Journal of Intelligent & Fuzzy Systems 43 (2022) 4601–4612
DOI:10.3233/JIFS-220066
IOS Press

Artificial intelligence model driven by transfer learning for image-based medical diagnosis

Idowu Paul Okuwobi[a,*], Zhixiang Ding[b], Jifeng Wan[b] and Shuxue Ding[a,*]
[a]*School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China*
[b]*Department of Ophthalmology, Affiliated Hospital of Guilin Medical University, Guilin, China*

Abstract. Artificial intelligent (AI) systems for clinical-decision support are an important tool in clinical routine. It has become a crucial diagnostic tool with adequate reliability and interpretability in disease diagnosis and monitoring. Undoubtedly, these models are faced with insufficient data challenges for training, which often directly determines the model's performance. In order word, insufficient data for model training leads to inefficiency in the model built. To overcome this problem, we propose an AI-driven model by transfer learning in accurate diagnosis for medical decision support. Our approach leverages the shortage of data with a pretrained model by training the neural network with a fraction of the new dataset. For this purpose, we utilized the VGG19 network as the backbone network to support our model in integrating known features with the newly learned features for accurate diagnosis and decision making. Integrating this trained model speeds up the training phase and improve the performance of the proposed model. Experimental results show that the proposed model is effective and efficient in diagnosing different medical diseases. As such, we anticipated that this diagnosis tool will ultimately aid in facilitating early treatment of these treatable diseases, which will improve clinical out-comes.

Keywords: Optical coherence tomography (OCT), choroidal neovascularization (CNV), diabetic macular edema (DME), age-related macular degeneration (AMD), convolutional neural networks (CNN), artificial intelligence (AI)

1. Introduction

Medical imaging is an important tool for clinicians in diagnosing and treating different diseases. It has become one of the most useful resources available to adequately monitor and care for the patients effectively [1]. Medical diagnosis uses various imaging techniques such as; computer tomography (CT), magnetic resonance imaging (MRI), ultrasound, X-ray, nuclear medicine imaging (including positron-emission tomography (PET)), and optical coherence tomography (OCT) [2–5]. Each of these imaging modalities operate technically different to generate the images of the body parts. Ultrasound [6] is a non-invasive medical imaging technique that is widely used in a wide range of applications. It is regarded as one of the safest imaging techniques that utilizes sound waves, rather than ionizing radiation that are commonly used by other imaging techniques. Ultrasound uses conducting gel to transmit the generated high frequency sound waves from the probe to the body. These high-frequency waves bounced back as they hit individual structures in the body, thereby generating a real-time image. Ultrasonic possess minimal risk, as a results of this, it is widely utilized in viewing the kidneys, pelvis, abdomen, fetus in pregnancy, blood vessels, muscles, breasts,

*Corresponding authors. Idowu Paul Okuwobi and Shuxue Ding, School of Artificial Intelligence, Guilin University of Electronic Technology, 1# Jijing Road, Guilin 541004, China. E-mails: paulokuwobi@guet.edu.cn and sding@guet.edu.cn.

ISSN 1064-1246/$35.00

joints, and bones. X-ray imaging is one of the oldest and most frequently used imaging techniques. Unlike Ultrasonic, X-rays imaging uses an electromagnetic radiation. It is mainly used in diagnosing tissues affecting the skeletal systems. In addition, X-rays are also employed in detecting other anatomies, such as digestive issues, breast cancer, etc. Digestive issues can be view through barrow swallows and enemas, while breast cancer is observed through the mammography. Nevertheless, there are known radiation risks associated with the X-rays imaging technique. Like the X-ray imaging, the CT imaging technique [7] adopts the X-rays technology to generates the cross-sectional images of the body. CT imaging uses a large scanner with a circular opening with a motorized table for the patients to lie on during the scanning process. As the patient lies on the motorized table, a source (X-ray) and a detector rotates around the patient, thereby generating a narrow beam of X-rays. These fan-shaped beam of X-rays traverse through the patient body section to produce scans that are combined into a single or multiple images of the body tissues or organs. One of the advantages of the CT imaging technique is its ability to generate a detailed and clearer image of the internal organs, soft tissues, blood vessels, and bones than the conventional X-rays. As such, the benefits of using CT scans far exceed the risks, which include an increased risk of cancer, reaction to the contrast agent or dye used. MRI imaging [8] technique uses magnetic field and radio waves to produce the detailed diagnostic images of the soft tissues, organs, ligaments, cartilage, and bones without the use of any harmful radiation. MRI imaging is widely deployed in examining internal structures of the body, and spinal cord abnormalities, brain damages, prostate, joint abnormalities, liver, breast, injuries, spinal injuries, and other abnormalities. It uses powerful magnetic field to subject the protons to the same field. After which, the radio frequency stimulates the protons and revolve them out of balance. These protons revert to their original position as soon as the frequency is turned off, thereby causing an energy released that is detectable by the MRI sensors. These protons from different tissues part return to their original alignment at different rates, providing the MRI the ability to distinguish between individual tissue and identify abnormality occurrence. The molecules flipping and returning to normal alignment movement and paths are recorded and processed to an image.

OCT is a novel non-invasive imaging technique, that differs from Ultrasound due to its usage of back-reflection of infrared light rather than sound. OCT imaging is equipped with the ability of 2D and 3D mode and its frame rate exceeds that of the video frame rate [9, 10]. Depending on the internal organ or the tissue, the OCT penetration depth could reach about 2 mm [10]. So, OCT imaging technique has substantial advantages over other clinical imaging modalities such as the Ultrasound and MRI, particularly in resolution. OCT imaging is a useful diagnosis tool with great potential in many clinical applications, especially in ophthalmology, arthritis, and cardiology. It has become an important tool in the daily routine of the ophthalmologist in imaging the retina, cornea, optic nerve head, and retinal layers. Due to its ability to accurately measure the retinal nerve fiber layer (RNFL), OCT imaging is now used in neuro-ophthalmology and multiple sclerosis clinical trials. Furthermore, it's a key diagnosis tool in various retinal disorders such as age-related macular degeneration (AMD) (including drusen), diabetic macular edema (DME), choroidal neovascularization (CNV), proliferative diabetic retinopathy (PDR), central serous chorioretinopathy (CSC), [11, 12] etc.

