CNN Models for Eye Disease Classification on Retinal OCT Images

Geng Niu[1*]

[1] Department of Electrical and Computer Engineering, New York University
GN2279@nyu.edu

Abstract

This project studies retinal disease classification from optical coherence tomography (OCT) images, a task that is essential for early ophthalmic diagnosis but traditionally relies on perfessional doctor interpretation. Using the publicly available OCT2017 dataset, we make the problem as a four-class classification task involving choroidal neovascularization (CNV), diabetic macular edema (DME), DRUSEN, and normal retinal conditions. We adopt a transfer learning approach based on convolutional neural networks, evaluating three widely used architectures: ResNet-18 (He et al. 2016), EfficientNet-b0 (Tan and Le 2019), and MobileNetV2 (Sandler et al. 2018), all initialized with ImageNet pre-trained weights. Models are trained on the official training split dataset and evaluated on offical split test set of 968 images using accuracy, macro-F1, macro-precision, macro-recall, and confusion matrix analysis. All three models achieve consistently strong performance, with test accuracy ranging from 99.28-percent to 99.59-percent and macro-F1 scores exceeding 0.99, indicating balanced classification across all disease categories. Confusion matrices show that misclassifications are rare and primarily occur between clinically blurry related classes, such as DRUSEN and CNV. To enhance interpretability, Grad-CAM is applied to visualize class-discriminative regions, demonstrating that the models focus on anatomically meaningful retinal structures, even when the prediction is wrong. Overall, the results confirm the effectiveness of transfer learning and highlight the importance of explainable models in medical imaging applications.

Keywords: deep learning, Convolutional Neural Networks, Medical Image Classification, Transfer Learning, Grad-CAM

1 Introduction

Optical coherence tomography (OCT) has become a widely used imaging modality in ophthalmology for the diagnosis of retinal diseases, as it provides high-resolution cross-sectional views of retinal structures. Accurate interpretation of OCT images is critical for identifying pathological conditions such as choroidal neovascularization (CNV), diabetic macular edema (DME), and DRUSEN, which are associated with vision loss if left untreated. However, human analysis of OCT image is time-consuming and requires clinical experience. So, it motivate the development of automatic and reliable diagnostic systems.

In this project, we focus on the problem of 4-class retinal disease classification from OCT images with four categories: CNV, DME, DRUSEN, and NORMAL. Convolutional neural networks (CNNs) have demonstrated strong performance in medical image analysis. Nevertheless, it's so important to understand how different CNN models perform under a unified experimental setting and whether their predictions are based on clinically meaningful visual patterns rather than meaningless correlations.

While previous work has shown that transfer learning can be effective for OCT images classification, comparative analysis across multiple CNN backbones are often limited. In particular, quantitative performance alone may not be enough for medical applications without accompanying interpretability.

Our main contributions are as follows:

- We investigate retinal disease classification on the OCT2017 dataset using transfer learning with multiple CNN models.
- We implement and compare ResNet-18, EfficientNet-B0, and MobileNetV2 under a unified training and evaluation framework.
- We evaluate model performance using accuracy and multiple macro-averaged metrics to ensure balanced assessment across disease classes.
- We apply Grad-CAM to analyze model interpretability, examine both correct predictions and failure cases and perform failure case and domain-shift analysis.

2 Background and Related Work

Deep learning, particularly convolutional neural networks (CNNs), has become a mainstream approach for visual recognition tasks by automatically learning image features from raw image data. CNNs are also effective in medical imaging because many diagnostic tasks rely on spatial patterns, textures, and structural variations, features that can be well captured through convolutional operations. Transfer learning is a commonly used method that leverages models pre-trained on large natural image datasets like ImageNet and fine-tunes them to suit medical tasks, thereby improving convergence stability.

Early research has shown that deep learning models can achieve performance comparable to human experts in vari-

[*] These authors contributed equally.

ous medical imaging applications. A landmark study by Esteva et al. (2018) (Kermany et al. 2018) demonstrated that CNN-based systems can directly identify medical diagnoses and disease types from images, achieving specialist-level performance in classifying ocular lesions. This work highlighted the clinical feasibility of image-based deep learning and spurred subsequent joint research in both medicine and computer science. In retinal imaging, similar CNN-based methods have been applied to optical coherence tomography (OCT) for detecting retinal abnormalities, often achieving high classification accuracy.

