercept or laser for diabetic macular edema: Post hoc analyses from VISTA and VIVID,'' *Ophthalmology*, vol. 125, no. 1, pp. 51–56, Jan. 2018.

[5] G. H. Bresnick, D. B. Mukamel, J. C. Dickinson, and D. R. Cole, ''A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy,'' *Ophthalmology*, vol. 107, no. 1, pp. 19–24, Jan. 2000.

[6] Z. Sun, D. Yang, Z. Tang, D. S. Ng, and C. Y. Cheung, ''Optical coherence tomography angiography in diabetic retinopathy: An updated review,'' *Eye*, vol. 35, no. 1, pp. 149–161, Jan. 2021.

[7] S. Sharma, S. Maheshwari, and A. Shukla, ''An intelligible deep convolution neural network based approach for classification of diabetic retinopathy,'' *Bio-Algorithms Med-Syst.*, vol. 14, no. 2, p. 45, Jun. 2018.

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, ''Backpropagation applied to handwritten zip code recognition,'' *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[9] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, ''Pedestrian detection with unsupervised multi-stage feature learning,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3626–3633.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, ''Large-scale video classification with convolutional neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[11] A. Toshev and C. Szegedy, ''DeepPose: Human pose estimation via deep neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[12] P. R. Asha and S. Karpagavalli, ''Diabetic retinal exudates detection using machine learning techniques,'' in *Proc. Int. Conf. Adv. Comput. Commun. Syst.*, vol. 2, S. C. Satapathy, A. Govardhan, K. S. Raju, and J. K. Mandal, Eds. Cham, Switzerland: Springer, Jan. 2015, pp. 573–578.

[13] S. Ding, X. Xu, and R. Nie, ''Extreme learning machine and its applications,'' *Neural Comput. Appl.*, vol. 25, nos. 3–4, pp. 549–556, Sep. 2014.

[14] R. Vij and S. Arora, ''A systematic review on diabetic retinopathy detection using deep learning techniques,'' *Arch. Comput. Methods Eng.*, vol. 30, no. 3, pp. 2211–2256, Apr. 2023.

[15] S. Bhandari, S. Pathak, and S. A. Jain, ''A literature review of early-stage diabetic retinopathy detection using deep learning and evolutionary computing techniques,'' *Arch. Comput. Methods Eng.*, vol. 30, no. 2, pp. 799–810, Mar. 2023.

[16] R. R. Maaliw, Z. P. Mabunga, M. R. D. De Veluz, A. S. Alon, A. C. Lagman, M. B. Garcia, L. L. Lacatan, and R. M. Dellosa, ''An enhanced segmentation and deep learning architecture for early diabetic retinopathy detection,'' in *Proc. IEEE 13th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Mar. 2023, pp. 0168–0175.

[17] B. N. Narayanan, M. S. De Silva, R. C. Hardie, N. K. Kueterman, and R. Ali, ''Understanding deep neural network predictions for medical imaging applications,'' 2019, *arXiv:1912.09621*.

[18] S. Masood, T. Luthra, H. Sundriyal, and M. Ahmed, ''Identification of diabetic retinopathy in eye images using transfer learning,'' in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, May 2017, pp. 1183–1187.

[19] S. Mohammadian, A. Karsaz, and Y. M. Roshan, ''Comparative study of fine-tuning of pre-trained convolutional neural networks for diabetic retinopathy screening,'' in *Proc. 24th Nat. 2nd Int. Iranian Conf. Biomed. Eng. (ICBME)*, Nov. 2017, pp. 1–6.

[20] A. M. Mutawa, S. Alnajdi, and S. Sruthi, ''Transfer learning for diabetic retinopathy detection: A study of dataset combination and model performance,'' *Appl. Sci.*, vol. 13, no. 9, p. 5685, May 2023.

[21] R. Vij and S. Arora, ''A novel deep transfer learning based computerized diagnostic systems for multi-class imbalanced diabetic retinopathy severity classification,'' *Multimedia Tools Appl.*, vol. 2023, pp. 1–38, Mar. 2023.

[22] Ö. Kasim, ''Ensemble classification based optimized transfer learning feature method for early stage diagnosis of diabetic retinopathy,'' *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 11337–11348, Aug. 2023.

[23] X. Wang, Y. Lu, Y. Wang, and W.-B. Chen, ''Diabetic retinopathy stage classification using convolutional neural networks,'' in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 465–471.

[24] C. Lam, C. Yu, L. Huang, and D. Rubin, ''Retinal lesion detection with deep learning using image patches,'' *Invest. Ophthalmol. Vis. Sci.*, vol. 59, no. 1, pp. 590–596, 2018.

[25] S. Wan, Y. Liang, and Y. Zhang, ''Deep convolutional neural networks for diabetic retinopathy detection by image classification,'' *Comput. Electr. Eng.*, vol. 72, pp. 274–282, Nov. 2018.

[26] M. T. Hagos and S. Kant, ''Transfer learning based detection of diabetic retinopathy from small dataset,'' 2019, *arXiv:1905.07203*.

[27] I. Kandel and M. Castelli, ''Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review,'' *Appl. Sci.*, vol. 10, no. 6, p. 2021, Mar. 2020.

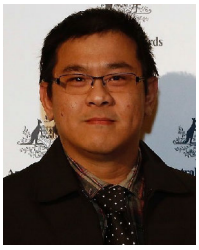
[28] J. D. Bodapati, V. Naralasetti, S. N. Shareef, S. Hakak, M. Bilal, P. K. R. Maddikunta, and O. Jo, ''Blended multi-modal deep ConvNet features for diabetic retinopathy severity prediction,'' *Electronics*, vol. 9, no. 6, p. 914, May 2020.

