Group 3 : Justified Referral in AI Glaucoma Screening

Sidhant(2021495), Mohit(2021339), Aryan(2021025)

April 28, 2024

Abstract

This study addresses the development of a machine learning model for the early detection of glaucoma from color fundus photographs (CFPs), aiming to mitigate one of the leading causes of irreversible blindness. We employed a dataset of over 110,000 annotated images to create and validate algorithms that enhance the accuracy and generalizability of glaucoma detection. The model capitalizes on a convolutional neural network (CNN) structure and is trained using innovative preprocessing techniques, including a novel cropping method and optimized grayscale conversion, with a focus on the disease's telltale signs. Our methodology incorporated data augmentation and class rebalancing to address significant class disparities, with a particular emphasis on achieving high model sensitivity. Evaluation metrics such as accuracy, precision, recall, and AUC were utilized to gauge the model's performance, with an ROC curve and confusion matrix providing further insight. Despite achieving an accuracy of over 96% and an AUC of approximately 92The study concludes that while the model shows promising results in training, its variability in validation accuracy calls for refinements to enhance stability and reliability in real-world applications. Through continued development, this model has the potential to revolutionize glaucoma screening and contribute to a significant reduction in vision impairment on a global scale.

1 Keywords

Glaucoma Detection, Convolutional Neural Networks, Fundus Photography, Image Preprocessing, Class Imbalance, Model Generalizability.

2 Introduction

2.1 Background

Glaucoma is one of the leading causes of blindness because it frequently goes untreated until it has already caused serious visual loss. Glaucoma is one of the most prevalent causes of blindness. It is crucial to diagnose the problem at an early stage using imaging techniques such as color fundus photographs (CFPs) and optical coherence tomography (OCT). This is necessary in order to prevent the progression of the disease and to preserve one's eyesight.

2.2 Literature Survey/Related Work

A number of previous research studies have studied the utilization of CFPs for artificial intelligence-based glaucoma screening, and the results have been favorable. On the other side, the performance of AI models typically declines when they are verified on other datasets. This underscores the requirement of building algorithms that are both robust and generalizable in order to address this issue.

2.3 Objective

Through the utilization of a special dataset consisting of more than 110,000 fundus images that have been annotated, we need to create and validate the machine learning algorithms for glaucoma screening purpose. Our objective is to enhance the accuracy and generalizability of artificial intelligence models in the detection of glaucoma and the course of the disease.

2.4 Scope

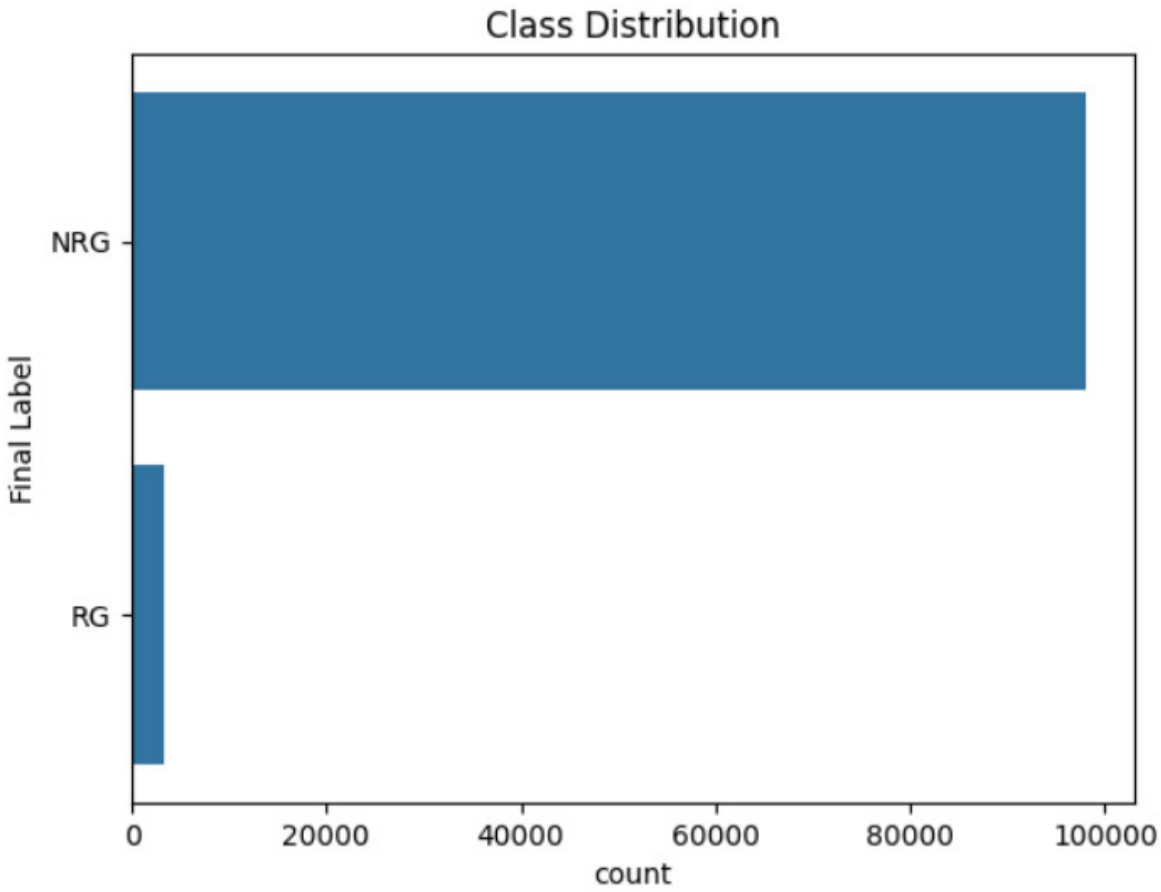
Included in the scope of the challenge is the multi-label categorization of 10 additional glaucomatous features as well as the binary classification of images into "referable glaucoma" or "no referable glaucoma."

