Addressing Class Imbalance in Eye Disease Classification: A Transfer Learning-based Approach for Deep Learning Models using Retinal Images

Jaymar R. Cezar
College of Computing and Information Technologies
National University
Manila, Philippines
cezarjr@student.national-u.edu.ph

Abstract

Eye diseases pose a significant risk to global health, leading to vision loss and imposing substantial economic and social challenges. This research explores the application of deep learning using retinal images for the detection and classification of eye diseases. The study adopts a transfer learning approach, utilizing the EfficientNetB5 pre-trained model as the base, and addresses class imbalances through data augmentation techniques. The performance of the models trained on imbalanced and augmented datasets is evaluated using accuracy, F1 macro score, and F1 micro score. The results demonstrate the effectiveness of the transfer learning approach in detecting eye diseases, with both models achieving good performance.

1 Introduction

Eye diseases pose a significant risk to billions of individuals worldwide, primarily because of their potential to lead to vision loss. This global burden necessitates immediate attention and advancements in treatments to mitigate its impact on quality of life and the substantial economic and social challenges it presents. According to World Health Organization, at least 2.2 billion people are affected by near or distance vision impairments globally. Regarding the global financial burden of eye diseases, it is estimated that US$411 billion is the annual cost associated with productivity losses due to it.

With a multitude of eye diseases known to exist, individuals can be susceptible to a wide range of conditions, each requiring specific and timely responses. Some are untreatable, fortunately, most can be treated with early detection and proper medication. Globally, as of 2022, the top 5 leading causes of visual impairment and blindness are cataracts with 94 million, unaddressed refractive error affecting around 88.4 million individuals, age-related macular degeneration impacting approximately 8 million people, glaucoma affecting around 7.7 million individuals, and diabetic retinopathy impacting about 3.9 million individuals (Bourne et al., 2021).

Diagnosing eye diseases heavily rely on the expertise of ophthalmologists, which are the conventional methods that can be time-consuming. They perform physical examinations with the use of ophthalmology instruments followed by a comprehensive interpretation. Retinal imaging is one of the many methods to diagnose an ophthalmology disease, as various common eye diseases manifest in the retina of the eye. The retina is the layered tissue lining the interior of the eye, it is specifically at the very back of the eyeball. Retina enables the conversion of the light that enters the eye into a neural signal that is suitable to be sent by the optic nerves to the visual cortex of the brain to be processed and create the images that we see.

The increasing burden on healthcare systems calls for efficient and cost-effective solutions to address the rising demand for eye care services. Nowadays, Artificial Intelligence (AI), particularly its subset machine learning (ML), has effectively produced automated solutions to a wide range of problems in the real world, with healthcare being one of the most significant areas of research for ML researchers to create automated disease prediction and classification systems. In recent years, with the progression of technology, the use of artificial neural networks as the algorithm in machine learning, or its subset, deep learning (DL), has been applied progressively in healthcare, particularly in computer-aided detection systems that have been successfully used for the speedy identification and classification of multiple diseases like heart disease (Baccouche et al., 2020) and also on a particular eye disease, diabetic retinopathy (Jiang et al., 2019). Deep learning models can automatically learn complex patterns or features directly from the raw data, unlike traditional machine learning algorithms that rely heavily on expert-crafted features. This capability can be beneficial in medical image analysis, where the interpretation of high-resolution radiological images (such as X-rays, CT scans, and MRI scans) is crucial for accurate diagnoses (Razzak et al., 2018).

This research aims to explore the application of deep learning using retinal images as the dataset for the detection and classification of eye disease. The study will implement a transfer learning approach or the use of pre-trained models as it aims to improve the performance of a model by transferring the knowledge contained from different but related sources. Transfer learning is a popular and promising area in deep learning (Zhuang et al., 2020). The study will use EfficientNetB5 as the base, a member of the EfficientNet family that achieved state-of-the-art performance first introduced by Tan and Le (2020). The study will also address class imbalances and will apply data augmentation techniques to address the class imbalances. The results of this study will give insight into how effective a transfer learning approach on working with imbalanced data in detecting and classifying some of the most eye diseases that are the leading cause of vision loss.

2 Related Works

With the rapid advancement of technology and the increasing availability of large amounts of data, ML researchers have demonstrated significant interest in developing medical expert systems to automate diagnostic processes (Marghalani and Arif, 2019; Daghrir et al., 2020; Wang et al., 2020). These sophisticated expert systems can generate accurate detections based on pre-defined rules and on features a neural network derived. Hence, medical field research particularly ophthalmology is now witnessing significant growth. This section provides a brief review of related work in this area.