[29] E. Kaya and I. Saritas, ''Performances of CNN architectures on diabetic retinopathy detection using transfer learning,'' in *Proc. 57th Int. Scientific Conf. Inf., Commun. Energy Syst. Technol. (ICEST)*, Jun. 2022, pp. 1–4.

[30] C. Zhang, T. Lei, and P. Chen, ''Diabetic retinopathy grading by a source-free transfer learning approach,'' *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103423.

[31] N. Shaukat, J. Amin, M. Sharif, F. Azam, S. Kadry, and S. Krishnamoorthy, "Three-dimensional semantic segmentation of diabetic retinopathy lesions and grading using transfer learning," *J. Personalized Med.*, vol. 12, no. 9, p. 1454, Sep. 2022.
[32] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey," *IEEE Access*, vol. 10, pp. 28642–28655, 2022.
[33] V. Lakshminarayanan, H. Kheradfallah, A. Sarkar, and J. J. Balaji, "Automated detection and diagnosis of diabetic retinopathy: A comprehensive survey," *J. Imag.*, vol. 7, no. 9, p. 165, Aug. 2021.
[34] S. Agarwal and A. Bhat, "A survey on recent developments in diabetic retinopathy detection through integration of deep learning," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 17321–17351, Sep. 2022.
[35] W. K. Wong, F. H. Juwono, and C. Apriono, "Vision-based malware detection: A transfer learning approach using optimal ECOC-SVM configuration," *IEEE Access*, vol. 9, pp. 159262–159270, 2021.
[36] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process.-Syst. Signal, Image, Video Technol.*, vol. 38, no. 1, pp. 35–44, Aug. 2004.
[37] O. Faust, R. U. Acharya, E. Y. K. Ng, K.-H. Ng, and J. S. Suri, "Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review," *J. Med. Syst.*, vol. 36, no. 1, pp. 145–157, Feb. 2012.
[38] W. Wong and C. I. Ming, "A review on metaheuristic algorithms: Recent trends, benchmarking and applications," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–5.
[39] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15, no. 1, pp. 4–31, Feb. 2011.
[40] N. Noman, D. Bollegala, and H. Iba, "An adaptive differential evolution algorithm," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2011, pp. 2229–2236.
[41] Z. Huang and Y. Chen, "An improved differential evolution algorithm based on adaptive parameter," *J. Control Sci. Eng.*, vol. 2013, no. 4, pp. 1–5, Jan. 2013.
[42] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, "Advantage and drawback of support vector machine functionality," in *Proc. Int. Conf. Comput., Commun., Control Technol.*, 2014, pp. 63–65.
[43] R. Sarki, S. Michalska, K. Ahmed, H. Wang, and Y. Zhang, "Convolutional neural networks for mild diabetic retinopathy detection: An experimental study," *bioRxiv*, Sep. 2019, Art. no. 763136.
[44] B. N. Narayanan, R. C. Hardie, M. S. De Silva, and N. K. Kueterman, "Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy," *J. Med. Imag.*, vol. 7, no. 3, May 2020, Art. no. 034501.
[45] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Diabetic retinopathy classification using a modified Xception architecture," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2019, pp. 1–6.
[46] M. M. Farag, M. Fouad, and A. T. Abdel-Hamid, "Automatic severity classification of diabetic retinopathy based on DenseNet and convolutional block attention module," *IEEE Access*, vol. 10, pp. 38299–38308, 2022.
[47] Md. R. Islam, M. A. M. Hasan, and A. Sayeed, "Transfer learning based diabetic retinopathy detection with a novel preprocessed layer," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jun. 2020, pp. 888–891.

**W. K. WONG** received the M.Eng. and Ph.D. degrees from Universiti Malaysia Sabah, in 2012 and 2016, respectively. Prior to joining academia, he was with the Telecommunication and Building Services Industry. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Curtin University Malaysia. His research interests include embedded system development, machine learning applications, and image processing.



**FILBERT H. JUWONO** (Senior Member, IEEE) received the B.Eng. degree in electrical engineering and the M.Eng. degree in telecommunication engineering from the University of Indonesia, Depok, Indonesia, in 2007 and 2009, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Western Australia, Perth, WA, Australia, in 2017. He is currently with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University. His research interests include signal processing for communications, wireless communications, power-line communications, smart grids and battery technology, machine learning applications, and biomedical engineering. He was a recipient of the Prestigious Australian Awards Scholarship, in 2012. He serves as an Associate Editor for IEEE Access, a Review Editor for *Frontiers in Signal Processing*, and an Editor-in-Chief for a newly established journal *Green Intelligent Systems and Applications*.

**CATUR APRIONO** (Member, IEEE) received the B.Eng. and M.Eng. degrees in telecommunication engineering from the Department of Electrical Engineering, Universitas Indonesia, Indonesia, in 2009 and 2011, respectively, and the Ph.D. degree in nanovision technology from Shizuoka University, Japan, in 2015. Since 2018, he has been an Assistant Professor of telecommunication engineering with the Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, where he is currently a Lecturer. His research interests include antenna and microwave engineering, terahertz waves technology, and optical communications. He was a member of the IEEE Antenna and Propagation Society (AP-S) and the IEEE Microwave Theory and Technique Society (MTT-S). He involved in the IEEE Joint Chapter MTT/AP Indonesia Section, as a Secretary, and a Treasurer, in 2017, 2018, and 2019, respectively, and also active in various Chapter activities, such as the first Indonesia–Japan workshop on Antennas and Wireless Technology (IJAWT), as a Secretary, and the 2019 IEEE International Conference on Antenna Measurements Applications (CAMA), Bali, in October 2019, as a Treasurer.