Most existing OCT classification systems use standard CNN architectures, such as ResNet, and often apply transfer learning to enhance performance. Evaluation typically uses accuracy and confusion matrices on the test set, and some studies also report precision and recall. However, relying only on accuracy can hide performance differences across different disease categories, especially in multi-class medical situation. Furthermore, while many studies report high predict performance, few systematically using interpretability techniques to analyze model behavior and validate whether predictions are based on clinically meaningful image regions.

Our project builds upon these findings, applying multiple transfer learning based CNN models to retinal OCT classification and focusing on comparative evaluation and interpretability. Instead of proposing new architectures, we concentrate on controlled condition comparisons of various modern CNN models within same experimental framework. Moreover, we introduce macro-average evaluation metrics and Grad-CAM visualizations to provide a more overall balanced and interpretable evaluation of model performance, making our work a comparative analysis in the existing study.

3 Problem Statement and Goals

3.1 Problem Description

This project aims to solve the problem of automatic classification of retinal diseases based on optical coherence tomography (OCT) images. Given an OCT scan image as input, the task is to predict the corresponding retinal condition based on a predefined set of disease categories. Specifically, we construct the problem as a multi-class image classification task, with output categories including: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal retinal condition.

The system input is a grayscale OCT image, which is converted to a three-channel representation and adjusted to a fixed spatial resolution suitable for a convolutional neural network. The output is a discrete category label indicating the predicted retinal condition. The model aims to learn discriminative visual features that capture clinically relevant structural patterns in OCT images, thereby accurately distinguishing between pathological and normal states, as well as different disease types.

Besides prediction accuracy, model interpretability is also an important aspect of this problem. Since OCT-based diagnosis is a medical application, it is necessary to understand whether the model's decisions are based on meaningful retinal structures, rather than spurious correlations. Therefore, in addition to classification performance, this project also considers post-hoc interpretability as an integral part of the problem.

3.2 Objectives and Scope

This project aims to design, implement, and evaluate a deep learning based image classification system for retinal diseases, using optical coherence tomography (OCT) images for classification, with a focus on comparsion analysis and interpretability. This project does not propose a new neural network architecture, but rather focuses on understanding the effectiveness of existing convolutional neural network models in a unified and controlled experimental environment.
Specifically, the project objectives are as follows:

- Compare the performance of various convolutional neural network (CNN) architectures (ResNet-18, EfficientNet-B0, and MobileNetV2) on the OCT2017 dataset using transfer learning.
- Evaluate model performance using balanced metrics, including accuracy, macro-mean precision, recall, and F1 score, to assess model performance across all disease categories.
- Visualize and analyze model behavior using Grad-CAM, and examine cases of correct predictions and misclassifications to evaluate interpretability.
- Report and discuss the quantitative results and qualitative insights obtained from confusion matrices and interpretability analysis.

The scope of this project is limited to in-distribution evaluation on carefully curated public datasets. This article does not cover clinical deployment, real-time inference, or cross-dataset generalization. Architectural modifications are also not discussed; instead, standard training settings are used to ensure fairness and reproducibility in model comparisons.

4 Approach

4.1 Overall Design

Our approach follows a modular deep learning pipeline for retinal disease classification from OCT images, combining transfer learning based image classification with interpretability analysis. The overall system is designed to ensure fair model comparison, reliable evaluation, and clinically meaningful interpretation of predictions.

At a high level, the pipeline consists of four main components: data preprocessing, model training, performance evaluation, and explainability analysis. These components interact sequentially and share common inputs and outputs to maintain consistency across experiments.

First, OCT images are preprocessed to match the input requirements of convolutional neural networks, including resizing, normalization, and data loading using a standardized pipeline. The processed images are then fed into CNN-based classifiers initialized with ImageNet pre-trained weights.