Lee et al. (2017) focused their research on distinguishing a normal from those patients with age-related macular degeneration (AMD) that can be seen in the optical coherence tomography (OCT) of the patients. They utilized a large OCT database linked with electronic medical records (EMRs) to train a modified version of the VGG16 convolutional neural network using Caffe and Python, achieving high accuracy. At the image, macula, and patient levels, the deep learning model achieved an area under the receiver operating characteristic curve (AUC-ROC) values of 92.78%, 93.83%, and 97.45%, respectively, with corresponding accuracies of 87.63%, 88.98%, and 93.45%. Their research demonstrates the effectiveness of deep learning in accurately categorizing OCT images and its contribution to computer-aided diagnosis in ophthalmology.

In another study, Nazir et al. (2020) tackled diabetes-based eye diseases. They presented a novel automated approach for the early detection of diabetic retinopathy, diabetic macular edema, and glaucoma. They utilized the combination of a Fast Region-based Convolutional Neural Network (FRCNN) algorithm with fuzzy k-means clustering (FKM) which resulted in achieving accurate disease localization and segmentation. The proposed method is evaluated and compared against state-of-the-art techniques using five different datasets such as Diaretdb1, MESSIDOR, ORIGA, DR-HAGIS, and HRF demonstrating its effectiveness in disease detection and segmentation.

In a study entitled "Discriminative Kernel Convolution Network for Multi-Label Ophthalmic Disease Detection on Imbalanced Fundus Image Dataset," Bhati et al. (2023) worked on an imbalanced dataset that contains around 5000 organized fundus images of the left and right eyes of patients. Their study introduced a discriminative kernel convolution network (DKCNet) that effectively captures region-wise features without introducing excessive computational complexity. They address the class imbalance by applying oversampling and undersampling techniques. The researchers integrated the DKCNet with an Inception-Resnet backbone network that resulted in promising results with an AUC of 96.08, an F1-score of 94.28, and a kappa score of 0.81. Furthermore, the proposed model exhibits robust performance on three publicly available benchmark datasets, even when tested on completely unseen fundus images, demonstrating its potential for broader application in the field.

In a study conducted by Sugeno et al. (2021), they utilized a transfer learning approach, particularly the EffcientNet-B3 for the lesion detection and severity grading of diabetic retinopathy. They trained the model using the publicly available dataset which is the APTOS 2019 Blindness Detection to train the model. After dataset preprocessing, the trained model achieved high specificity and sensitivity values (>0.98) for identifying DR retinas. For severity grading, classification accuracy values of 0.84, 0.95, and 0.98 were achieved for the 1st, 2nd, and 3rd predicted labels, respectively. Their study demonstrates the effectiveness of EfficientNet-B3 in severity grading, along with the corresponding retinal areas. Lesion extraction successfully captured red and white lesions, including soft and hard exudates. The extracted lesion areas were validated against ground truth using the DIARETDB1 database images, demonstrating overall accuracy. These proposed simple and easily applicable methods hold promise for aiding in the detection and severity grading of DR, assisting in the selection of appropriate treatment strategies.

3 Methodology

3.1 Dataset

The dataset used in this study is obtained from Kaggle, entitled "eye_diseases_classification." The dataset is a diverse collection of retinal images, with each image belonging to one of four distinct classes:

- **Cataract:** This class consists of images depicting cataracts.
- **Diabetic Retinopathy:** This class includes images representing diabetic retinopathy.
- **Glaucoma:** This class contains images associated with glaucoma.
- **Normal:** This class comprises normal retinal images.

Each class in the dataset has approximately 100 images. Figure 1 shows some of the images in the dataset.

3.2 Data Preparation

In this study, the focus is on working with an imbalanced dataset. To address this, a random sampling approach was implemented. It was applied to choose 300 images from each class, except for the normal class, which retained all 1074 images. As a result, the dataset achieved a new distribution ratio of 10:3:3:3. Figure 2 provides a clear visualization of the percentage distribution for each class in the entire dataset. Notably, the normal class accounts for over half of the dataset with 54.4%, while each of the eye disease classes represents 15.2%.

The dataset was split using the Python library 'splitfolders' to create distinct folders for the training, validation, and test set folders. The created subsets of the dataset have a ratio of 7:2:1, with 70% of the data allocated for the training set, 20% for the validation set, and 10% for the test set.