2.5 Impact

For the purpose of glaucoma screening, the effective development of machine learning algorithms has the potential to transform both the early diagnosis and management of the condition. Through the implementation of screening methods that are both more precise and more broad, we intend to lessen the impact of glaucoma-related vision impairment and blindness on a global scale.

3 Materials and Methods

3.1 Dataset



The dataset for the Justified Referral in AI Glaucoma Screening (JustRAIGS) challenge consists of a large collection of fundus photographs. The dataset includes over 110,000 carefully annotated fundus images obtained from approximately 60,000 individuals undergoing screening.
Each fundus image is labeled as either "referable glaucoma" (RG) or "no referable glaucoma" (NRG). There are two subsets of the dataset: a training subset with 101,442 gradable fundus images and a test subset with 9,741 fundus images, which is not available to us. Furthermore, all fundus images from referable glaucoma eyes have been extensively annotated with up to ten additional labels associated with various glaucomatous features.
These additional labels are related to specific characteristics or abnormalities observed in the fundus images of glaucoma patients.

3.2 Methodology

3.2.1 Task 1:

Image Preprocessing: Initially, images were resized to 10% of their original dimensions for efficient loading and processing. Subsequently, images underwent cropping to emphasize the central areas of the eye, crucial for glaucoma analysis. Cropping was achieved by identifying the midpoint and selecting the widest region symmetrically around it with a specific aspect ratio. Following cropping, images were converted to grayscale using the OpenCV library to focus on structural changes rather than color variations.

Figure 1: Original Image

Figure 2: Cropped and Grayscaled Image

Data Splitting: The dataset was divided into training, validation, and test sets. Initially, 10% of the data was reserved for the test set with stratification to maintain class balance. The remaining data was further split, allocating 10% for validation. This ensured an equal distribution of glaucoma and non-glaucoma images across all sets.

Data Augmentation: To enhance model robustness and prevent overfitting, data augmentation techniques were applied to the training images using TensorFlow's ImageDataGenerator. This involved rescaling pixel values for normalization and applying transformations such as shifting, zooming, and flipping.

Resampling for Class Imbalance: Due to significant class imbalance (originally 98,000 NRG images to 3,000 RG images), resampling techniques were utilized to balance the classes in the training set. This involved increasing the number of RG images to 15,000 and reducing NRG images to 60,000, resulting in a new class imbalance ratio of 4:1.

Model Selection: Various models were evaluated, and the one with the highest sensitivity at 95% specificity was chosen. These models included custom architectures and pre-trained VGG16 models with different configurations.

Model Training: The selected model was trained using a data generator, with callbacks for early stopping and model checkpoint to enhance training efficiency and prevent overfitting. Training was performed in batches, with metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) monitored during training.