In this work, we utilized the OCT image modality for building and testing our artificial intelligence (AI) model. AI has shown to have great potential in revolutionizing disease monitoring and diagnosis by outperforming human expert in various difficult medical tasks, especially in classification problems. In addition, AI has the capacity to rapidly analyze and classify massive amount of data in a manner and time faster than human experts, showing its efficiency and effectiveness over human experts. Even though, its highly potential, clinical interpretability and feasible preparation of AI is a challenge. Basically, traditional algorithmic technique for image-based classification analysis depends solely on (1) feature extraction based on handcraft approach and object segmentation, (2) segmented objects identification based on specifically designed statistical classifiers for each objects class, and (3) image classification. Optimization of multiple classifiers required many human experts, and it is computationally expensive. As such, a new technique is required. The proposed model addressed the issue with lack of data in a domain, by leveraging data from a similar domain. This technique proved to be highly effective, especially when faced with domains with limited data. Instead of training a completely new neural network, we used a feed-forward technique to fix the weights in the lower levels, which are optimized already to recognize the structures in the images. In addition,

we retrained the weights of the upper levels using back propagation. As such, making the model to recognize discriminate features of a specific category of images. This makes the model operates much faster and with fewer training examples and significantly less computational power. Our contributions include:

(1) We proposed a diagnostic tool based on a trainable AI model for the screening of patients with common treatable blinding retinal diseases.
(2) We proposed a visualization model that assist clinicians in visualizing the diseased region, and a supportive tool in monitoring the detection process.
(3) We proposed a model that is trainable with small dataset and achieve a detection accuracy that is comparable with human experts.

2. Related works

Advancement in the neural networks architectures has led to significant improvement in accuracy, speed, and computational complexity of image-based classification models. With the development of different convolutional neural network (CNN) based deep networks great achievements have been observed in image-based problems. Specifically, the ImageNet challenger results, which is regarded as one of the most significant challenges in the image classification and segmentation field [27]. Kataoka et al. [13] evaluates feature on different deep learning networks for object recognition and detection. The authors performed experimentation on both VGGNet and AlexNet architectures and concluded that VGGNet outperformed the AlexNet architecture. In addition, the authors fine-tuned the features by concatenating some layers of both the architectures and transformed them using principal component analysis (PCA). Using the Caltech101 and Daimler Pedestrian Benchmark datasets, the authors presented an accuracy of 91.8%. Mahmood et al. [14] presented a hybrid model for the image classification process. The authors used the ResNet model for feature extraction and PCA-SVM (support vector machine) to fine-tune the features. They used the Caltech-256, Caltech-101, MLC, and MIT-67 datasets for the experimentation. Their model was trained on 30 images from each class, and experimentation results show that the model outperforms other methods. Ren et al. [15] proposed a mixed approach, using the features acquired through CNN architecture and extreme gradient boost (XGB) classifier. They used MNIST and CIFAR-10 dataset to test their model, and experimental results show that the model is efficient. Srivastava et al. [16] proposed an ensemble deep learning approach based on various pretrained CNNs for feature extraction. The authors combined various feature extraction with scale-invariant feature transform (SIFT). A majority voting scheme using SVM was adopted during the training of the model to detect the image. On the CIFAR-10 dataset, the model achieved an accuracy of 91.8%.

Shaha et al. [17] proposed a technique based on a pretrained VGG19 model that is fused with the SVM classifier. The oretrained model performed the feature extraction and the fused SVM performed the classification process. The authors compared the performance of the model with different models such as VGG16, AlexNet, VGG19 on Caltech256 and GHIM10K datasets. The authors experimental results show that VGG19 performed better than AlexNet and VGG16 architectures. Xin and Wang [18] proposed a comprehensive evaluation of different classification models (SVM, stochastic gradient boosting trees, CNN, random forest, and k-nearest neighbor) for the feature extraction and classification. Pandey et al. [19] proposed a method based on three network models such as region-based (R-CNN), R-FCN (fully convolutional network), and CNN, called common sense knowledge (CSK) for object detection. The experimental results obtained by the authors show that the model is efficient. Singh et al. [20] presented a method using fused handcrafted features. The technique was tested on SIMPLIcity, Soccer, PASCAL VOC2005, SIMPLIcity, Caltech-101, and Flower datasets. They performed a compared their experimental results with the AlexNet, and their model shows better performance than AlexNet. Yadav et al. [21] performed a comparative evaluation of different CNN models performance based on Inception and VGG16 models with the traditional approach of image classification. The traditional methods used for this analysis includes oriented fast and rotated brief (ORB) and SVM. The authors utilized the popular transfer learning approach to improve the accuracy of the classification process. Based on the authors report, the transfer learning-based technique achieved the better results.

Karthikeyan et al. [22] demonstrated the suitability of the pretrained models such as RestNet101, VGG19, and VGG16 using the transfer learning technique on a large X-ray images dataset from patients with COVID-19, pneumonia, and healthy

patients. These models are efficient and effective in the classification process, with an accuracy higher than conventional approaches. Talaat et al. [23] proposed an approach that uses swarm-based algorithm for features selection and a CNN based approach that perform automated feature extraction. This hybrid approach performed better than the conventional methods. Liu et al. [24] proposed a CNN model that is integrated with both global and local feature pyramids fir multiclass pest detection. This model performs the operation in two stages (pest detection and classification). The pest detection phase relies on the global activation feature pyramid network (GaFPN) to aggregate each convolutional block for screening and activating depth and spatial information from feature maps produced by each block. While the classification phase utilizes the feature map generated in the pest detection phase to design a local activated feature pyramid network (LaFPN) that was adopted for pest classification and position regression. In addition, the authors used fully connected layers to localize and classify the images. Experimental results based on the authors in-house pest dataset show that the model is better than Faster R-CNN and FPN. Kumar et al. [25] proposed a comprehensive analysis of deep learning and classical machine learning classifiers based on performance variations with different feature vector representations. The authors developed an ensemble method based on deep learning and classical machine learning methods for the classification problem. The authors aim at improving the performance of single models through the experimentation. Seemendra et al. [26] proposed an approach using various pretrained CNN such as VGG16, VGG19, ResNet, DenseNet, MobileNet and EfficientNet models to detect and classify invasion ductal carcinoma. The results presented by the authors show that VGG19 model achieved the best result.

3. Material and methods

3.1. Datasets

We used the dataset provided by Kermany et al. [27] for both the training and testing of the proposed model. The training dataset consists of a total of 108,309 OCT images (37,205 with choroidal neovascularization, 11,348 with diabetic macular edema, 8,616 with drusen, and 51,140 normal), and the testing dataset consist of a total of 1,000 OCT images (250 with choroidal neovascularization, 250 with diabetic macular edema, 250 with drusen, and 250 normal). The dataset consists of 4 classes namely - CNV, DME, DRUSEN and NORMAL. In Fig. 1, we illustrate the four classes of the images. We created a subset dataset from the dataset due to significant amount of wrongly captured images in the dataset. The model was trained with a total of 46,000 OCT images (20,000 with CVN, 5,000 with DME, 6,000 with DRUSEN, and 15,000 NORMAL), and tested with 968 OCT images (242 from each category) from 633 patients.

3.2. Environmental setup

3.2.1. Hardware

We train and test the proposed model on Intel (R) Core™ i7 CPU @ 2.90GHz with a NVIDIA GeForce RTX 3090 GPU. Using a GPU instead of CPU shows significant acceleration in the training and testing phase of the process.

3.2.2. Software

Python is used for the programming of the proposed model for both training and testing phase. Furthermore, Adam optimizer is utilized for training the neural layers of the proposed network. The

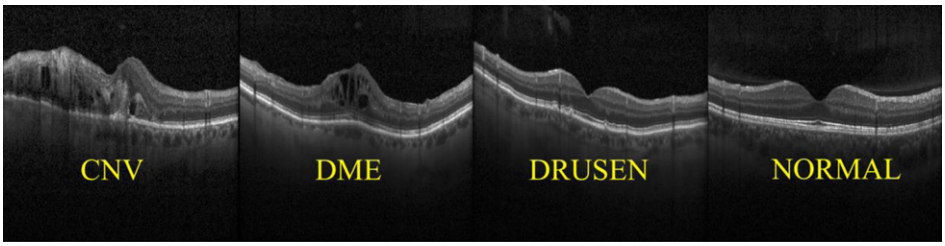

Fig. 1. The four classes of the image in the dataset.

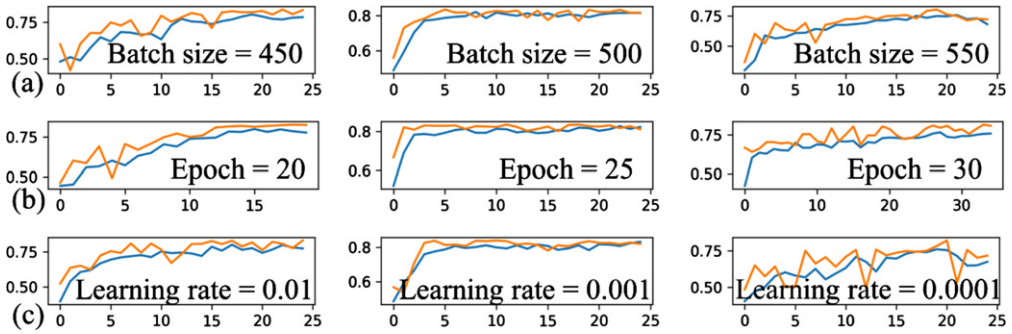

Fig. 2. Comparison of different hyper-parameter for the training of the model. (a) Varying batch size at fixed epoch and learning rate, (b) Varying epoch at fixed batch size and learning rate, and (c) Varying learning rate at fixed epoch and batch size. (At fixed values, batch size = 500, epoch=25, and learning rate=0.001).

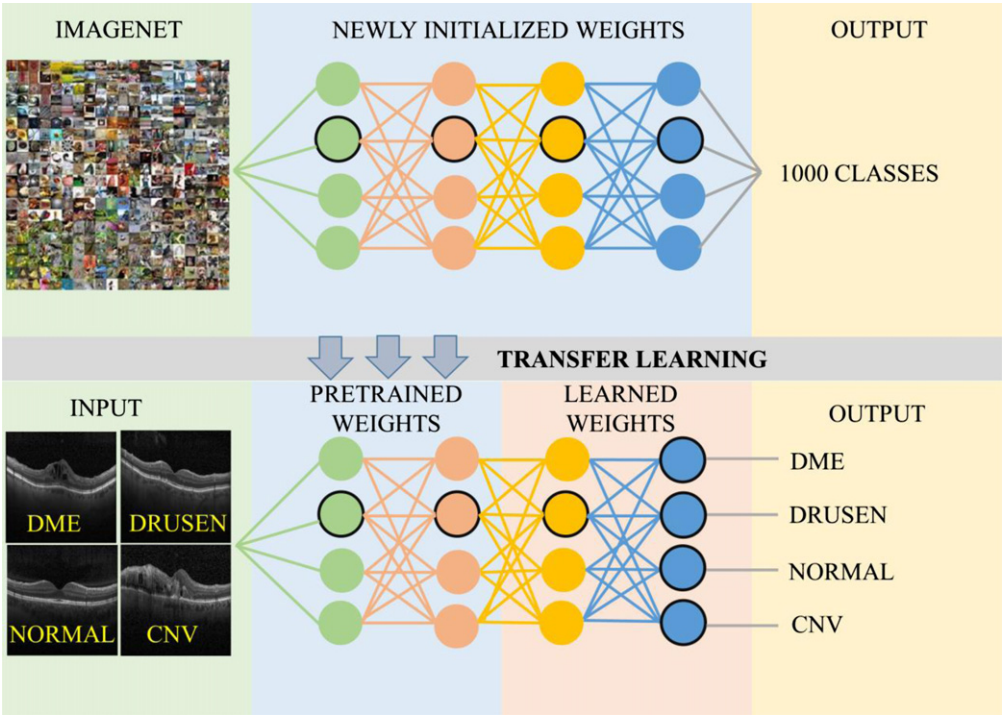

Fig. 3. The architecture of the proposed AI model for image-based medical diagnosing.

training phase required approximately two hour and testing phase required few seconds. The proposed model was configured as follows: batch size of 500, epoch of 25, and learning rate of 0.001.

In Fig. 2, we present the comparison made before fixing the hyper-parameter values for our experimentation. We performed experimentation on some of the OCT data that were not part of the training or testing dataset reported in this work but were used for fine tuning the proposed model to ascertain optimal values for the hyper-parameters. As shown in Fig. 2, we observed that batch size = 500, epoch=25, and learning rate=0.001 obtained the highest accuracy.

3.3. Methods

Figure 3 shows the proposed model architecture for medical diagnosis based on image analysis. Data augmentation technique is used to compensate for the wrongly capture data in the dataset. Figure 4 shows some of the images in the dataset that were wrongly captured, which without data augmentation

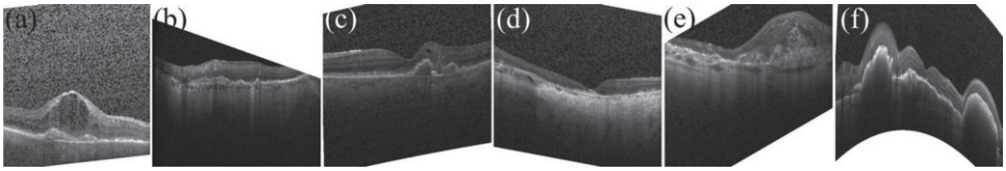

Fig. 4. Varying wrongly captured images in the dataset.

would be less useful in training the model. As such, we performed (a) position augmentation and (b) color augmentation before training the model. For the position augmentation, we performed (1) scaling for such image as Fig. 4a, (2) cropping (Fig. 4b), (3) flipping (Fig. 4c and d), (4) rotation (Fig. 4e), and (5) translation (Fig. 4f). Furthermore, we performed brightness and contrast modification using the color augmentation technique. Data augmentation is utilized on the data to improve the models' accuracy and provides the proposed model with a better generalization capability.

We freeze the lower layers of the pretrained VGG19 model weights, excluding the last 5 upper layers. We performed backpropagation to update the weights of the last 5 layers. This technique provides the proposed model with adequate transfer learned knowledge of low-level features that are generally applicable among all images. Furthermore, the applied technique retained the extracted general image features such as lines, edges, and curves which are universal and relevant for the proposed model. In addition, we modified the upper layers of the VGG19 network to a custom classifier that output 4 classes based on our data. While we kept the weights of the lower layers intact, we control the network to focus mainly on learning the new dataset (OCT images) features in the successive upper layers. The custom classifier and subsequent layers are retrained through backpropagation, while the remaining layers were freeze. One of the advantages of backpropagation is to propagates error, while updating the hidden layers to reduce such errors through gradient descent approach. This error occurs because of differences between the actual and target outputs prediction of the model. In this work, we utilized a transfer learning approach on a network trained on a totally different group of datasets with different task to solve our own task based on the general low-level features shared between images.

Through the transfer learning approach, we successfully customized the upper layers of the proposed model with the specific features of the new dataset (OCT images). Moreso, to compensate for the small size of the OCT image dataset, we freeze the weights of the lower layers since they have been previously trained on a large dataset. These new learned weights by the upper layers send the new features to the dense layers for adequate learning of the OCT image dataset. A newly designed SoftMax layer is built to replace the VGG19 SoftMax activation for the output layer. The new output layer reduces the classification category of the VGG19 output layer (1000 categories) to 4 categories for the purpose of the 4 retinal diseases used in this work.

4. Results

4.1. Evaluation metrics

(a) Accuracy is the degree at which the result of a measurement conforms to true value. It can be obtained as:

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

(a) where TP=True positive; TN=True negative; FP=False positive; and FN=False negative.

(b) Precision is the degree of closeness or dispersity of measurement to each other. It is expressed as:

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

(c) Recall is the degree of measurement at which a model correctly identifies the true positives. It is defined as:

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

(d) F_1 score is the degree of measurement of a test accuracy. It is expressed as:

$$F_1\ score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \quad (4)$$

(e) Cohen Kappa score (κ) is the degree of measurement of intra-rater and inter-rater reliability for a test item [28].

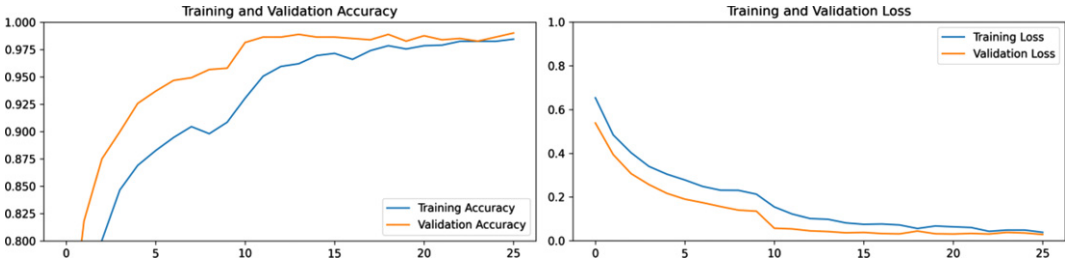

Fig. 5. The training accuracy and loss against epoch of the proposed model. Accuracy vs Epoch (left), and Loss vs Epoch (right).

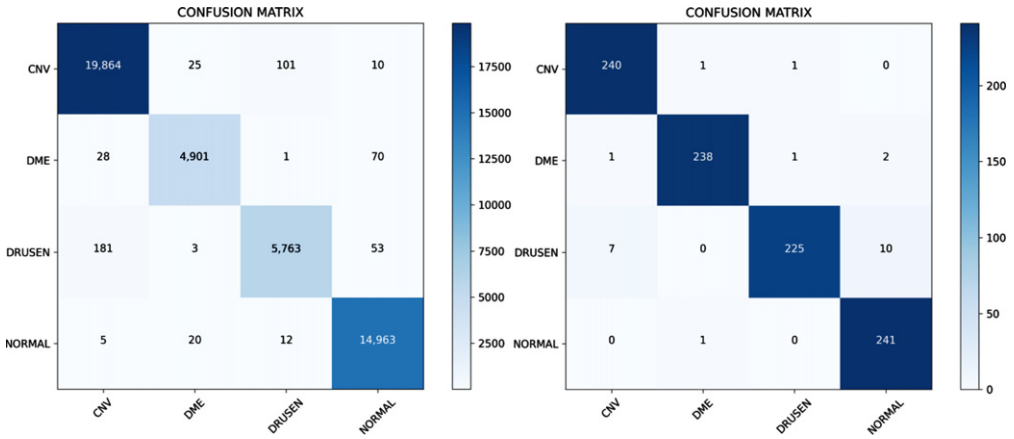

Fig. 6. The confusion matrix for the training and testing phase of the proposed model.

4.2. Experimental results

We evaluated the proposed model in diagnosing the leading causes of blindness diseases using key distinguishable features of each of the diseases. To obtain better classification results, these new features obtained from the learned weights were utilized for training the model, and pretrained weights were incorporated to provide adequate universal features for the proposed model. Figure 5 shows the accuracy and loss function of the training phase of the model. The training phase of the proposed model achieve a commendable accuracy as shown in the confusion matrix in Fig. 6 (right). Here, it is clearly visible that CNV and DRUSEN possess a big challenge for the proposed model due to the feature similarity of the diseases. In addition, both DME and DRUSEN at early stage are difficult to differentiate with NORMAL as shown in Fig. 6. In Fig. 6 (right), the proposed model was tested on 968 OCT images and the confusion matrix shows the detail of the proposed model performances in each of the category of diseases. It could be observed that the proposed model could effectively differentiate between early stage DRUSEN and CNV, which was a challenge during the training phase. Unfortunately, some of the DRUSEN are still misclassified as CNV due largely to the feature similarity between late DRUSEN and CNV. More so, some of the early stage DME were misclassified as NORMAL with significant reduction in quantity between the training and testing phase.

For the multi-class comparison between all the diseases (CNV, DME, DRUSEN, and NORMAL) the results are presented in Tables 1 and 2.

In Table 1, we presented the TP, TN, FP, and FN for each of the class (CNV, DME, DRUSEN, and NORMAL). We observed that NORMAL has the highest TP and lowest FP, followed by CNV, DME, and DRUSEN. In addition, both the DME and DRUSEN

Table 1
Experimental results of the proposed model based on the testing phase experimentations

Metrics	CNV	DME	DRUSEN	NORMAL
TP	240	238	225	241
TN	718	724	724	714
FP	2	4	17	1
FN	8	2	2	12

Table 2
Multi-class evaluation results of the proposed model based on the testing phase experimentations

	PRECISION	RECALL	F_1 SCORE
CNV	0.99	0.97	0.98
DME	0.98	0.99	0.99
DRUSEN	0.93	0.99	0.98
NORMAL	1.00	0.93	0.96
Accuracy			0.98
Macro Avg.	0.98	0.97	0.98
Weighted Avg.	0.96	0.97	0.94
Cohen's Score (κ)			0.97

have the lowest FN, followed by the CNV and NORMAL. In Table 2, we presented for each of the class the precision, recall, and F_1 score, and the final accuracy and Cohen Kappa Score for the testing phase of the proposed model. The proposed model achieved a precision of 99% for CNV, 98% for DME, 93% for DRUSEN, and 100% for NORMAL. Also, a recall of 97% for CNV, 99% for DME, 99% for DRUSEN, and 93% for NORMAL. A F_1 score of 98% for CNV, 99% for DME, 98% for DRUSEN, and 96% for NORMAL. Furthermore, we obtained an accuracy of 98% and Cohen's Score (κ) of 97% during the testing phase. This indicates that the model effectively learned the features of the new datasets and efficiently incorporates the pretrained model intuitive knowledge for supportive decision making in classifying each class of the diseases. These results show an improvement in the model capacity in using the rich features learned during the training phase to classify these diseases accurately. Both the accuracy, precision, and recall show that the model accurately classified the images, while the F_1 score show the model's performance, by indicating the performance of the model based on each image class. The high κ indicates that the model classification accuracy is almost perfect since it is almost close to the ground truth.

5. Discussion

In this work, we implemented an AI-driven model for the diagnosis and referral of leading causes of vision blindness (CNV and DME). To compensate for the shortage of datasets, we employed a transfer learning technique to support our model in learning the new features for accurate decision making in the classification process. The competitive performance of the proposed model through experimentations shows that, despite the shortage of dataset for training the model, the model was able to learn the OCT image features adequately and efficiently classified the images. We achieved this using the transfer learning technique that transfers knowledge learned from a task to boost the accuracy of a related task, thereby minimizing the training data required. With this approach, the proposed model can effectively predict and identifies different retinal diseases using the knowledge learned from other unrelated datasets. In this work, we show that an AI-driven model is comparable to human experts (with significant clinical experience), with the high degree of correctness at which the model identifies the different diseases. In addition, we demonstrated the power of the transfer learning approach to make highly effective and precise classifications amidst very limited training dataset. The results of our work evidence that AI model has the potential to effectively learn and possess the capacity to transfer learned knowledge from an increasing complicated data with a considerable degree of generalization using comparatively small dataset. As such, AI presents an effective model that requires further exploration to support clinicians in their daily routine. In addition, it also presents an automated application to assist in diagnosis such as in disease detection [29, 30], image-based disease enhancement [31, 32], and classification and prevalent disease severity grading [33].

Experimentation results show that the proposed model still faces challenges in some difficult situation in differentiating between some of the disease classes, such as in these 3 instances: (1) early DRUSEN and NORMAL, (2) late DRUSEN and CNV, and (3) early DME and NORMAL. In Fig. 7, we show the images of the four classes with each class possessing different features at different stages but exhibiting similar features with other image class. Figure 7a (row 1-4) shows the four image classes with close similarity in features at each stage of the disease. Figure 7a (row 1) shows all the image class exhibiting close similarity with normal image, without any tangible feature to distinguish each image class from the normal image. Figure 7b (row 2) shows all the image class with features and physical appearance like drusen image, which is also relatively difficult for experience expert to differentiate if patients' history is not available. Figure 7c (row 3) shows all the image class exhibiting close similarity in features with DME images. It is clearly visible that, except normal images, other image class possesses DME, as such these images contain DME features. While Fig. 7d (row 4) show each of the image class

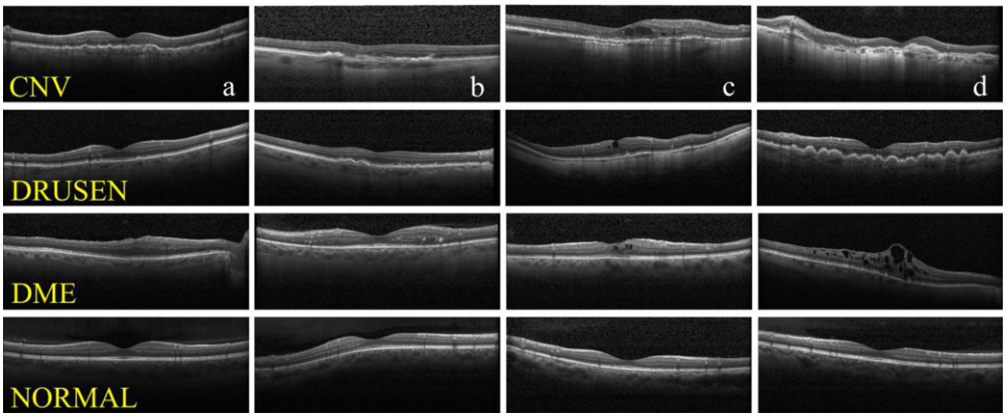

Fig. 7. Images from the dataset with high feature similarity.

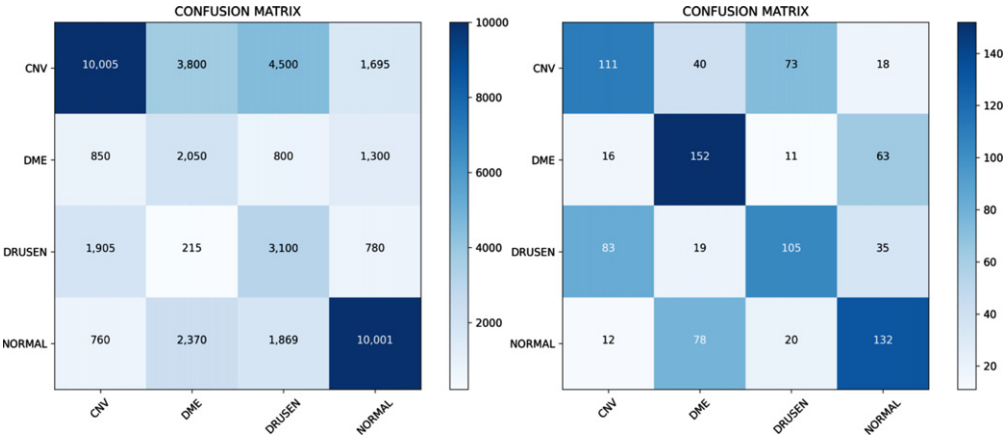

Fig. 8. The Confusion matrix for both the training and testing phase of the CNN technique. Training phase (left) and testing phase (right).

that classically depicted their distinguishable features accurately for proper identification. These types of images provide adequate features needed by the AI model to properly learn and utilize such knowledge for efficient classification. Unfortunately, the dataset used comprises of mixtures of the images shown in Fig. 7, with significant amount falling in this category. This is evidence in the results of the proposed model especially in the training phase. It is clearly observed that: (1) the CVN images were misclassified as DRUSEN and DME, and none as NORMAL. We observed that both CNV and DRUSEN possess similar features, especially disrupted retinal pigment epithelium (RPE) layer as shown in Fig. 7. At any stages of the CNV, it is very feasible to observe the disrupted RPE with pigments, while DRUSEN at early stage lack such feature. As such, it is a key feature for differentiating between the two diseases for the proposed model. Even with this key difference, there are cases where both CNV and DRUSEN are typical unidentifiable, which results in the misclassification of the proposed model. The problem of misclassification of the CNV as DME by the proposed model is typically due to the feature (hyperreflective foci, HRF), found in CNV, DRUSEN, and DME as shown in Fig. 7. HRFs are biomarker associated with retinal diseases [29–31, 34]. This feature similarity makes it difficult for the proposed model to differentiate accurately between the three class of diseases, (2) the drusen images were misclassified as NORMAL, followed by CVN, and none as DME. Early DRUSEN are like NORMAL, with no clear or distinct feature for identification. DRUSEN at a later stage possess similar feature with CNV, with close distinct feature that is difficult to differentiate properly. The stages of DRUSEN determined its associated features, as such, these features are also found in other retinal diseases, such as in CNV, and DME. This gives rise to misclassification in the proposed model, and (3) DME images were misclassified as NORMAL,

followed by CNV, and DRUSEN. Early DME have similar feature with NORMAL, with no clear feature for identification. It is easy to classify early DME as NORMAL, even by an experienced expert. With the progression of DME, one of the key features is the increasing in the quantity of HRFs which is found also in CNV and DRUSEN as shown in Fig. 7. This feature leads to misclassification of these diseases by the proposed model. Significant improvement is observed in the testing phase results, showing the extraordinary capacity of our model in learning the new features of the OCT images. Here, we observed significant drop in the misclassification of all the image class, but DME images misclassification as NORMAL and drusen images misclassification as CNV still remains a challenge.

In clinical setting, we established the fact that AI-driven approach in medical image (OCT) is feasible, and results obtained are reliable for clinician usage for interpretation and decision making. Furthermore, we incorporated a referral module into the proposed model to (1) inform the patients of the state of their eyes, and (2) guide the patients on the subsequent step needed after the results of the diagnosis. To effectively guide the patients, the referral module categorized the diseases based on the results (output) of the proposed model and refer such patients as follows: (1) CNV and DME images as immediate referrals, (2) DRUSEN images as regular referral, and (3) NORMAL images as monitoring. CNV and DME conditions demand an urgent referral to an ophthalmologist for immediate treatment, else the patient is at the risk of complication such as vessel bleeding and scarring, that could lead to permanent vision loss. DRUSEN is categorized as regular referral, since special medications are not urgently required for dry macular degeneration (AMD) such as DRUSEN, so such referral to an ophthalmology is less urgent. NORMAL condition is regarded as monitoring, which implies that the patient should visit ophthalmology when available for physical check-up.

Our model is effective in identifying potential pathology in the images, with a supportive referral decision making module, that facilitates timely and effective diagnosis of the leading causes of permanent vision loss. One of the key characteristics of the OCT imaging is the visualization of subretinal and/or intraretinal fluids associated with many retinal disorders. This retinal fluid can only be view in OCT image, thereby making OCT imaging a powerful diagnosis tool for the clinician. Additionally, associated retinal fluids visualized in the OCT image typically correlated well with the visual acuity. Therefore, retinal exudative diseases are recommended to be guided by the used of OCT imaging rather than by clinical examination or color fundus photography, which shows that the work presented here is significantly and clinically relevant. It is important to extend the usage of AI beyond diagnosis or detection of various diseases, but in making treatment recommendations and disease treatments. These potential areas are crucial and hold promising prospects for future investigation and exploration.

Table 3
Evaluation results of the CNN model based on the testing phase experimentations

	PRECISION	RECALL	F_1 SCORE
CNV	0.46	0.50	0.48
DME	0.63	0.53	0.58
DRUSEN	0.43	0.50	0.46
NORMAL	0.50	0.53	0.54
Accuracy			0.52
Macro Avg.	0.50	0.52	0.52
Weighted Avg.	0.51	0.52	0.52
Cohen's Score (κ)			0.72

We compare the proposed model with the state-of-the-art (SoA) CNN model. The CNN model uses a multilayer system that consists of the input layer, hidden layer, and an output layer. The hidden layer comprises of three layers namely (1) convolutional layers, (2) pooling layers, and (3) fully connected layers. We trained the SoA CNN model from the scratch with the OCT dataset, without using any pretrained model, to ascertain the efficiency of the proposed method. We reported the results of the comparison between the SoA CNN and the proposed model that is integrated with a pretrained model as its backbone architecture, to determine the effectiveness of both models. Figure 8 shows the confusion matrix of the state-of-the-art CNN model for both the training and testing phase. In addition, we present the results of the SoA CNN model in Table 3. Based on these results, it is obvious that the pretrained model integrated into the proposed model has significant effect on the overall accuracy of the proposed model. With an accuracy of 98% as compared with the SoA CNN model accuracy of 52%, we concluded that transfer learning is a great tool in improving the accuracy of a neural network.

Finally, the goal of this research work is to develop and investigate an AI-driven model in medical diagnosis and evaluate its soundness. According to the results of the proposed model, it is expected that the proposed model will facilitate screening programs

and provide a better referral system that will assists clinicians, especially in remote or low-resource areas. This will lead to a positive impact on public health, especially in clinical routine. Our future studies include using OCT images from various manufacturing vendors in both the training and testing datasets to make the proposed model a multi-vendor model that is universally useful. The proposed model has a great potential and could be used in a wide range of medical task, and in multidisciplinary field.

6. Conclusions

In this study, we proposed an AI-driven model by transfer learning approach for diagnosing and classification of medical images into four class of diseases. This image-based AI model classifies two major severe vision diseases (macular degeneration and diabetic retinopathy) using OCT images and has proven to have potential application in medical decision making. The main contribution of the proposed model include: (a) we built an AI model that transferred learned knowledge from a domain rich in data to the healthcare domain that suffers from shortage of data, (b) we incorporated a referral module into the proposed AI model to inform the patients on the state of their eyes, and guide the patients on further treatments recommendation, (c) to effectively guide the patients, we incorporated an AI referral module that categorized the diseases based on the results (output) of the proposed model and refer such patients as follows: (1) CNV and DME images as immediate referrals, (2) DRUSEN images as regular referral, and (3) NORMAL images as monitoring, and (d) we proposed a visualization model that assist clinicians in visualizing the diseased region, and a supportive tool in monitoring the detection process. Hence, we anticipated that this model would become a crucial diagnostic tool for clinicians in automated identification and referral systems for retinal diseases.

Our future work is to integrate attention mechanism to assist the model in classifying the disease accurately by focusing mainly on each class of diseases for adequate localization and identification. This will assist the model in differentiating distinctively between the various classes.

Author contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing–original draft preparation, IP Okuwobi.; writing– review and editing, IP Okuwobi, Z. Ding, J. Wan, S. Ding.; visualization, IP Okuwobi; supervision, S. Ding.; project administration, funding acquisition, IP Okuwobi and S. Ding.

Funding

This work was supported by the Youth Science and Technology Fund of Guangxi Natural Science Foundation (2021GXNSFBA220075), a grant from the Guangxi Postdoctoral Special Support Fund (C21RSC90ZN02), Scientific Research Fund (YXRSZN03 and UF20035Y), Scientific Research Fund (UF19008Y), and Guangxi Postdoctoral Special Support Fund (C22RSC90ZN01).

Data availability statement

This work uses the dataset provided by Kermany et al. [27], which is publicly available.

Conflicts of interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] R. Aggarwal, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis (2021) 23.

[2] W.-K. Chan et al., Using deep convolutional neural networks for enhanced ultrasonographic image diagnosis of differentiated thyroid cancer, *Biomedicines* **9**(12) (2021), Art. no. 12. doi: 10.3390/biomedicines9121771.

[3] H. Guedri, M. Ben Abdallah, F. Echouchene and H. Belmabrouk, Novel computerized method for measurement of retinal vessel diameters, *Biomedicines* **5**(2) (2017), Art. no. 2. doi: 10.3390/biomedicines5020012.

[4] A.-L. Deleu, M.J. Sathekge, A. Maes, B. De Spiegeleer, M. Sathekge and C. Van de Wiele, Characterization of FDG PET images using texture analysis in tumors of the gastrointestinal tract: a review, *Biomedicines* **8**(9) (2020), Art. no. 9. doi: 10.3390/biomedicines8090304.

[5] G. Pandolfo et al., Sharing the same perspective. Mental disorders and central serous chorioretinopathy: a systematic review of evidence from 2010 to 2020, *Biomedicines* **9**(8) (2021), Art. no. 8. doi: 10.3390/biomedicines9081067.

[6] T. Robins, J. Camacho, O.C. Agudo, J.L. Herraiz and L. Guasch, Deep-learning-driven full-waveform inversion for

ultrasound breast imaging, *Sensors* **21**(13) (2021), Art. no. 13. doi: 10.3390/s21134570.

[7] S. Lee et al., Deep-learning-based coronary artery calcium detection from CT image, *Sensors* **21**(21) (2021), Art. no. 21. doi: 10.3390/s21217059.

[8] K. Grąt, M. Grąt and O. Rowiński, Usefulness of different imaging modalities in evaluation of patients with non-alcoholic fatty liver disease, *Biomedicines* **8**(9) (2020), Art. no. 9. doi: 10.3390/biomedicines8090298.

[9] K. Nagib, B. Mezgebo, N. Fernando, B. Kordi and S.S. Sherif, Generalized image reconstruction in optical coherence tomography using redundant and non-uniformly-spaced samples, *Sensors* **21**(21) (2021), Art. no. 21. doi: 10.3390/s21217057.

[10] M. Siger, M. Owidzka, M. Świderek-Matysiak, W. Omulecki and M. Stasiołek, Optical coherence tomography in the differential diagnosis of patients with multiple sclerosis and patients with MRI nonspecific white matter lesions, *Sensors* **21**(21)(2021), Art. no. 21. doi: 10.3390/s21217127.

[11] R. Sacconi et al., Multimodal imaging assessment of vascular and neurodegenerative retinal alterations in Type 1 diabetic patients without fundoscopic signs of diabetic retinopathy, *Journal of Clinical Medicine* **8**(9) (2019), Art. no. 9. doi: 10.3390/jcm8091409.

[12] A. Arrigo et al., Choroidal patterns in stargardt disease: correlations with visual acuity and disease progression, *Journal of Clinical Medicine* **8**(9) (2019), Art. no. 9. doi: 10.3390/jcm8091388.

[13] H. Kataoka, K. Iwata and Y. Satoh, Feature evaluation of deep convolutional neural networks for object recognition and detection, *arXiv:1509.07627 [cs]*, Sep. 2015, Accessed: Nov. 28, 2021. [Online]. Available: http://arxiv.org/abs/1509.07627

[14] A. Mahmood, M. Bennamoun, S. An and F. Sohel, Resfeats: Residual network based features for image classification, in *2017 IEEE International Conference on Image Processing (ICIP)* (2017), 1597–1601. doi: 10.1109/ICIP.2017.8296551.

[15] X. Ren, H. Guo, S. Li, S. Wang and J. Li, A novel image classification method with CNN-XGBoost model, in *Digital Forensics and Watermarking*, Cham (2017), 378–390. doi: 10.1007/978-3-319-64185-0_28.

[16] S. Srivastava, P. Mukherjee, B. Lall and K. Jaiswal, Object classification using ensemble of local and deep features, in *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)* (2017), 1–6. doi: 10.1109/ICAPR.2017.8593056.

[17] I. Kandel and M. Castelli, Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review, *Applied Sciences* **10**(6) (2020), Art. no. 6. doi: 10.3390/app10062021.

[18] Research on image classification model based on deep convolution neural network | EURASIP Journal on Image and Video Processing | Full Text. https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-019-0417-8 (accessed Nov. 28, 2021).

[19] A. Pandey, M. Puri and A. Varde, Object Detection with Neural Models, Deep Learning and Common Sense to Aid Smart Mobility, in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)* (2018), 859–863. doi: 10.1109/ICTAI.2018.00134.

[20] C. Singh and J. Singh, Geometrically invariant color, shape and texture features for object recognition using multiple kernel learning classification approach, *Information Sciences* **484** (2019) 135–152. doi: 10.1016/j.ins.2019.01.058.

[21] Deep convolutional neural network based medical image classification for disease diagnosis | Journal of Big Data | Full Text. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0276-2 (accessed Nov. 28, 2021).

[22] D. Karthikeyan, A.S. Varde and W. Wang, Transfer learning for decision support in Covid-19 detection from a few images in big data, in *2020 IEEE International Conference on Big Data (Big Data)* (2020), 4873–4881. doi: 10.1109/BigData50022.2020.9377886.

[23] A.T. Sahlol, D. Yousri, A.A. Ewees, M.A.A. Al-qaness, R. Damasevicius and M.A. Elaziz, COVID-19 image classification using deep features and fractional-order marine predators algorithm, *Sci Rep* **10**(1) (2020), 15364. doi: 10.1038/s41598-020-71294-2.

[24] L. Liu et al., Deep learning based automatic multiclass wild pest monitoring approach using hybrid global and local activated features, *IEEE Transactions on Industrial Informatics* **17**(11) (2021), 7589–7598. doi: 10.1109/TII.2020.2995208.

[25] V. Kumar, D.R. Recupero, D. Riboni and R. Helaoui, Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes, *IEEE Access* **9** (2021), 7107–7126. doi: 10.1109/ACCESS.2020.3043221.

[26] A. Seemendra, R. Singh and S. Singh, Breast cancer classification using transfer learning, in *Evolving Technologies for Computing, Communication and Smart World*, Singapore (2021), 425–436. doi: 10.1007/978-981-15-7804-5_32.

[27] D.S. Kermany et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell* **172**(5) (2018), 1122–1131.e9. doi: 10.1016/j.cell.2018.02.010.

[28] M.L. McHugh, Interrater reliability: the kappa statistic, *Biochem Med (Zagreb)* **22**(3) (2012), 276–282.

[29] I.P. Okuwobi, Z. Ji, W. Fan, S. Yuan, L. Bekalo and Q. Chen, Automated quantification of hyperreflective foci in SD-OCT with diabetic retinopathy, *IEEE Journal of Biomedical and Health Informatics* **24**(4) (2020), 1125–1136. doi: 10.1109/JBHI.2019.2929842.

[30] I.P. Okuwobi et al., Automated segmentation of hyperreflective foci in spectral domain optical coherence tomography with diabetic retinopathy, *JMI* **5**(1) (2018), 014002. doi: 10.1117/1.JMI.5.1.014002.

[31] I.P. Okuwobi, Y. Shen, M. Li, W. Fan, S. Yuan and Q. Chen, Hyperreflective foci enhancement in a combined spatial-transform domain for SD-OCT images, *Translational Vision Science & Technology* **9**(3) (2020), 19. doi: 10.1167/tvst.9.3.19.

[32] I.P. Okuwobi, Z. Ding, J. Wan and S. Ding, DPSF: a novel dual-parametric sigmoid function for optical coherence tomography image enhancement, *Med Biol Eng Comput* **60**(4) (2022), 1111–1121. doi: 10.1007/s11517-022-02538-8.

[33] Z. Qian, C. Wu, H. Chen and M. Chen, Diabetic retinopathy grading using attention based convolution neural network, in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* **5** (2021), 2652–2655. doi: 10.1109/IAEAC50856.2021.9390963.

[34] S. Liu, D. Wang, F. Chen and X. Zhang, Hyperreflective foci in OCT image as a biomarker of poor prognosis in diabetic macular edema patients treating with Conbercept in China, *BMC Ophthalmology* **19**(1) (2019), 157. doi: 10.1186/s12886-019-1168-0.