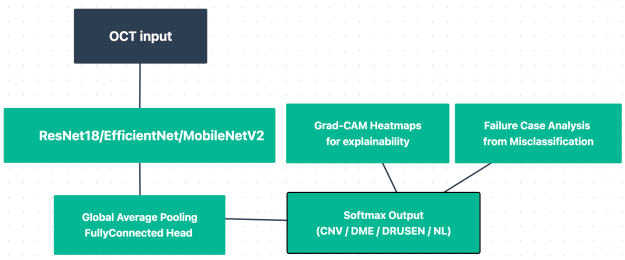


Figure 1: Model architecture workflow

Multiple backbone architectures—ResNet-18, EfficientNet-B0, and MobileNetV2—are evaluated under the same training configuration to enable controlled comparison. During training, model parameters are fine-tuned on the OCT training set using supervised learning.

After training, each model is evaluated on a held-out test set using quantitative metrics such as accuracy and macro-averaged precision, recall, and F1 score, along with confusion matrix analysis. These metrics provide a balanced assessment of performance across all disease categories. Finally, Grad-CAM is applied as an explainability module to visualize class-discriminative regions in OCT images. The interpretability analysis is performed on both correctly classified samples and misclassified cases, linking model predictions back to clinically relevant retinal structures.

4.2 Models, Methods, or System Components

Model Architectures This project adopts convolutional neural networks (CNNs) with transfer learning as the core modeling approach for retinal disease classification. Three representative CNN backbones are evaluated: ResNet-18, EfficientNet-B0, and MobileNetV2. These architectures differ in depth, parameter efficiency, and design philosophy, enabling a systematic comparison under a unified experimental framework.

All models are initialized with ImageNet pre-trained weights. The original classification layers are replaced with a task-specific fully connected layer that outputs predictions for four retinal categories: CNV, DME, DRUSEN, and NORMAL. Apart from the final classification layer, all network parameters are fine-tuned on the OCT dataset, allowing the models to adapt general visual features to retinal imaging characteristics while maintaining consistent architecture settings across experiments.

Training Procedure and Optimization Model training is performed as a multi-class classification task. During training, input OCT images are passed through the network to produce class predictions, which are compared against ground-truth labels using a standard multi-class classification loss. Optimization is carried out using the Adam optimizer, which provides adaptive parameter updates and stable convergence during fine-tuning.

To ensure fair comparison across models, all architectures are trained using the same learning rate, batch size, and number of epochs. Training is intentionally limited to a small number of epochs, as transfer learning enables rapid convergence on the OCT dataset and also meets the requirement of project. No additional regularization techniques or architecture-specific tuning are introduced, allowing performance differences to primarily reflect the underlying model designs.

4.3 Data Pipeline and Implementation

The data pipeline is implemented using PyTorch's dataset and data loading utilities. OCT images are resized and normalized using ImageNet statistics to ensure compatibility with pre-trained models. All models share identical preprocessing and data splits. This design ensures reproducibility and eliminates confounding factors when comparing different architectures.

Evaluation and Analysis Methods Model performance is evaluated on a held-out test set using accuracy and macro-averaged precision, recall, and F1 score. Macro-averaged metrics treat each class equally and are therefore more appropriate for multi-class medical classification tasks than accuracy alone. Confusion matrices are further used to analyze class-wise prediction behavior and identify misclassification patterns.

To enhance result interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied as a post-hoc analysis technique. Grad-CAM produces visual explanations by highlighting image regions as heatmap that contribute most strongly to a model's prediction. This analysis is performed on both correctly classified samples and misclassified cases, enabling direct assessment of whether the models focus on clinically meaningful retinal structures.

Implementation All experiments are implemented using the PyTorch deep learning framework, with pre-trained CNN architectures obtained from the Torchvision model library. Evaluation metrics, including accuracy and macro-averaged precision, recall, and F1 score, are computed using scikit-learn, while visualization and analysis are performed with Matplotlib and Seaborn. Grad-CAM–based interpretability analysis is conducted using the pytorch-grad-cam library.

To ensure reproducibility and fair comparison, all models are trained under the same configuration, using the Adam optimizer with a fixed learning rate and batch size for a small number of training epochs, as transfer learning enables rapid convergence on the OCT dataset. Experiments are run on Colab with A100 GPU acceleration, significantly reducing training time, with each model completing training within 30 minute.

Figure 2 summarizes the final test-set performance of all evaluated models, demonstrating consistently strong and balanced results across disease categories.

5 Data and Experimental Setup

5.1 Dataset

This project uses the OCT2017 dataset, a publicly available optical coherence tomography (OCT) dataset on kaggle commonly used for retinal disease classification research.

	Model	Accuracy	Macro-F1	Macro-Precision	Macro-Recall
0	ResNet-18	0.992769	0.992789	0.992972	0.992769
1	EfficientNet-B0	0.993802	0.993801	0.993952	0.993802
2	MobileNetV2	0.995868	0.995874	0.995935	0.995868

Figure 2: Performance Table

The dataset consists of labeled OCT images categorized into four clinically relevant classes: choroidal neovascularization (CNV), diabetic macular edema (DME), DRUSEN, and normal retinal condition (NORMAL). The dataset provides standardized splits for training, validation, and testing, making it suitable for controlled experimental evaluation. In total, the dataset contains a large number of OCT images, with 83,484 images in the training set, 32 images in the validation set, and 968 images in the test set. Given the extremely small size of the official validation split, the validation set is used only as a sanity check during training, while all quantitative performance evaluation is conducted on the held-out test set.

All images are preprocessed to match the input requirements of convolutional neural networks. Specifically, OCT images are resized to a fixed spatial resolution and converted to a three-channel format. Standard normalization using ImageNet statistics is applied to ensure compatibility with pretrained models. No additional data augmentation techniques are employed, allowing the experiments to focus on the effect of model architecture and transfer learning rather than augmentation strategies.

5.2 Baselines or Comparison Points

We use ResNet-18 as the primary baseline model. ResNet-18 is a widely adopted convolutional neural network architecture and has been commonly used in medical image classification tasks. As a result, it serves as a strong and representative baseline for retinal OCT classification with transfer learning. In addition to the baseline, we compare ResNet-18 against two CNN architectures: EfficientNet-B0 and MobileNetV2. These models are selected to provide meaningful comparison. EfficientNet-B0 represents a parameter-efficient architecture that emphasizes compound scaling, while MobileNetV2 is a lightweight model designed for efficiency and reduced computational cost. All models are evaluated under the same experimental conditions, including identical data splits, preprocessing steps, and training configurations.

This project focuses on a controlled within-dataset comparison. This design choice avoids confounding effects due to differences in preprocessing, evaluation protocols, or dataset versions, and allows for a fair assessment of relative model behavior and performance.

5.3 Evaluation Protocol

All model trained for 5 epoches with fixed random seeds. Model performance is evaluated using test set provided by the OCT2017 dataset. We report accuracy as an overall performance measure, along with macro-averaged precision, recall, and F1 score to ensure balanced evaluation across all disease categories. Confusion matrices are used to analyze class-wise prediction behavior and identify systematic misclassification patterns.

Beyond aggregate metrics, qualitative evaluation is performed using Grad-CAM visualizations. These visualizations are applied to both correctly classified samples and misclassified cases to examine whether model predictions are supported by clinically meaningful retinal regions. This combination of quantitative and qualitative evaluation provides a comprehensive assessment of both model performance and interpretability.

6 Result

6.1 Quantitative Results

Figure 2 shows the test set performance of all evaluated models. Overall, all three CNN architectures achieve strong results, confirming the effectiveness of transfer learning for retinal OCT classification. MobileNetV2 achieves the best overall performance, followed closely by EfficientNet-B0 and the ResNet-18 baseline. Performance differences are consistent across accuracy and macro-averaged metrics, indicating balanced classification across disease categories. Obviously, the lightweight MobileNetV2 outperforms the larger baseline model, suggesting that higher model complexity is not required to achieve high accuracy on this dataset.

6.2 Illustrative result

Figure 3 shows the correct classification in baseline ResNet-18 model with Grad-CAM activation heatmap. Figure 4 shows the Grad-CAM visualizations for three representative cases across different models. Which shows a shared failure case where all three models (ResNet-18, EfficientNet-B0, and MobileNetV2) incorrectly predict CNV for an image labeled as DRUSEN. Despite architectural differences, the activation heatmap consistently highlight similar sub-retinal image regions, suggesting that partial structural patterns contribute to this boundary error. Figure 5 and 6 shows both correctly and wrongly classified examples for EfficientNet-B0 and MobileNetV2, respectively. In these cases, Grad-CAM emphasizes illness related regions more distinctly and symmetrically, aligning well with known OCT characteristics. Together, these examples demonstrate that the models rely on similar regions across architectures, and that misclassifications primarily in visually cases rather than from model failures.

