Electronic Research Archive

ERA, 31(8): 4843–4861.
DOI: 10.3934/era.2023248
Received: 01 June 2023
Revised: 17 June 2023
Accepted: 19 June 2023
Published: 07 July 2023

http://www.aimspress.com/journal/ERA

Research article

Retinal diseases classification based on hybrid ensemble deep learning and optical coherence tomography images

Kuntha Pin[1,†], Jung Woo Han[2,†] and Yunyoung Nam[3,*]

[1] Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea
[2] Department of Ophthalmology, Bucheon Hospital, Soonchunhyang University College of Medicine, Bucheon 14584, Korea
[3] Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea

* **Correspondence:** Email: ynam@sch.ac.kr; Tel: +821093232233; Fax: +82415301282.

† These two authors contributed equally.

Abstract: Optical coherence tomography (OCT) is a noninvasive, high-resolution imaging technique widely used in clinical practice to depict the structure of the retina. Over the past few decades, ophthalmologists have used OCT to diagnose, monitor, and treat retinal diseases. However, manual analysis of the complicated retinal layers using two colors, black and white, is time consuming. Although ophthalmologists have more experience, their results may be prone to erroneous diagnoses. Therefore, in this study, we propose an automatic method for diagnosing five retinal diseases based on the use of hybrid and ensemble deep learning (DL) methods. DL extracts a thousand constitutional features from images as features for training classifiers. The machine learning method classifies the extracted features and fuses the outputs of the two classifiers to improve classification performance. The distribution probabilities of two classifiers of the same class are aggregated; then, class prediction is made using the class with the highest probability. The limited dataset is resolved by the fine-tuning of classification knowledge and generating augmented images using transfer learning and data augmentation. Multiple DL models and machine learning classifiers are used to access a suitable model and classifier for the OCT images. The proposed method is trained and evaluated using OCT images collected from a hospital and exhibits a classification accuracy of 97.68% (InceptionResNetV2, ensemble: Extreme gradient boosting (XG-Boost) and k-nearest neighbor (k-NN). The experimental results show that our proposed method can improve the OCT classification performance; moreover, in the case of a limited dataset, the proposed method is critical to develop accurate classifications.

Keywords: OCT image; retinal disease; deep learning; machine learning; ensemble classifiers; hybrid machine learning and deep learning

1. Introduction

Common macular and vascular diseases include age-related macular degeneration (ARMD), diabetic macular edema (DME), branch retinal vein occlusion (BRVO), central retinal vein occlusion (CRVO), and central serous chorioretinopathy (CSCR), which are the leading causes of visual impairment and blindness worldwide [1–3]. According to the World Health Organization (WHO), DME, which primarily affects working-age adults, affected 425 million people worldwide in 2017 and is expected to affect 629 million people by 2045 [4]. The WHO also estimates that 196 million people had ARMD in 2020; this number is expected to rise to 288 million by 2040 [5]. The prevalence of ARMD in elderly people is 40% at the age of 70 years, rising to 70% at the age of 80 years. Rogers et al. [6] discovered that BRVO and CRVO affected 13.9 million and 2.5 million of the world's population aged 30 years and older, respectively, in 2008. Men have a higher prevalence of CSCR compared to women [7]. A large population is afflicted by these diseases, and projections suggest that this number will escalate in the future. However, the first stage of these diseases can be treated, and patients can recover their vision loss through early detection and treatment [8–10].

Optical coherence tomography (OCT) is a noninvasive imaging modality that provides high-resolution information within a cross sectional area. OCT retinal imaging enables visualization of the thickness, structure, and detail of various layers of the retina. In addition, when the retina develops a disease, OCT enables the visualization of abnormal features and damaged retinal structures [11]. Therefore, retinal OCT images are widely used in the medical field to monitor information in medical images prior to treatment or for the diagnosis of various diseases.

For several years, ophthalmologists have analyzed the comprehensive information inside the retina for retinal care services, treatment, and diagnosis using retinal OCT images in clinical settings. The clinician performs these tasks manually and wait for each process. As a result, manual analysis is time consuming when there are numerous OCT images. Even if the clinician has great expertise, this analysis may not be accurate [12]. An automated technique based on deep learning (DL) or machine learning using artificial intelligence has been proposed as a solution to overcome this limitation.

Recently, computer algorithms based on artificial intelligence, DL, and machine learning have been proposed for the automatic diagnosis of various retinal diseases and have been applied in clinical health care. Han et al. [13] modified three well known convolutional neural network (CNN) models to gain access to normal and three subtypes of neovascular age-related macular degeneration (nAMD). The classification layers of the original CNN models were replaced by new layers: four fully connected layers and three dropout layers, along with a Leaky rectified linear activation unit (ReLU) as an activation function. The modified models were trained using the transfer learning technique and tested on 920 OCT images; the VGG-16 model achieved an accuracy of 87.4%. Sotoudeh-Paima et al. [14] classified OCT images to identify normal, AMD, and choroidal neovascularization (CNV) using a multiscale CNN. This CNN was evaluated and achieved a classification accuracy of 93.40% on the public dataset. Elaziz et al. [15] developed a four-class classification method for accessing retinal diseases from OCT images based on an ensemble DL model and machine learning. First, the features

are extracted from the two models, MobileNet and DenseNet, and were concatenated as full features of the input images. Then, feature selection was performed to remove irrelevant features and to input the useful features into machine learning to classify the input data. A total of 968 OCT images were used to evaluate classification performance, and an accuracy of 94.31% was achieved. Another study by Liu et al. [16] used a DL model to extract attention features from OCT images. It used the extracted features as guiding features for CNV, DME, drusen, and normal. The classification performance was assessed using public datasets, and an average accuracy of 95.10% was achieved. Minagi et al. [17] used transfer learning with universal adversarial perturbations (UAPs) for classification with a limited dataset. Three types of medical images, including OCT images, were used to assess diseases, and DL models were trained using the ImageNet dataset. The UAPs algorithm was used to generate a training set based on the data provided to train the DL model. There were 11,200 OCT images utilized in training and assessing the model's performance, and a classification accuracy of 95.3% was achieved for the four classes: CNV, DME, drusen, and normal. Tayal et al. [18] presented four ocular disease classifications based on three CNN models using OCT images. Images were enhanced before being fed to CNN models. To assess the performance of the presented method, 6,678 publicly available OCT images were evaluated. The method achieved an accuracy of 96.50% with a CNN model which compressed nine layers. The performance of the CNN models with nine layers outperformed the experimented CNN models with five and seven layers. Adversarial retraining is an algorithm used to improve the performance of DL models based on classification.