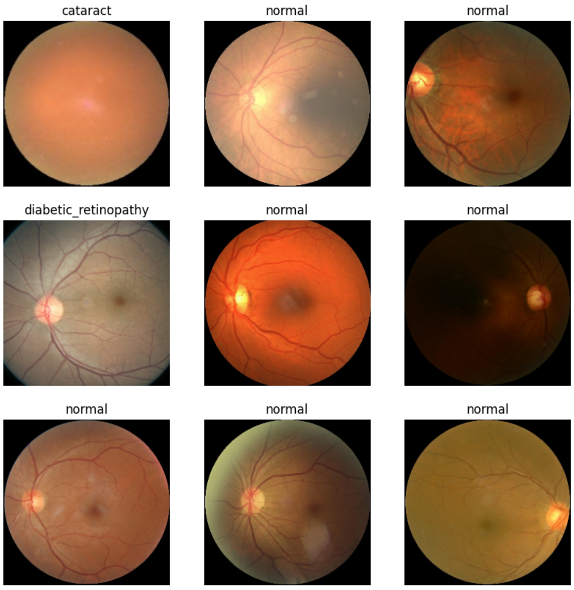


Figure 1: Sample images from the dataset.

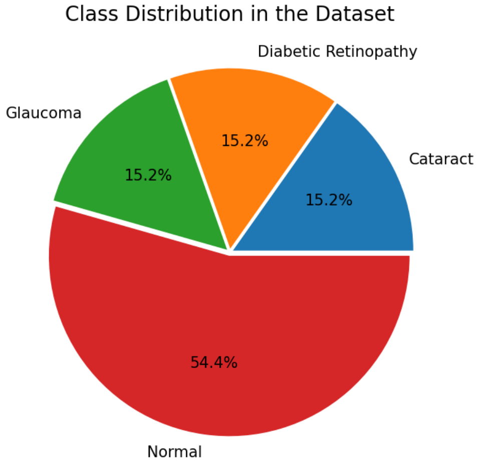


Figure 2: Class Distribution in the dataset.

Subsequently, the dataset was loaded to a TensorFlow dataset using the 'tf.data.dataset' API. To ensure consistency, the images were resized to a uniform size of 224 pixels in height and 224 pixels in width. Additionally, a batch size of 64 was set to process multiple images in parallel during training iterations, optimizing computational resources and accelerating the training process.

3.3 Preprocessing

For the preprocessing, the rescaling of the pixel values of the images to a range of [0,1] is not applied because the pre-trained model expects the inputs to be float tensors of pixels with values in the [0-255] range. Nonetheless, to enhance performance and efficiency, the dataset was further optimized using the 'Dataset. prefetch' method, which leverages autotuning capabilities to overlap data preprocessing and model execution.

3.4 Model

In this study, the researcher follows a transfer learning approach for the model. The pre-trained model that will be used as the base model is the EfficientNetB5, a member of the EfficientNet family, which is a series of convolutional neural network (CNN) models known for their efficiency and state-of-the-art accuracy. The pre-trained model is trained on more than a million images from the ImageNet database (Deng et al., 2009). The classification head of the pre-trained model is excluded to customize it for the task at hand. To fine-tune the model, a custom network is added on top of the trained base model. A subset of layers is selectively made trainable, starting from layer 540, since the base model has 574 layers. This fine-tuning process allows the model to adapt and learn task-specific features.

Following the base model architecture setup, additional layers are added. The first layer is batch normalization, followed by a fully connected or dense layer that has 256 neurons and ReLU activation for feature extraction, and a dropout layer with a rate of 0.5 for regularization. Finally, the classification head consists of a dense layer with a softmax activation function, containing 4 units, where each unit represents a class of either an eye disease or a normal class in the dataset. After constructing the model, it is compiled using the 'adamax' optimizer, 'sparse_categorical_crossentropy' loss function, and the 'accuracy' metric.

The created model will be trained on both the imbalanced training set and the augmented training set with both having a maximum number of 30 epochs. The researcher applied an early stopping method to stopped the training phase and restore the best weights when the validation accuracy metric stopped improving from the last 7 epochs. The researcher will then evaluate the model based on the chosen evaluation metrics.