Figure 7,8 and 9 shows the confusion matrices for ResNet-18, EfficientNet-B0, and MobileNetV2 on the OCT2017 test set. All three models achieve near-perfect classification performance, with no misclassifications observed for CNV, DME, or NORMAL categories. The only errors consistently occur when DRUSEN samples are predicted as CNV, with the number of such situations decreasing from ResNet-18 to EfficientNet-B0 and MobileNetV2. This shared error pattern means that the confusion from visual similarity between DRUSEN and CNV rather than model-specific weaknesses. Overall, the confusion matrix confirm that the models' highly balanced performance

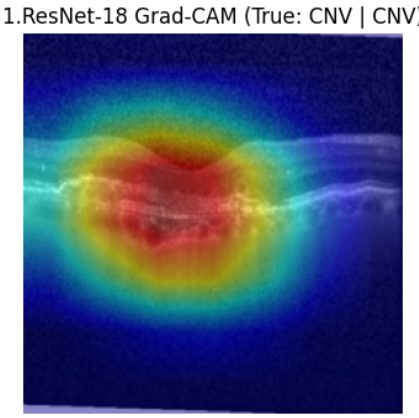
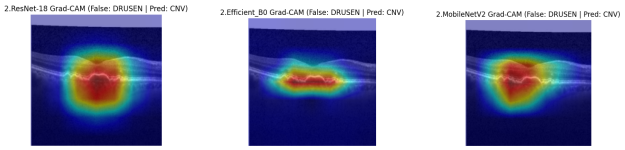


Figure 3: True classification example in ResNet-18



(a) ResNet-18 (b) EfficientNet-B0 (c) MobileNetV2

Figure 4: Wrong classification of all three CNN models.

across classes, while the remaining errors are limited to clinically related boundary cases, which can be imporve when professional doctor supervisioned it.

7 Analysis and Discussion

7.1 Error Analysis and Failure Modes

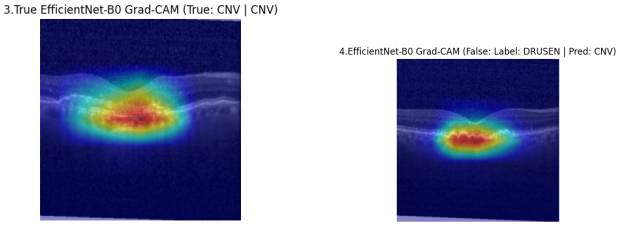
Analysis of the confusion matrices for all three models revealed a high degree of consistency in error patterns. For all three models, only cases of predicting DRUSEN images as CNV were observed; no errors were observed for CNV, DME, or NORMAL samples. This consistency suggests that the failure is not model-specific but reflects the inherent difficulty in distinguishing these two clinically relevant categories.

From a medical imaging perspective, DRUSEN and CNV may exhibit overlapping structural patterns in OCT scans, particularly in the similarity of the lower boundary which is the core of dignosis. While the lower boundary of DRUSEN is regular, the lower boundary of CNV is blury, but overall pattern are similar from human visually inspected.

Therefore, some samples may be located near the decision boundary, and these samples possess inherent blurriness even for trained models. The fact that all three architectures misclassified the same subset of DRUSEN samples suggests that these cases represent hard or borderline samples.

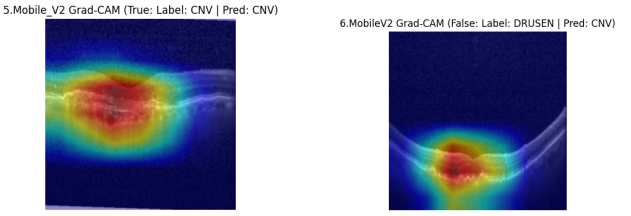
7.2 Interpretation

The nearly identical confusion patterns across different architectures further suggest that, despite differences in model, they rely on similar image feature patterns for classification.



(a) True example of EfficientNet-B0 model (b) False example of EfficientNet-B0 model

Figure 5: Examples for EfficientNet-B0 on the OCT2017.



(a) True example of MobileNetV2 model (b) False example of MobileNetV2 model

Figure 6: Examples for MobileNetV2 on the OCT2017.

Grad-CAM visualizations support this observation, showing that all models focus on similar regions when predicting correct, even for incorrect samples. Therefore, the common failure cases likely rooted from shared feature representations learned from the pre-trained data, rather than architectural limitations.

Importantly, these errors are extremely rare relative to the total size of the test set, with all models maintaining overall performance above 99 percent. This indicates that the learned feature representations are robust to most samples but still hard to handle a few blury cases.

8 Limitations and Ethical Considerations

Despite the overall good experimental results, this project has several significant limitations. First, all experiments were conducted on the OCT2017 dataset, a selected publicly available benchmark dataset. Therefore, the reported

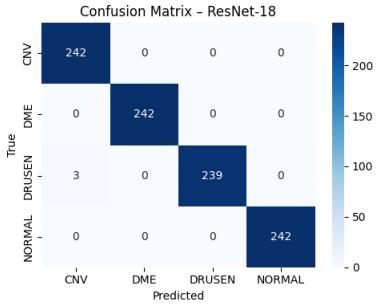


Figure 7: ResNet-18 Confusion Matrix

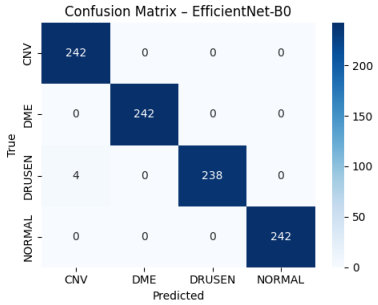


Figure 8: EfficientNet-B0 Confusion Matrix

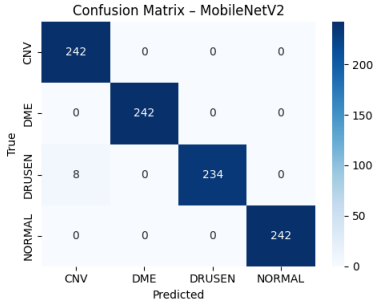


Figure 9: MobileNetV2 Confusion Matrix

results may not be fully generalizable to real world clinical situations, as OCT images vary depending on imaging equipment, imaging methods and patient populations. This potential domain variability can lead to performance degradation when the model is applied to situations outside the training set.

Second, the modeling approach relies on standard transfer learning based on existing CNN model. As observed in the error analysis, certain clinically relevant categories (e.g., CNV and DRUSEN) may has overlapping visual features, leading to unavoidable boundary cases. Without expert doctor involved validation, such misclassifications could be a risk in medical decision making.

From an ethical perspective, deep learning based diagnostic systems should be viewed as decision support tools, not replacements for clinical doctors. Overreliance on model predictions without proper validation could lead to misdiagnosis or delayed treatment. Furthermore, biases present in the dataset may be implicitly learned by the model, affecting fairness among different patient groups.

9 Conclusion and Future Work

This project presents a comparative study of deep learning based approaches for retinal disease classification from OCT images using transfer learning. By evaluating ResNet-18, EfficientNet-B0, and MobileNetV2 under a unified experimental framework, we shows that modern CNN architectures can achieve highly accurate and balanced performance on the OCT2017 dataset. Quantitative results and confusion matrix analysis show near perfect classification, while Grad-CAM visualizations provide interpretable evidence that model predictions are grounded in clinically meaningful retinal structures. Importantly, the consistent failure cases across models highlight intrinsic blurry between certain disease categories rather than models' weakness.

Future work may focus on improving robustness and clinical applicability by incorporating larger and more diverse multicenter datasets, modeling prediction uncertainty for conner cases, and integrating doctor validation. Exploring domain adaptation techniques could further enhance generalization to real-world clinical settings.

References

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Kermany, D. S.; Goldbaum, M.; Cai, W.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR.

Code: https://github.com/Max-CCpersonal/DL-final-project