According to the literature, retinal OCT classification was developed using DL and DL based methods such as transfer learning, smoothing generative adversarial networks, adversarial retraining, and multi-scale CNN. This method is used to improve the model's performance by fine-tuning previous task knowledge using the OCT image problem, increasing the dataset size for training, applying the technique of inputting data for the training model, and changing the training input image sizes. However, the classification methods can achieve an accuracy of less than 97.00%, indicating their potential for further improvement. Moreover, these studies classify retinal diseases into fewer than five classes. This study aims to improve the classification accuracy and detect five classes of retinal diseases, which are more than the previous studies highlighted in the literature.

In this study, we propose an automatic method based on a hybrid of deep learning and ensemble machine learning for screening five different retinal diseases from OCT images to improve the performance of OCT image classification. The proposed method improves classification accuracy, outperforming standalone classifiers without a hybrid. In addition, it can be trained using a smaller dataset from our hospital, which has been strictly labelled by experts. Moreover, the proposed method enables deployment with a web server for open access to test the evaluation performance within seconds.

2. Materials and methods

2.1. Dataset

All OCT images were collected from Soonchunhyang University's Bucheon Hospital. The OCT images were collected and normalized after approval by the Bucheon Hospital's Institutional Review Board (IRB). OCT images were captured using DRI-OCT (Topcon Medical System, Inc., Oakland,

NJ, USA). The scan range was 3–12 mm in the horizontal and vertical directions, with a lateral resolution of 20 μm and an in-depth resolution of 8 μm. The shooting speed was 100,000 A-scans per second. The OCT images utilized were collected twice; the first comprised 2,000 images that were captured between April and September 2021, while the second consisted of 998 images, and took place over a period of approximately five months from September 2021 to January 2022. Therefore, the total number of OCT images collected twice was 2,998; these were labeled by ophthalmologists for five retinal diseases (ARMD:740, BRVO:450, CRVO:299, CSCR:749, and DME:760) as the ground truth.

This study was approved by the Institutional Review Board (IRB) from Soonchunhyang University Bucheon Hospital, Bucheon, Republic of Korea (IRB approval number: 2021-05-001). All methods were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects.

2.2. OCT image preprocessing

Image processing is a technique for performing various operations on the original images to convert it into a format suitable for DL models or to extract useful features. In image classification based on deep learning, image processing is an essential initial process to change an image before feeding it to the CNN model. The CNN model requires a unique size for the image input, and higher-resolution images demand longer computing times. To shorten the operating time and the suitable size required by the CNN models, all OCT images were downsized to 300 pixels in height and 500 pixels in width. The OCT image dataset was split into an 80% training set and 20% testing set. The training set was used to train the deep learning model and the testing set was used to assess performance.

2.3. Data augmentation

The size of the dataset has a significant impact on the DL performance. Therefore, a larger dataset may enable a better performance. However, in the medical field, most medical dataset has size limits. Data augmentation is a technique developed to overcome the limitations of a dataset by performing different operations on the data provided and creating new data, thereby enhancing the dataset size. Additionally, data augmentation is used to enhance performance [19], generalize the model [20], and avoid overfitting [21]. We utilize data augmentation techniques from the Python library imgaug including like vertical flip, rotation, scale, brightness, saturation, contrast, enhance and contrast, and equalization. The OCT images were augmented at angles of 170, 175, 185, and 190. The selected angle is suitable for rectangle shape representation without loss of information from the original OCT images; scale image with a random range between 0.01 to 0.12; the level of brightness from 1 to 3; the saturate operation, which ranges from 1 to 5, increases by one with each level; random contrast with contrast values ranging from 0.2 to 3; enhance and contrast with levels ranging from 1 to 1.5; and image equalization with levels ranging from 0.9 to 1.4. At the end of the data augmentation process, one OCT image can serve as the basic for generating 29 augmented images. Therefore, the training set comprised a total of 69,455 OCT images, including samples. The acquired OCT and augmented images are shown in Figure 1. Applying data augmentation, only the training set is used for training the proposed method. After finishing the augmentation operation, the OCT images are passed through the 10-fold cross-validation technique to partition the data into folds for the training model (training data) and to test the

model after finishing every epoch (validation data).