3.5 Data Augmentation

The researcher applied a data augmentation technique to battle the imbalances in the dataset using the 'ImageDataGenerator' class. The ImageDataGenerator was used to generate images from the existing images in the minority class of the training set by applying various transformations. These transformations include rotation with a range of 45 degrees, shifting the width and height by 20%, applying a shear transformation with a range of 0.1, zooming with a range of 0.2, and performing horizontal flipping. The fill mode is set to 'constant' with a value of 0 meaning that the gaps will be filled with black color. The ImageDataGenerator generates augmented batches of images by applying the defined transformations. The augmented images are saved to a copy of the directory of the original training set thus adding it with the original images. Figure 3 shows some generated images.

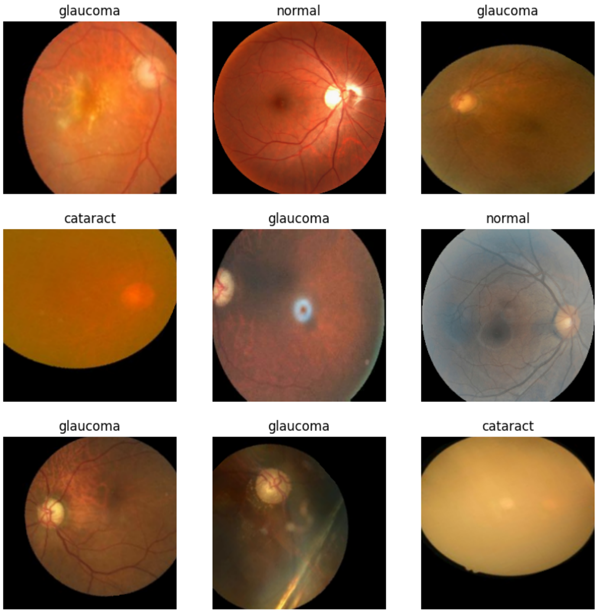


Figure 3: Generated images from the Augmentation.

3.6 Evaluation Metrics

To measure the performance of the model in the training phase, the following metrics were used:

Sparse-Categorical Cross-Entropy loss: The loss represents the discrepancy between the predicted outputs of the model and the true labels in the training dataset. Lower loss values indicate that the model's predictions are closer to the actual labels. Sparse-Categorical Cross-Entropy (scce) is used in multi-class classification. It produces a category index of the most likely matching category.

$$J(w) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}^i) + (1 - y_i)\log(1 - \hat{y}^i)\right]$$

Accuracy: Accuracy measures the percentage of correctly classified instances in the training dataset. It is calculated by dividing the number of correct predictions by the total number of instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Instances}}$$

The researcher also evaluated the model's performance on a separate test set. The metrics used are F1 macro, F1 micro, and accuracy.

F1 Score: The F1 score is the weighted average of precision and recall. It combines both precision and recall into a single score, providing a balanced view of the model's performance. The F1 score is the harmonic mean of precision and recall and is computed with the following formula:

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Macro F1 Score: The Macro F1 score is the unweighted mean of the F1 scores calculated per class. It is the simplest aggregation for the F1 score.

Micro F1 Score: The Micro F1 Score is the F1 score calculated using the total number of True Positives (TP), False Positives (FP), and False Negatives (FN), instead of individually for each class.

4 Results and Discussion

The created neural network derived from EfficientNet and added custom layers was first trained on the imbalanced training set. Figure 4 shows the learning curve of the neural network. As stated earlier in the methodology section, the training process was run on 30 epochs with a total of 22 batches in each epoch and each batch consisted of 64 images. The changes in the training phase of the model were closely monitored based on loss and accuracy in both the training set and validation set. As we can see in the figure, in the initial epochs the model displayed promising performance as indicated by the continues to decrease in both training loss and validation loss. The accuracy in both the training set and validation set also showed a continuously increasing trend. As the training process continued, the model also continuously exhibit remarkable performance. The trends in the initial epochs indicate that the model is effectively learning from the training set and its generalization ability is also good as seen by the continues increasing in validation accuracy and decreasing validation loss. However, after the model recorded its highest validation accuracy and the lowest loss at epoch 18, the curve started to show slight fluctuations in validation accuracy and display an increase in validation loss. These observations in the training phase indicate overfitting. But with the implemented early stopping mechanism, the best weights based on the validation accuracy which are achieved in epoch 18 were restored although the training process continued until the 25th epoch where the model stopped. At the best epoch, the model achieved a training accuracy of 98.55 with a loss of 0.00382 and a validation accuracy of 93.91% with 0.2760. Overall, the results demonstrate that the neural network model achieved high accuracy and relatively low loss despite having imbalanced training and validation data.