Validation and Performance Evaluation: Validation metrics were calculated for each model, including accuracy, precision, recall, F1-score, and AUC. Sensitivity at 95% specificity was particularly emphasized for model selection, ensuring accurate detection of glaucoma cases.

Early Stopping: We employ callbacks for early stopping and model checkpoint to enhance training efficiency and prevent overfitting. Training occurs in batches of 64 images over a maximum of 30 epochs.

Table 1: Comparison of Model Performances on Validation Data

Model	Precision	Recall	AUC	Sensitivity at Spec. 95%	Description
0	0.2254	0.0556	0.7772	0.2812	Mix of oversampling (15k) and undersampling (60k); custom network with batch normalization and dropout.
1	0.1722	0.4826	0.8375	0.4271	Full oversampling (76k+76k); VGG16 followed by dense layers, focusing on extensive data augmentation for training.
2	0.1979	0.2127	0.4954	0.0463	Mix of oversampling (15k) and undersampling (60k); simpler custom model with convolutional and pooling layers, targeting a balanced approach between classes.
3	0.1119	0.7578	0.8396	0.5917	Mix of oversampling (10k) and undersampling (10k); VGG16 combined with dropout and dense layers, aiming for high sensitivity in detection.
4	0.4558	0.3391	0.9240	0.6851	Mix of oversampling (15k) and undersampling (60k); VGG16 with additional dropout and dense layers, optimized for higher precision and recall balance.

3.2.2 Task 2:

In this task, the preprocessing steps applied to images labeled as "Glaucoma" from the provided CSV file mirror those of Task 1. These steps include resizing, cropping, and conversion to grayscale using the same methods described earlier. Additionally, a function named `select_values` is implemented to organize key information from each row of the dataset based on severity labels of glaucoma (Label G1, Label G2, Label G3). This function first checks if glaucoma is labeled as 'RG' (Referable Glaucoma) under Label G3. If so, it assigns the prefix 'G3' to the features associated with this label. If Label G3 isn't 'RG', the function then checks if Label G1 is filled (not null). If it is, the prefix 'G1' is set, indicating consideration of features associated with this label. If neither condition is met, the prefix 'G2' is used, implying a different or lesser severity that may not require immediate referral. Following organization of the dataset to include necessary labels and transformations, an image processing pipeline is established for the neural network model. TensorFlow's ImageDataGenerator class is utilized to create two generators—one for training and another for validation. These generators play a crucial role in rescaling pixel values of images from a range of 0-255 to 0-1, thereby enhancing numerical stability and model performance.

Model Architecture: The CNN model architecture is constructed using the Keras library with TensorFlow as the backend. It consists of several layers:

- Two pairs of convolutional and max-pooling layers for feature extraction, employing 32 and 64 filters, respectively.
- A flattening layer to transform the 2D feature maps into a 1D feature vector.
- Two dense layers with 128 units each for classification, interleaved with batch normalization and a dropout rate of 0.1 to prevent overfitting.
- An output layer with 10 units and sigmoid activation, suitable for multi-class or multi-label classification scenarios.

The model is compiled with the Adam optimizer, utilizing categorical cross-entropy as the loss function. Evaluation metrics include accuracy, precision, recall, and AUC.

Training Procedure: The model is trained using a data generator, a common approach to handling large datasets and augmentations in image classification tasks. A model checkpoint callback is employed to save model weights with the best validation loss. The training process is monitored through loss graphs, with training loss steadily decreasing while validation loss may indicate overfitting beyond a certain epoch. Validation metrics, including accuracy, precision, recall, F1-score, and AUC, provide insight into the model's performance. Sensitivity at specificities such as 95% is particularly emphasized to ensure the model's effectiveness in detecting glaucoma cases.

3.3 Novelty

Our research introduces several novel methodologies in the domain of early glaucoma detection using machine learning:

- **Custom Image Cropping:** We implemented a novel cropping technique to focus on the central areas of the eye, crucial for glaucoma analysis. This method dynamically identifies the significant content of the fundus images, ensuring symmetrically centered crops while maintaining specific aspect ratios.
- **Optimized Grayscale Conversion:** Our approach to grayscale conversion was optimized for glaucoma detection, emphasizing textural and structural changes rather than color variations. By utilizing the OpenCV library, we transformed RGB images into grayscale, enhancing computational efficiency while preserving diagnostic relevance.
- **Class Imbalance Handling:** To address the significant class imbalance in our dataset, we applied resampling techniques to balance the classes. By increasing the number of glaucoma images and reducing non-glaucoma images, we created a more representative training set, crucial for training unbiased models sensitive to detecting glaucoma.
- **Exploration of Custom and Pre-trained Architectures:** We experimented with various custom and pre-trained convolutional neural network (CNN) architectures tailored for glaucoma detection. This exploration included adapting dense layers, modifying architectures, and integrating pre-trained models such as VGG16. By leveraging both custom and pre-trained architectures, we aimed to optimize model performance and generalizability.
- **AUC as Main Evaluation Metric:** In our training process, we prioritized the area under the curve (AUC) as the primary evaluation metric. AUC serves as a robust measure of a model's ability to differentiate between glaucoma and non-glaucoma cases, offering insights into its discriminative power and suitability for clinical use.

These novel methodologies collectively contribute to advancing the state-of-the-art in early glaucoma detection, offering promising avenues for enhancing model accuracy, generalizability, and clinical relevance.

3.4 Evaluation Metric(s)

3.4.1 Task 1:

We evaluate our model's performance using several key metrics to ensure it's both accurate and reliable for clinical use, particularly in detecting glaucoma. We use accuracy to gauge the overall correctness of the model, precision to assess the accuracy of positive glaucoma detections, and recall to check how well the model identifies all glaucoma cases. The F1-score helps us balance precision and recall, making it a crucial metric when both aspects are equally important. We also calculate the AUC to measure the model's ability to differentiate between patients with and without glaucoma, aiming for a higher score for better performance.

We visually represent our results using a confusion matrix, which shows the breakdown of correct and incorrect predictions, and a ROC curve, which evaluates the model's diagnostic ability at various thresholds. Additionally, we determine a specific threshold that allows the model to achieve 95% specificity, meaning it can accurately identify 95% of non-glaucoma cases, while also assessing how many actual glaucoma cases it can detect at this setting. This comprehensive evaluation ensures the model minimizes the risk of misdiagnosis, making it a reliable tool in clinical settings.

3.4.2 Task 2:

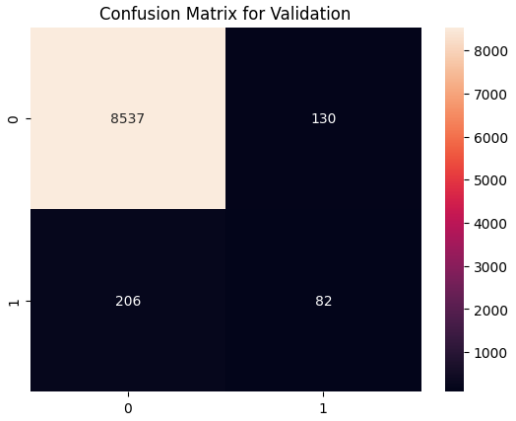
The chosen metrics—accuracy, precision, recall, and AUC—offer a comprehensive assessment of the model's performance. However, specific numerical values for these metrics are not provided in the graph. Graphical Interpretation: The spikes in the validation curves for both loss and accuracy indicate potential issues with the model's training regime. These could be addressed through hyperparameter tuning, implementing cross-validation, or modifying the model's architecture.

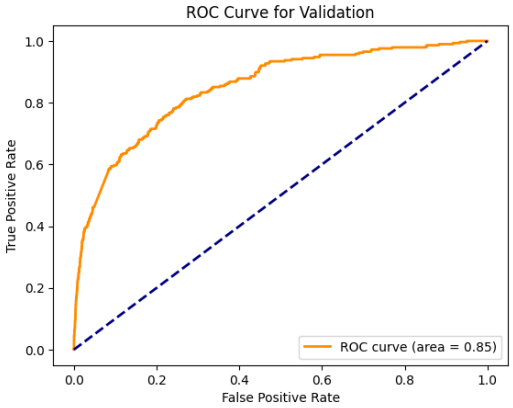
4 Results

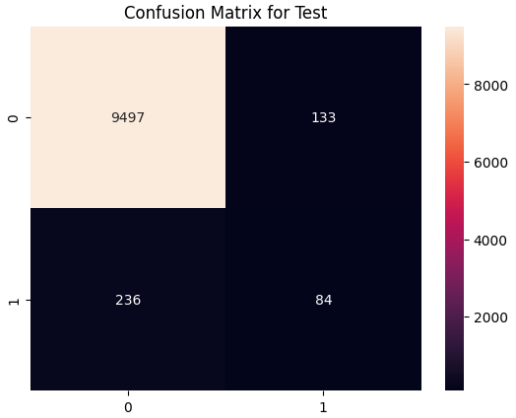
4.1 Task 1:

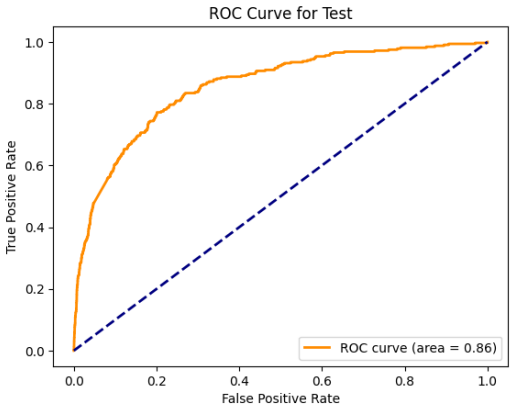
Table 2: Performance Metrics for Validation and Test Data

Metric	Validation Data	Test Data
Accuracy	0.9657	0.9666
Precision	0.4558	0.4818
Recall	0.3391	0.3696
F1-Score	0.3889	0.4183
AUC	0.9240	0.9200
Threshold at Specificity 95.00%	0.2718	0.2718
Sensitivity at Specificity 95.00%	0.6851	0.7205









4.2 Task 2:

Metrics for Training Data (last recorded epochs):

- AUC: Ranges from 0.9388 to 0.9509
- Precision: Approximately 0.67
- Recall: Ranges from 0.9782 to 0.9868
- Accuracy: Increases up to 0.4155
- Loss: Decreases to 5.4754

The model initially appears to perform well on the training data; however, the trends in validation loss and accuracy suggest a need for better regularization techniques to improve generalization.

5 Discussion

5.1 Task 1 Discussion:

The results indicate that while the model achieved high accuracy (96.56% for validation and 96.66% for test data) and a substantial Area Under the Curve (AUC) of around 92%, there are still significant areas for improvement, particularly regarding model sensitivity and the potential issue of overfitting.//

The model's precision, which reflects its ability to correctly identify referable glaucoma cases among those predicted as such, was relatively low at approximately 45.58% for validation and 48.18% for test data. This suggests that while the model is efficient at screening out non-glaucoma cases (as indicated by high specificity), it struggles to correctly identify true glaucoma cases among the positives it flags. This is further highlighted by the recall rates (33.91% for validation and 36.96% for test data), indicating that a significant number of actual glaucoma cases are being missed.//

The relatively low F1-Score, which balances precision and recall, emphasizes the need to improve both metrics to ensure the model is not only selective but also sensitive enough to be clinically useful. The AUC values, though decent, suggest that while the model discriminates between classes to a good extent, there's room for enhancing its capability to differentiate between non-glaucoma and glaucoma cases more distinctly.//

Moreover, the presence of a high accuracy alongside lower precision and recall raises concerns about the model potentially overfitting the training data. Overfitting is likely occurring here due to the model learning specific details from the training set that do not generalize well to unseen data, as suggested by the higher performance on training data compared to validation and test datasets.//

5.2 Task 2 Discussion:

The observed spikes in validation loss and the variations in accuracy during training suggest potential overfitting. The model appears to learn specific details from the training data that do not generalize well, likely due to the complexity and variability of glaucoma features across different images. This issue could be mitigated by:

- Implementing more sophisticated data augmentation techniques to introduce a greater variety of training scenarios.
- Enhancing the dropout rate or introducing other forms of regularization to reduce the model's complexity.
- Using cross-validation during training to ensure that the model performs consistently across different subsets of the data.

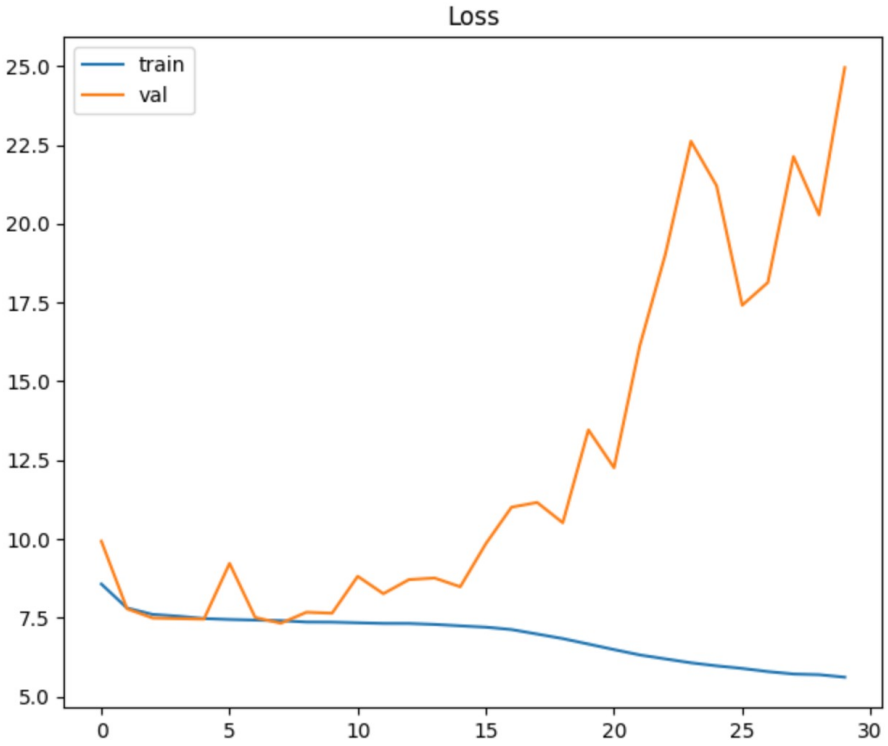


Figure 3: Graphical representation of overfitting issues and corrective actions for Task 2.

6 Work to be done

Ensembling Techniques: Implementing ensemble methods, such as bagging, boosting, or stacking, could enhance model performance by combining predictions from multiple base models. Ensembling can help mitigate the effects of overfitting and improve overall predictive accuracy by leveraging the diversity of individual models.//

Exploration of Other Pre-trained Models: While the study explored custom architectures and the VGG16 model, further investigation into other pre-trained convolutional neural network (CNN) architectures, such as ResNet, Inception, or EfficientNet, could uncover models better suited for glaucoma detection. Experimentation with different architectures may reveal superior features for extracting glaucomatous characteristics from fundus images. // Validation on External Datasets: To validate the generalizability of the model, testing on external datasets beyond the JustRAIGS challenge dataset is imperative. Evaluating the model's performance on diverse datasets acquired from different populations, imaging devices, and clinical settings can provide insights into its real-world utility and robustness across varied scenarios. // Clinical Validation and Integration: Collaborating with ophthalmologists and healthcare professionals for clinical validation of the model's predictions is essential before integrating it into clinical workflows. Conducting prospective studies to assess the model's performance in real-world clinical settings can validate its efficacy and safety as a diagnostic tool for glaucoma screening.//

7 Distribution of work among group members

Throughout the project, each member of the team contributed significantly to various aspects of the model development, training, evaluation, and documentation. While the overall contributions were nearly equal, the following highlights the major contributions of each member:

- Mohit: Data Preprocessing, Model Building, Model Training for Task 1, Model Evaluation.
- Sidhant: Model Building, Model Training for Task 1, Model Evaluation, Report Writing.

- Aryan: Handling Image Formats, Model Training for Task 1, Model Training for Task 2.

By leveraging the collective expertise and contributions of each team member, we successfully developed and evaluated multiple machine learning models with varied sampling techniques and model architectures, but finalized 5 best models as stated above, out of which we chose the 4th model as our best model and worked on it for Task 1 and also used it for Task 2.

References

[1] ResearchGate, *Examples of the original and cropped fundus images. All images were labeled as normal or*, https://www.researchgate.net/figure/Examples-of-the-original-and-cropped-fundus-images-All-images-were-labeled-as-normal-or_fig1_348666988

[2] Medium, *Use Weighted Loss Function to Solve Imbalanced Data Classification Problems*, by *Zergtant*, https://medium.com/@zergtant/use-weighted-loss-function-to-solve-imbalanced-data-classification-problems-749237f38b75