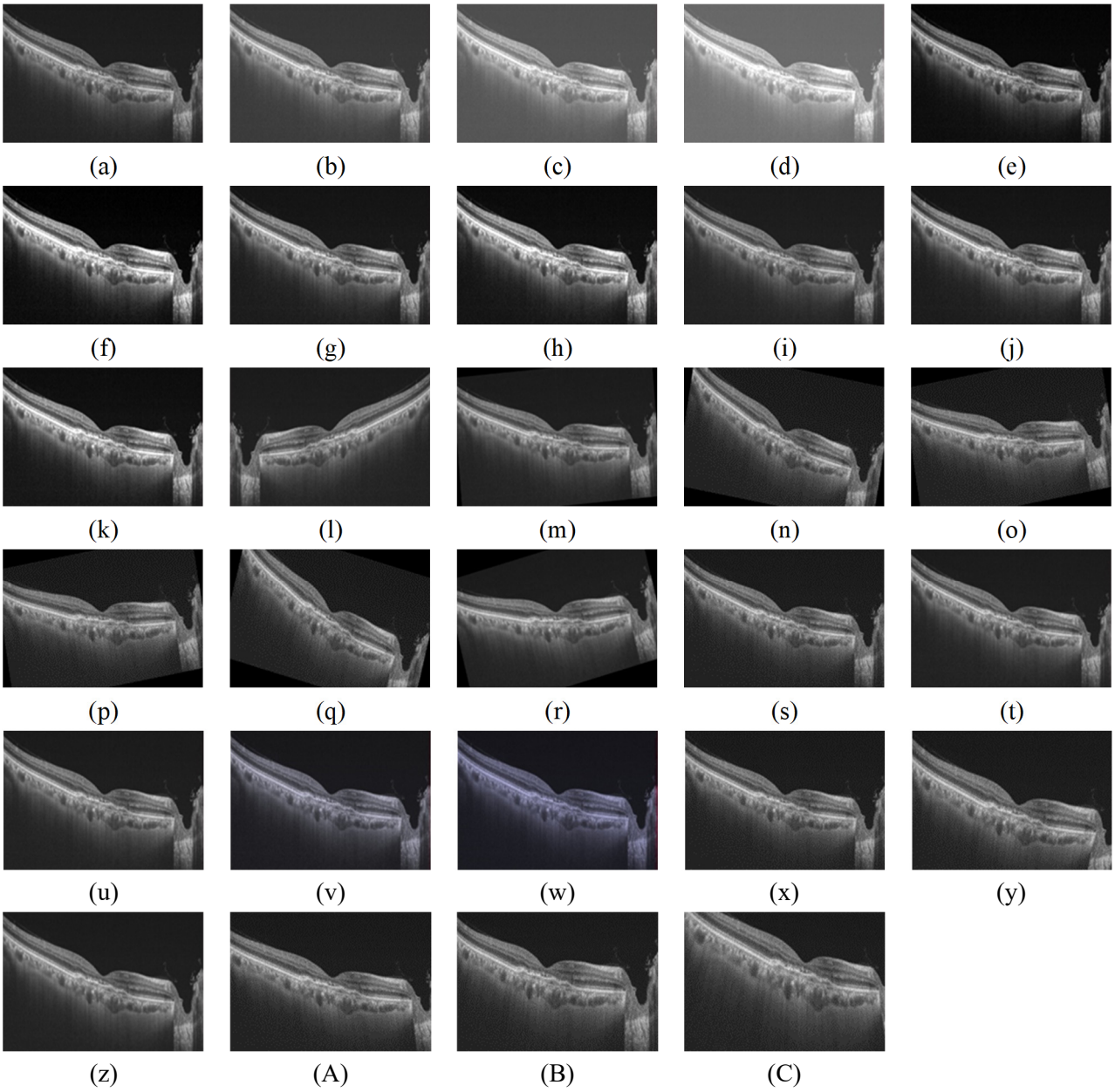

Figure 1. OCT images before and after performing data augmentation. (a) represents the original OCT image. (b), (c), and (d) illustrate brightness adjustments. (e) and (f) demonstrate contrast modifications. (g), (h), and (i) display contrast enhancement. (j) and (k) depict equalization. (l) represents a vertical flip. (m), (n), (o), (p), (q), and (r) indicate angle rotations. (s), (t), (u), (v), and (w) illustrate saturation changes. (x), (y), (z), (A), (B), and (C) represent scaling variations.

2.4. System architecture of proposed method

Figure 2 shows the architecture of the proposed method that comprises three significant blocks: feature extraction, classification, and boosting performance. First, transfer learning based on CNN models extracts one thousand features from the OCT images. Second, various machine learning algorithms are used to classify the OCT images based on the features extracted by the CNN model. Finally, the ensemble algorithm fuses the distribution probabilities of the same class and predicts the retinal disease class based on probability. Each block of the proposed architecture is described in detail in the following subsections.

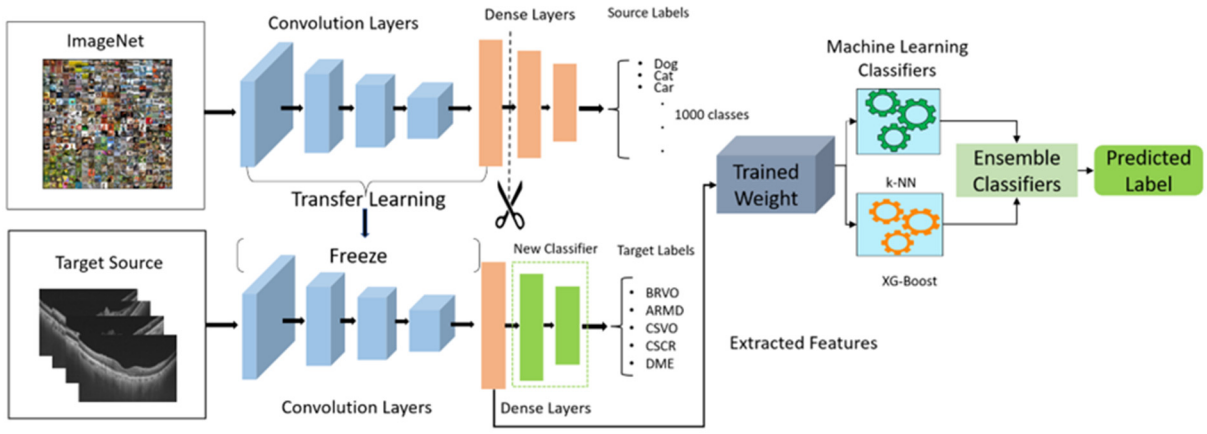

Figure 2. System architecture overview of the proposed method. The proposed method accepts images with resolution of 500 pixels in width and 300 pixels in height. CNN models extract features from OCT images and classify them using machine learning algorithms. Voting classifier ensemble output probabilities for predicting retinal disease.

2.4.1. Feature extraction based on transfer learning

Transfer learning is a technique used to transform the knowledge of a related task that has already been studied to improve the learning of a new task. Training a CNN model from scratch is computationally expensive and time consuming; moreover, an extensive dataset is required to achieve a better performance. Therefore, transfer learning has been developed to overcome DL's drawbacks [22]. To retrain the model with new tasks based on prior knowledge, pretrain was refined, small top layers were trained, and the final layers were frozen. In this study, the transfer learning CNN (TL-CNN) models EfficientNetB0 [23], InceptionResNetV2 [24], InceptionV3 [25], ResNet50 [26], VGG16 [27], and VGG19 [28] are selected and updated. The modification names of the CNN models start with TL, indicating transfer learning, and ends with the original names of the CNN models, including TL-EfficientNetB0, TL -InceptionResNetV2, TL-InceptionV3, TL-ResNet50, TL-VGG16, and TL-VGG19. The original CNN models were created for generic image classification tasks. They were trained and tested on a large dataset (ImageNet) to categorize 1000 different types of images. To use a CNN model with the transfer learning technique and classify retinal OCT images, each CNN model

must modify its classification layers to adapt to the target classes. One specific problem is the categorization of OCT images. The new classification layer is modified with continued stacking of GlobalAveragePooling2D, one Normalization layer, and two Dense layers. The first Dense layer consists of 1,024 with the ReLU activation function and the final dense layer with a five output- vectors. Finally, the updated model is pretrained, and the pretrain model is retrained to fine-tune the previous feature representation in the base model to make it more relevant for OCT image classification. The output consists of five vectors representing the distribution class probabilities using the Softmax activation function. As mentioned previously, a CNN model based on transfer learning is used to extract convolutional features from the OCT images. Therefore, the convolutional features were extracted from the TL-CNNs models where the GlobalAveragePooling2D layers of the classification layer. These features are one-dimensional. Different models provide various features and numbers based on the structure and convolution filters of the model.

2.4.2. Hybrid of deep learning models and machine learning classifiers

Six TL-CNN models independently extracted the features. At the GlobalAverage-Pooling2D layers, the features were extracted (TL-EfficientNetB0: 1,280 features, TL-InceptionResnetV2: 1,536 features, TL-InceptionV3: 2,048 features, TL-ResNet50: 2,048 features, TL-VGG16: 512 features, TL-VGG19: 512 features). Then, the extracted features of each TL-CNN model were used as the input to five popular machine learning classifiers: support vector machine (SVM) [29], k-nearest neighbors (k-NN) [30], decision tree (DT) [31], Random Forests (RF) [32], Naïve Bayes [33], and XGBoost [34]. Various machine learning classifiers use different techniques for learning and distinguishing the different classes of data.

2.4.3. Ensemble voting classifier

Individual machine learning classifiers provide different identification accuracies. This is because each classifier has its own learning ability to identify classes based on the given features. Therefore, an ensemble method is used to aggregate the distribution probabilities of the two classifiers. The proposed method selects two higher prediction classifiers (k-NN and XGBoost) based on an experiment to perform aggregation. An ensemble is a type of soft voting that performs better than other models [35]. Soft voting predicts the final class label as the class label most frequently predicted by classifiers. In soft voting, class labels are predicted by averaging the probability p of the class. Table 1 presents the proposed algorithm, which includes image processing, splitting data, data augmentation, feature extraction, classification, and an ensemble of classifiers:

$$\boldsymbol{y_{FC} = argmax_i \sum_{k=1}^{m} w_k p_{ik}} \tag{1}$$

where w_k is the weight of the machine learning classifiers, which can be either k-NN or XGBoost; it automatically learns from disease features in OCT images and then identifies the type of disease based on the input data; i represents the class label of the retinal diseases, where i $\in$ {0: ARMD, 1: BRVO, 2: CRVO, 3: CSCR, 4: DME}; and p_ik represents the probability of machine-learning weight k for class i.

Table 1. Algorithm of the proposed method.

Algorithm 1: Proposed OCT images Classification
1: p**rocedure** OCT IMAGES PROCESSING
2: **return** preprocessed-images
3: **procedure** SPLIT-DATA (OCT-data)
4: train-data, test-data, train-labels, test-labels = split (OCT-images, labels)
5: p**rocedure** DATA AUGMENTATION (train-data)
6: augmented images = augmentation (vertical flip, rotation, scale, brightness, saturation, contrast, enhance and
7: contrast, and equalization)
8: **return** augmented-images
9: **procedure** 10-FOLD_CROSS_VALIDATION (augmented images, labels)
10: Fold1, Fold2, ……Fold10 = train_test_split(augmented images, labels)
11: **return** Fold1-10
12: **procedure** FEATURE_EXTRACTION (Fold1-10, test-data, test-labels)
13: TL-CNN models = modify the convolutional neural network (CNN) models
14: pre-train the TL-CNN models, small top layers are trained, and the final layers are frozen.
15: extracted features = TL-CNN model at GlobalAveragePooling2D layers
16: **return** extracted features saved in csv format
17: **procedure** CLASSIFICATION (extracted features, labels)
18: classifiers = [svm, k-NN, DT, RF, Naïve-Bayes, and XGBoost]
19: for clsf in range (0,6):
20: predicted-labels = classifiers[clsf]. fit (extracted-features)
21: training-accuracy = accuracy (predicted-labels, labels)
22: save_train_weight
23: voting = “soft”
24: ML1 = k-NN (train-data, train-labels, test-data)
25: ML2 = XGBoost (train-data, train-labels, test-data)
26: **procedure** ENSMEBLE_CLASSIFIERS (train-data, train-labels, test-data)
27: ensemble-classifiers = concadenate(ML1, ML2)
28: ensemble-classifers.fit (train-data, train-labels)
29: predictions = ensemble-classifers.predict(test-data)
30: save_training_weights, results_visualization

3. Experiments

The proposed OCT image classification method was developed using Python 3.7 and TensorFlow 2.6.0. In addition, Scikit Learn was operated on a personal computer running the Windows 10 operating system powered by an Intel(R) Xeon (R) Silver 4114 @ 2.20GHz CPU, 192GB RAM, and an NVIDIA TITAN RTX 119GB GPU.

The proposed OCT image classification method was trained using augmented OCT images and evaluated using a test set. There were two types of training. First, six TL-CNN models were trained to perform feature extraction from OCT images.

Six TL-CNN models were separately trained using a combination of the training set and the augmented images of the training set. The combination data were split using a 10-fold cross-validation algorithm to separate the images for training, validate the model during training, and prevent overfitting. Furthermore, the TL-CNN models were individually trained with a fixed batch size of 64, epochs of size 100, and an Adam optimizer with a learning rate of 0.0001. The learning rate was selected based on the standard learning rate provided by the TensorFlow library. For example, with a setting of 100 epochs, each model must be trained 100 times on the same data. Therefore, the performance is improved by updating the weight based on the information lost through repetitions of a training session. The weights of each TL-CNN model were recorded in a separate file after training and were utilized to extract features from the training and testing sets. Then, the machine learning models were trained with the convolution features extracted by the TL-CNN models to access the class probabilities. Six machine learning models were separately trained, and the weights were recorded after the training completed. Finally, an ensemble method based on soft voting was applied to the average class probabilities of the two classifiers to obtain an effective final class prediction.

4. Results and discussion

The results of the proposed OCT image classification are divided into three parts: classification results, deployment of the classification results to web services, and a comparison of the results with similar studies in terms of classification accuracy.

4.1. Classification

A test set was used to evaluate the performance of the proposed method after training the model. The same preprocessing was performed on both the test dataset and the training dataset without data augmentation. The test set contained 601 OCT images, which were used to assess the classification performance. Six TL-CNN models were individually trained to extract features from the OCT images and store the extracted features in pickle format. Six machine learning classifiers were utilized to discriminate the classes of OCT images based on the features extracted by the TL-CNN. Statistical theories were analyzed to measure the classification ability among the classes, sensitivity, specificity, precision, and accuracy. The relationship between the sensitivity and specificity of various categories was shown through a receiver operating characteristic (ROC) curve. Moreover, the confusion matrix was analyzed, which indicated the correct and incorrect class predictions. Table 2 lists the test results of using TL-EfficientNetB0 as an extractor and seven types of classifiers, including an ensemble classifier, the classification result outperformed the ensemble classifier with a sensitivity, specificity, precision, and accuracy of 96.17, 98.92, 95.89 and 95.85%, respectively. The second highest performance was achieved with the k-NN classifier, which achieved a sensitivity, specificity, precision, and accuracy of 87.37, 96.95, 88.82 and 88.89%, respectively. The classification results for the other machine learning classifiers are unstable, both increasing and decreasing randomly.

Table 2. Shown are OCT images before and after performing data augmentation. (a) represents the original OCT image. (b), (c), and (d) illustrate brightness adjustments. (e) and (f) demonstrate contrast modifications. (g), (h), and (i) display contrast enhancement. (j) and (k) depict equalization. (l) represents a vertical flip. (m), (n), (o), (p), (q), and (r) indicate angle rotations. (s), (t), (u), (v), and (w) illustrate saturation changes. (x), (y), (z), (A), (B), and (C) represent scaling variations.

TL-CNN model	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
TL-EfficientNetB0	SVM	86.79%	96.78%	88.64%	88.39%
	k-NN	87.37%	96.95%	88.82%	88.89%
	DT	85.80%	96.41%	84.95%	86.90%
	RF	66.20%	92.60%	81.08%	75.95%
	Naive Bayes	86.11%	96.40%	86.58%	86.90%
	XGBoost	85.86%	96.45%	86.38%	87.23%
	Ensemble	**96.17%**	**98.92%**	**95.89%**	**95.85%**

Table 3 shows the classification results when using TL-InceptionResnetV2 as an extractor and seven classifiers, showing that the result outperforms the ensemble classifier with a sensitivity, specificity, precision, and accuracy of 97.42, 99.40, 97.49 and 97.68%, respectively. The second highest performance was achieved with the k-NN classifier, with a sensitivity, specificity, precision, and accuracy of 87.37, 96.48, 88.19 and 87.56%, respectively. In addition, with the same extractor, the classification performance of XGBoost was similar to that of the k-NN classifier. Table 4 lists the evaluation results when using the TL-InceptionV3 extractor and seven machine learning classifiers, including the ensemble classifier, which outperformed other methods with a sensitivity, specificity, precision, and accuracy of 91.34, 97.59, 91.03 and 91.04%, respectively. The second highest performance was achieved by XGBoost, with a sensitivity, specificity, precision, and accuracy of 84.42, 95.10, 82.88, and 82.91%, respectively. Table 5 lists the classification results when using the TL-ResNet50 model as a feature extractor and classifying those features by seven different classifiers, which indicates that using ensemble classifiers outperforms the obtained a sensitivity, specificity, precision, and accuracy of 96.46, 99.14, 96.76 and 96.68%, respectively. The second highest performance was achieved by XGBoost, with a sensitivity, specificity, precision, and accuracy of 87.63, 96.59%, 88.27 and 87.73%, respectively. The performances of the other two classifiers, SVM and k-NN, were comparable and better than those of the three classifiers in the experiments. Table 6 lists the test results of the proposed classification with VGG-16 as a feature extractor and seven machine learning classifiers, the ensemble classifier exhibited the best performance, with a sensitivity, specificity, precision, and classification accuracy of 92.07, 98.00, 92.60 and 92.54%, respectively. The XGBoost classifier had the second highest performance for TL-VGG16 as a feature extractor; it obtained a sensitivity, specificity, precision, and accuracy of 80.48, 94.91, 81.44 and 82.26%, respectively. A similar performance was observed for SVM and k-NN. Table 7 lists the classification test results of the TL-VGG19 model for feature extraction and classification using these features by various classifiers. Ensemble classifiers algorithm outperformed the five other classifiers; its sensitivity, specificity, precision, and accuracy are 93.86, 93.40, 93.44 and 93.86%, respectively. The

second- and third-highest performances were achieved by XGBoost and SVM, respectively.

Table 3. Performance summary of proposed classification through feature extraction using TL-InceptionResnetV2, six classifiers, and ensemble voting classifiers. Various sensitivities, specificities, precisions, and accuracies are obtained using different classifiers. The proposed classification method with ensemble classifiers outperforms all statistic measurements.

TL-CNN model	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
TL-InceptionResnetV2	SVM	86.27%	96.13%	86.86%	86.40%
	k-NN	87.37%	96.48%	88.19%	87.56%
	DT	83.77%	95.54%	83.37%	84.41%
	RF	72.93%	93.79%	80.67%	79.27%
	Naive Bayes	79.66%	93.41%	78.25%	77.78%
	XGBoost	87.29%	96.47%	88.05%	87.56%
	Ensemble	**97.42%**	**99.40%**	**97.49%**	**97.68%**

Table 4. Performance summary of proposed classification through feature extraction using TL-InceptionV3, six classifiers, and ensemble voting classifiers. Various sensitivities, specificities, precisions, and accuracies are obtained when using different classifiers. The proposed classification method with ensemble classifiers outperforms all statistic measurements.

TL-CNN models	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
TL-InceptionV3	SVM	82.42%	94.50%	80.85%	81.09%
	k-NN	83.05%	94.61%	80.94%	81.43%
	DT	81.54%	94.18%	79.65%	80.09%
	RF	79.50%	94.01%	80.52%	79.77%
	Naive Bayes	65.72%	86.52%	2.58%	61.33%
	XGBoost	84.42%	95.10%	82.88%	82.91%
	Ensemble	**91.34%**	**97.59%**	**91.03%**	**91.04%**

Six TL-CNN models were compared, and TL-InceptionResNetV2 achieved a better performance than the other five models used in this study, with a sensitivity, specificity, precision, and accuracy of 97.42, 99.40, 97.49 and 97.68 %, respectively. The ensemble algorithm always outperformed all the TL-CNN models. The individual k-NN and XGBoost classifiers performed better than the three individual classifiers. Thus, ensembled k-NN and XGBoost also achieved better performance than k-NN and XGBoost.

Figure 3 shows the ROC result of the proposed classification method, which outperforms TL-InceptionResnetV2 with ensemble classifiers (k-NN and XGBoost). The ROC among each class ARMD, BRVO, CRVO, CSCR, and DME is 0.99, 0.96, 0.99, 0.99, and 0.98, respectively. The relationship between sensitivity and specificity of the five classes is most important. The confusion

matrix is implemented by using the Sklearn library in Python. The size of test data is essential to present the robustness of classification. The confusion matrix shows the number of correct and incorrect predictions among all classes. Figure 4 shows the confusion matrix of the proposed method which exhibited best performance; 148 of 149 OCT images of ARMD class are correctly predicted, 85 of 91 images of BRVO class are correctly predicted (ARMD:3 and DME:3 are incorrect predictions), 59 of 60 images are correctly predicted as CRVO and one image that is incorrectly predicted as BRVO, 148 of 150 images are correctly predicted and two images are incorrectly predicted as ARMD, and 149 of 153 are correctly predicted, and four are incorrectly predicted (ARMD: 1, BRVO:1, and CRVO:2 are incorrect prediction).

Table 5. Performance summary of proposed classification through features extraction using TL-ResNet50, six classifiers, and ensemble voting classifiers. Various sensitivities, specificities, precisions, and accuracies are obtained using different classifiers. The proposed classification method with ensemble classifiers outperforms all statistic measurements.

TL-CNN model	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
	SVM	86.25%	96.02%	85.12%	85.95%
	k-NN	85.75%	96.00%	86.26%	85.74%
	DT	82.04%	94.94%	82.34%	82.59%
TL-ResNet50	RF	52.90%	88063%	71.77%	65.67%
	Naive Bayes	67.71%	89.82%	72.49%	64.68%
	XGBoost	87.63%	96.59%	88.27%	87.73%
	Ensemble	**96.46%**	**99.14%**	**96.76%**	**96.68%**

Table 6. Performance summary of proposed classification through features extraction using TL-VGG16, six classifiers, and ensemble voting classifiers. Various sensitivities, specificities, precisions, and accuracies are obtained using different classifiers. The proposed classification method with ensemble classifiers outperforms all statistic measurements.

TL-CNN model	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
	SVM	76.39%	93.49%	76.96%	78.28%
	k-NN	74.55%	92.33%	75.70%	74.96%
	DT	57.42%	84.86%	55.72%	58.37%
TL-VGG16	RF	50.55%	86.58%	38.39%	63.18%
	Naive Bayes	59.53%	85.08%	58.84%	59.20%
	XGBoost	80.48%	94.91%	81.44%	82.26%
	Ensemble	**92.07%**	**98.00%**	**92.60%**	**92.54%**

Table 7. Performance summary of proposed classification through features extraction using TL-VGG19, six classifiers, and ensemble voting classifiers. Various sensitivities, specificities, precisions, and accuracies are obtained using different classifiers. The proposed classification method with ensemble classifiers outperforms all statistic measurements.

TL-CNN model	Machine Learning	Sensitivity	Specificity	Precision	Accuracy
TL-VGG19	SVM	79.90%	93.74%	78.77%	78.82%
	k-NN	69.16%	90.99%	70.56%	71.64%
	DT	53.23%	82.94%	53.41%	54.73%
	RF	48.29%	85.62%	37.71%	60.36%
	Naive Bayes	56.41%	82.96%	54.70%	54.89%
	XGBoost	82.44%	95.30%	81.90%	83.58%
	Ensemble	**93.86%**	**93.40%**	**93.44%**	**93.86%**

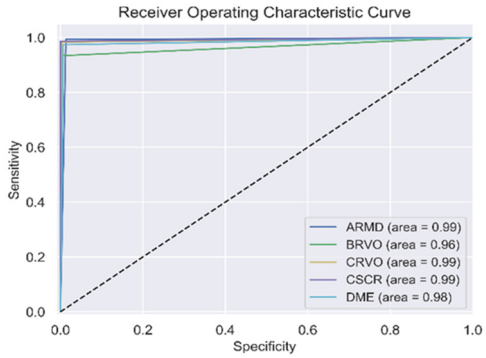

Figure 3. ROC curve of the proposed classification method, which exhibits best accuracy on TL-InceptionResnetV2 model and ensemble classifiers.

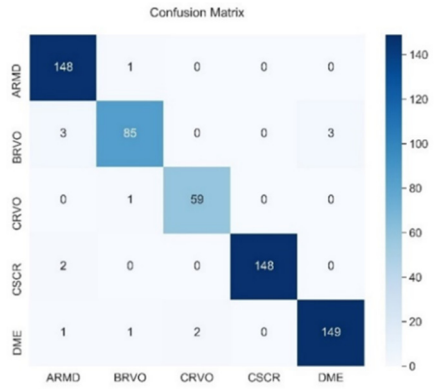

Figure 4. Confusion matrix of the proposed method when it exhibits best performance on TL-InceptionResNetV2 and ensemble classifiers.

4.2. OCT image classification web service

To render the proposed method applicable and accessible from outside through an Internet

connection, we deployed the proposed OCT image classification to a web server using the Flask framework. The web server receives one image input at a time and inputs it into the proposed classification method to predict retinal diseases. The input image is an OCT image consisting of three channels with a resolution of 300 pixels in height and 500 pixels in width. When inputting an OCT image through a web service user interface (UI), the image is transferred to a computer server that runs a DL classification model. First, the computer server performs image processing which is the same to the processes used in both the train and test sets. Second, the preprocessed image is inputted into the proposed classification weights for prediction. Finally, the predicted results are forwarded to the web service using the Flask framework. The prediction results consist of the image input, distribution probabilities among the five classes, the retinal disease diagnosis class, and prediction times of an image. The prediction time is the time taken to input an image to a web service to predict and return the prediction result. Figure 5 shows the initial UI of the web server. The prediction results obtained after inputting the OCT images are shown in Figure 6.

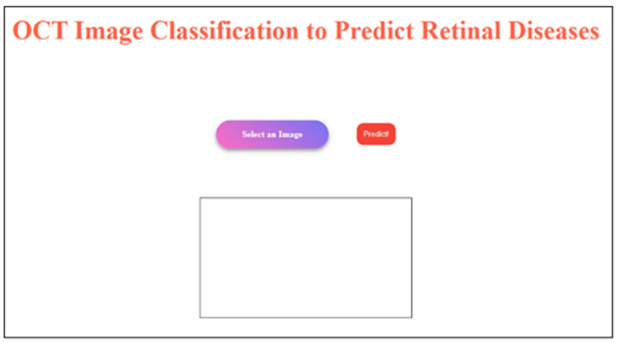

Figure 5. Initial user interface of the developed web service for OCT image classification. The "Select an Image" button allows the user to browse to the location of a stored image and upload it to the webservice, and the "Predict" button sends the image to a deep learning server and receives the diagnosis class.

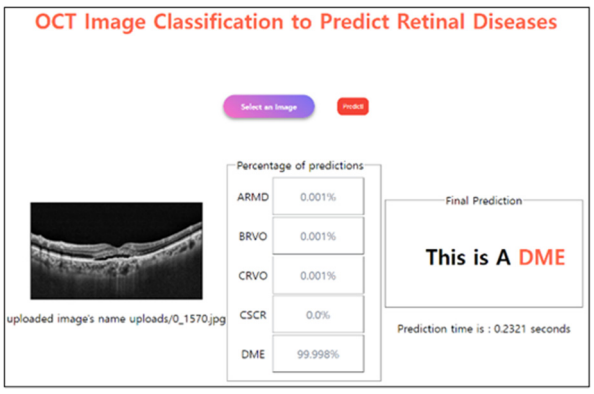

Figure 6. Prediction results from the development web service for OCT image classification. The predicted OCT image, distribution probabilities among five classes of retinal diseases in percent, a final predicted class based on higher probability, and time prediction are represented.

4.3. Comparison results

The higher accuracy of the proposed OCT image classification method is compared with that of the recent studies reviewed in the literature review section, as listed in Table 8. These studies focused on transfer learning, developing new models, and combining well known CNN models with machine learning. All the listed studies used either different OCT databases or a combination of these datasets. Moreover, the number and type of classification classes were different, with at most four classes. We classify retinal diseases into five classes using a dataset obtained from a hospital. An additional number of classes can affect the performance of the classification methods. Table 8 lists the methods and algorithms that have been presented, including the suggested model with transfer learning, the multiscale DL model, and transfer learning using existing CNN models. However, the results as listed in the literature review have shown an accuracy of < 97%. Instead of focusing on a single classifier, this study combines two machine-learning classifiers and the DL as a feature extractor. Our study exhibits an accuracy of 97.68%, which is greater than the accuracy of the aforementioned studies. In addition, the number of classification classes is higher than that of the studies reviewed.

Our study classifies retinal OCT images with disease classes that differ from the reviewed studies and are not available in the public dataset. We hope that these retinal diseases will become available in the future, and we will evaluate the proposed OCT image classification system using a public dataset.

Table 8. Results comparison.

Author	Year	Method	Disease type	Dataset size	Accuracy
Han et al. [13]	2022	Transfer learning with a modification of the well-known CNN models	4-class: PCV, RAP, nAMD, and NORMAL	4749	87.4%
Sotoudeh-Paima et al. [14]	2022	Deep learning: multi-scale convolutional neural network	3-class: AMD, CNV, NORMAL	120,961	93.4%
Elaziz et al. [15]	2022	Ensemble deep learning model for feature extraction, features selection, machine learning as classifier.	4-class: DME, CNV, DRUSEN, and NORMAL	84,484	94.32%
Liu et al. [16]	2022	Deep learning based on method and lesions segmentation model.	4-class: CNV, DME, DRUSEN, and NORMAL	86,134	95.10%
Minagi et al. [17]	2022	Transfer learning with DNN models	4-class: CNV, DME, DRUSEN, and NORMAL	11,200	95.3%
Tayal et al. [18]	2022	Deep learning-based method	4-class: DME, CNV, DRUSEN, NORMAL	84,484	96.5%
Proposed method	–	Hybrid of deep learning and machine learning + ensemble machine learning classifiers.	5-class: ARMD, BRVO, CRVO, CSCR, DME	2,998	**97.68%**

5. Conclusions

This study presents a hybrid ensemble OCT image classification method for the diagnosis of five classes of retinal diseases. The proposed method employs an ensemble machine learning classifier as the classifier and a hybrid deep learning model as the feature extractor. We identified the deep learning model and ensemble classifiers that were most suitable for OCT image classification. The proposed model outperformed an individual classifier. With an accuracy of 97.68%, the best deep learning model and ensemble machine learning classifiers of the proposed method were TL- InceptionResnetV2 and the aggregation of KNN and XGBoost. This classification can be deployed to web services for convenient access to diagnose retinal disease from outside the Internet. Moreover, the prediction time in seconds was short, reducing the time required for prediction. This study contributes to the development of accurate multiclass OCT image classification. In the future, we aim to improve the classification performance. If datasets with the same class as in our study are made public, we will assess the proposed method on these datasets to broaden their applicability. In the medical field, improved performance can be used to automatically classify OCT images and eliminate time-consuming tasks, and this classification can also aid in the prevention of vision loss.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Data availability

The data used to support this study have not been made available access because they are real clinical data from Soonchunhyang Bucheon Hospital, and patient's privacy should be protected, it enables to detect people through this data, but they are available from the corresponding author on reasonable request.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1010362) and the Soonchunhyang University Research Fund.

Conflict of interest

The authors declare no competing interests.

References

1. I. Pearce, A. Clemens, M. H. Brent, L. Lu, R. Gallego-Pinazo, A. M. Minnella, et al., Real-world outcomes with ranibizumab in branch retinal vein occlusion: The prospective, global, luminous study, *PloS One*., **15** (2020), 0234739. https://doi.org/10.1371/journal.pone.0234739

2. F. Tang, F. Xu, H. Zhong, X. Zhao, M. Lv, K. Yang, et al., Comparison of subfoveal choroidal thickness in eyes with crvo and brvo, *BMC Ophthalmol.*, **19** (2019), 1–7. https://doi.org/10.1186/s12886-019-1143-9
3. S. Ţălu, S. D. Nicoara., Malfunction of outer retinal barrier and choroid in the occurrence and progression of diabetic macular edema, *World J. Diabetes*, **12** (2021), 437. https://doi.org/10.4239/wjd.v12.i4.437
4. I. J. Orlando, B. S. Gerendas, S. Riedl, C. Grechenig, A. Breger, M. Ehler, et al., Automated quantification of photoreceptor alteration in macular disease using optical coherence tomography and deep learning, *Sci. Rep.*, **10** (2020), 1–12. https://doi.org/10.1038/s41598-020-62329-9
5. A. F. Borkenstein, E. Borkenstein, Cataract surgery with implantation of a high-add intraocular lens lentis® max ls-313 mf80 in end-stage, age-related macular degeneration: A case report of magnifying surgery, *Clin. Case Rep.*, **7** (2019), 74–78. https://doi.org/10.1002/ccr3.1912
6. S. Rogers, R. L. McIntosh, N. Cheung, L. Lim, J. J. Wang, P. Mitchell, et al., The prevalence of retinal vein occlusion: pooled data from population studies from the United States, Europe, Asia, and Australia, *Ophthalmology*, **117** (2010), 313–319. https://doi.org/10.1016/j.ophtha.2009.07.017
7. F. Xu, S. Liu, Y. Xiang, Z. Lin, C. Li, L. Zhou, et al., Deep learning for detecting subretinal fluid and discerning macular status by fundus images in central serous chorioretinopathy, *Front. Bioeng. Biotechnol*., **9** (2021), 651340. https://doi.org/10.3389/fbioe.2021.651340
8. R. Dandona, L. Dandona, R. K. John, C. A. McCarty, G. N. Rao, Awareness of eye diseases in an urban population in southern india, *Bull. World Health Organ.*, **79** (2001), 96–102.
9. N. Eladawi, M. Elmogy, M. Ghazal, O. Helmy, A. Aboelfetouh, A. Riad, et al., Classification of retinal diseases based on OCT images, *Front. Biosci.-Landmark*, **23** (2018), 247–264. https://doi.org/10.2741/4589
10. N. Rajagopalan, V. Narasimhan, S. K. Vinjimoor, J. Aiyer, Deep CNN framework for retinal disease diagnosis using optical coherence tomography images, *J. Ambient Intell. Human. Comput.*, **12** (2021), 7569–7580. https://doi.org/10.1007/s12652-020-02460-7
11. B. Keerthiveena, S. Esakkirajan, K. Selvakumar, T. Yogesh, Computer-aided diagnosis of retinal diseases using multidomain feature fusion, *Int. J. Imaging Syst. Technol.*, **30** (2020), 367–379. https://doi.org/10.1002/ima.22379
12. T. Kurmann, S. Yu, P. Márquez-Neila, A. Ebneter, M. Zinkernagel, M. R. Munk, et al., Expert-level automated biomarker identification in optical coherence tomography scans, *Sci. Rep.*, **9** (2019), 1–9. https://doi.org/10.1038/s41598-019-49740-7
13. J. Han, S. Choi, J. I. Park, J. S. Hwang, J. M. Han, H. J. Lee, et al., Classifying neovascular age-related macular degeneration with a deep convolutional neural network based on optical coherence tomography images, *Sci. Rep*., **12** (2022), 1–10. https://doi.org/10.1038/s41598-022-05903-7
14. S. Sotoudeh-Paima, A. Jodeiri, F. Hajizadeh, H. Soltanian-Zadeh, Multi-scale convolutional neural network for automated amd classification using retinal OCT images, *Comput. Biol. Med.*, **144** (2022), 105368. https://doi.org/10.1016/j.compbiomed.2022.105368
15. M. A. Elaziz, A. Mabrouk, A. Dahou, S. A. Chelloug, Medical image classification utilizing ensemble learning and levy flight-based honey badger algorithm on 6g-enabled internet of things, *Comput. Intell. Neurosci*., **2022** (2022). https://doi.org/10.1155/2022/5830766

16. X. Liu, Y. Bai, J. Cao, J. Yao, Y. Zhang, M. Wang, Joint disease classification and lesion segmentation via one-stage attention-based convolutional neural network in OCT images, *Biomed. Signal Process. Control*, **71** (2022), 103087. https://doi.org/10.1016/j.bspc.2021.103087
17. A. Minagi, H. Hirano, K. Takemoto, Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning, *J. Imaging*, **8** (2022), 38. https://doi.org/10.3390/jimaging8020038
18. A. Tayal, J. Gupta, A. Solanki, K. Bisht, A. Nayyar, M. Masud, DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases, *Multimedia Syst*., (2021). https://doi.org/10.1007/s00530-021-00791-9
19. P. Cao, X. Li, K. Mao, F. Lu, G. Ning, L. Fang, et al., A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation, *Biomed. Signal Process. Control*, **56** (2020), 101675. https://doi.org/10.1016/j.bspc.2019.101675
20. K. Kong, G. Li, M. Ding, Z. Wu, C. Zhu, B. Ghanem, et al., Robust optimization as data augmentation for large-scale graphs, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 60–69. https://doi.org/10.1109/CVPR52688.2022.00016
21. C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, *J. Big Data*, **8** (2021). https://doi.org/10.1186/s40537-021-00492-0
22. T. Le, M. T. Vo, T. Kieu, E. Hwang, S. Rho, W. S. Baik, Multiple electric energy consumption forecasting using a cluster-based strategy for transfer learning in smart building, *Sensors*, **20** (2020), 2668. https://doi.org/10.3390/s20092668
23. M. O. El Zein, M. M. Soliman, A. K. Elkholy, N. I. Ghali, Transfer learning based model for pneumonia detection in chest x-ray images, *Int. J. Intell. Eng. Syst.*, **14** (2021), 56–66. https://doi.org/10.22266/ijies2021.1031.06
24. T. Kaur, T. K. Gandhi, Deep convolutional neural networks with transfer learning for automated brain image classification, *Mach. Vision Appl*., **31** (2020), 1–16. https://doi.org/10.1007/s00138-020-01069-2
25. S. Elmuogy, A. N. Hikal, E. Hassan, An efficient technique for CT scan images classification of COVID-19, *J. Intell. Fuzzy Syst*., **40** (2021), 5225–5238. https://doi.org/10.3233/JIFS-201985
26. S. A. A. Ismael, A. Mohammed, H. Hefny, An enhanced deep learning approach for brain cancer mri images classification using residual networks, *Artif. Intell. Med.*, **102** (2020), 101779. https://doi.org/10.1016/j.artmed.2019.101779
27. D. Theckedath, R. R. Sedamkar, Detecting affect states using vgg16, resnet50 and se-resnet50 networks, *SN Comput. Sci*., **1** (2020). https://doi.org/10.1007/s42979-020-0114-9
28. M. Bansal, M. Kumar, M. Sachdeva, A. Mittal, Transfer learning for image classification using vgg19: Caltech-101image data set, *J. Ambient Intell. Humanized Comput*., (2021), 1–12. https://doi.org/10.1007/s12652-021-03488-z
29. A. Mechelli, S. Vieira, *Machine learning Methods and Applications to Brain Disorders*, Academic Press, London, 2019.
30. C. Wang, Y. Long, W. Li, W. Dai, S. Xie, Y. Liu, et al., Exploratory study on classification of lung cancer subtypes through a combined k-nearest neighbor classifier in breathomics, *Sci. Rep.*, **10** (2020), 5880. https://doi.org/10.1038/s41598-020-62803-4
31. A. Suresh, R. Udendhran, M. Balamurgan, Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers, *Soft Comput*., **24** (2020), 7947–7953. https://doi.org/10.1007/s00500-019-04066-4

32. A. Subudhi, M. Dash, S. Sabut, Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier, *Biocybern. Biomed. Eng.*, **40** (2020), 277–289. https://doi.org/10.1016/j.bbe.2019.04.004
33. V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. K. Kumar, B. S. Balaji, S. Pandiyan, Skin disease detection and segmentation using dynamic graph cut algorithm and classification through naive bayes classifier, *Measurement*, **163** (2020), 107922. https://doi.org/10.1016/j.measurement.2020.107922
34. Y. X. Liew, N. Hameed, J. Clos, An investigation of xgboost-based algorithm for breast cancer classification, *Mach. Learn. Appl.*, **6** (2021), 100154. https://doi.org/10.1016/j.mlwa.2021.100154
35. S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, *Int. J. Cognit. Comput. Eng.*, **2** (2021), 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001

AIMS Press