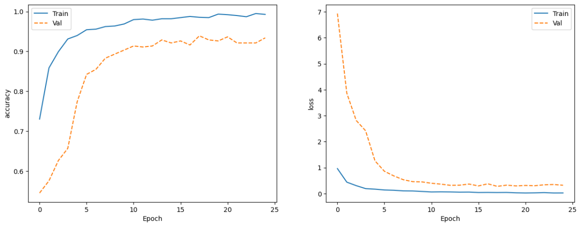


Figure 4: Learning Curve of the model trained on an imbalanced data.

As described in the methodology section, Data augmentation techniques were applied in the minority classes in the training set. 540 images were generated for each minority class, and after adding it to the original images of 210, each minority class has now 750 images. Consequently, the number of instances for each eye disease class is now nearly equivalent to the number of images in the normal class, which contains 751 instances. The augmented training set's class distribution is depicted in Figure 5. Notably, the figure illustrates that each class now represents approximately 25% of the total, achieving a balanced training set.

After applying the augmentation on the training set, the created model was also applied to the augmented training set to address the issue of dataset imbalance, aiming to improve the model's performance. Same as training in an imbalanced dataset, the model is set to train in 30 epochs, but now each epoch has a total of 47 batches, and the batch size is still 64. We can see in Figure the learning curve of the model trained on a balanced training set. The figure shows the model has steady progress in the initial epochs particularly until the 6th epoch, with the validation accuracy and training accuracy continuing to increase, and the training loss and

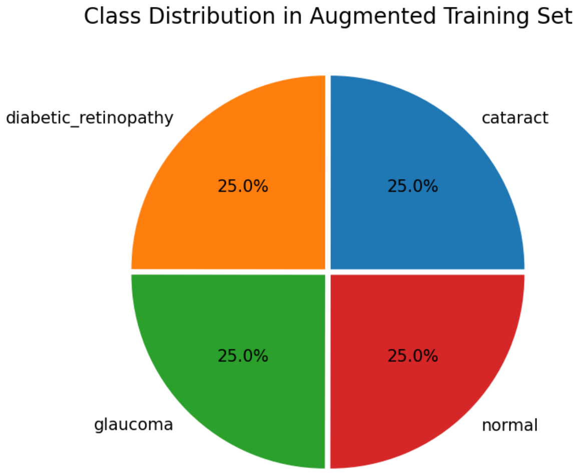


Figure 5: Class distribution in the Augmented Training Set.

validation loss continuing to decrease. However, at the start of the 7th epoch, the trend in validation loss and validation accuracy started to fluctuate. The model still reached its highest validation accuracy in the 8th epoch, but after this, the model's accuracy on validation data started to slowly decrease which indicates overfitting, and it resulted in early stopping in the epoch 15. However, the best weights were still restored. The performance of the model that is trained on the augmented training set which is now balanced did not increase compared to training in the imbalanced data. The model just converges faster to the local minima.

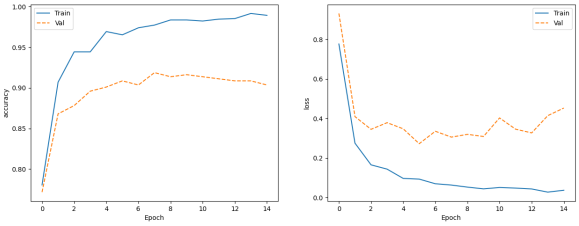
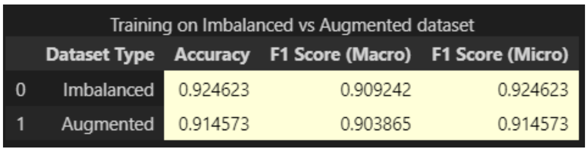


Figure 6: Class distribution in the Augmented Training Set.

Both the models trained on imbalanced training data and augmented training data were used in detecting eye disease in separate test data. The accuracy, f1 macro score, and f1 micro score of the models can be seen in Figure 7. The imbalanced dataset achieved an accuracy of 0.924623, indicating that it correctly classified approximately 92.46% of all the instances and it yielded an F1 score (macro) of 0.909242. The F1 score (micro) has the same value with accuracy, which is 0.924623. It is worth noting that the f1 macro score is the better score for an imbalanced dataset as macro F1 gives each class equal importance. The macro F1 score will still reflect true model performance even when the classes are imbalanced.


Training on Imbalanced vs Augmented dataset

	Dataset Type	Accuracy	F1 Score (Macro)	F1 Score (Micro)
0	Imbalanced	0.924623	0.909242	0.924623
1	Augmented	0.914573	0.903865	0.914573

Figure 7: Training on Imbalanced vs Augmented training data.

On the other hand, the augmented dataset achieved a slightly lower accuracy of 0.914573, implying that it correctly classified around 91.46% of the instances. The F1 score (macro) of 0.903865 indicates a similar balanced performance across classes as the imbalanced dataset. The F1 score (micro) of 0.91457 also demonstrates a comparable ability to capture true positives, false positives, and false negatives as the imbalanced dataset.

Overall, despite implementing augmentation techniques to address class imbalances, the performance of the model did not improve significantly. However, it is important to highlight that both models based on a transfer learning approach, trained on the imbalanced dataset and the augmented dataset, demonstrated good performance, achieving accuracy and f1 scores of at least 90%.

5 Conclusion

This research focused on addressing class imbalances in a retinal image dataset for the detection of eye diseases. Two approaches were explored: training on an imbalanced dataset and training on an augmented dataset.

A transfer learning approach was adopted, using the EfficientNetB5 pre-trained model as the base model. Custom layers were added on top of the base model, and selective fine-tuning was performed to adapt the model to the specific classification task. The trained models were evaluated on test data, and their accuracy, F1 macro score, and F1 micro score were analyzed. The results showed that both the model trained on the imbalanced dataset and the model trained on the augmented dataset achieved good performance, with accuracy and F1 scores exceeding 90%. Although the augmented dataset did not significantly improve the model's performance compared to the imbalanced dataset, both approaches proved effective in detecting eye diseases.

These findings highlight the effectiveness of the transfer learning approach in addressing class imbalance and the potential of data augmentation techniques in improving model performance. The research contributes to the field of eye disease classification by providing insights into training models with imbalanced datasets and exploring the impact of augmentation techniques.

This study concludes that accurate classification of eye diseases can be achieved even with imbalanced datasets using a transfer learning approach. Future research can further investigate other strategies for addressing class imbalance and explore the impact of different data augmentation techniques on model performance.

References

Asma Baccouche, Begoña Zapirain, Cristián Castillo, and Adel Elmaghraby. 2020. Ensemble deep learning models for heart disease classification: A case study from mexico. *Information*, 11:207.

Amit Bhati, Neha Gour, Pritee Khanna, and Aparajita Ojha. 2023. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset. *Computers in Biology and Medicine*, 153:106519.

R Bourne, JD Steinmetz, S Flaxman, et al. 2021. Gbd 2019 blindness and vision impairment collaborators; vision loss expert group of the global burden of disease study. trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study. *Lancet Glob Health*, 9(2):e130–e143.

Jinen Daghrir, Lotfi Tlig, Moez Bouchouicha, and Mounir Sayadi. 2020. Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach. In *2020 5th international conference on advanced technologies for signal and image processing (ATSIP)*, pages 1–5. IEEE.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Hongyang Jiang, Kang Yang, Mengdi Gao, Dongdong Zhang, He Ma, and Wei Qian. 2019. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2045–2048. IEEE.

Cecilia S. Lee, Doug M. Baughman, and Aaron Y. Lee. 2017. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. *Ophthalmology Retina*, 1(4):322–327.

Bashayer Fouad Marghalani and Muhammad Arif. 2019. Automatic classification of brain tumor and alzheimer's disease in mri. *Procedia Computer Science*, 163:78–84.

Tahira Nazir, Aun Irtaza, Ali Javed, Hafiz Malik, Dildar Hussain, and Rizwan Ali Naqvi. 2020. Retinal image analysis for diabetes-based eye disease detection using deep learning. *Applied Sciences*, 10(18).

Muhammad Razzak, Saeeda Naz, and Ahmad Zaib. 2018. *Deep Learning for Medical Image Processing: Overview, Challenges and the Future*, pages 323–350.

Ayaka Sugeno, Yasuyuki Ishikawa, Toshio Ohshima, and Rieko Muramatsu. 2021. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in Biology and Medicine*, 137:104795.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

Wu Wang, Junho Lee, Fouzi Harrou, and Ying Sun. 2020. Early detection of parkinson's disease using deep learning and machine learning. *IEEE Access*, 8:147635–147646.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning.