Multi-label Retinal Disease Classification on a High-Class Imbalanced Fundus Image Dataset

Garth Dustin Ayang-ang, Church Chill Parco, Kirk Patrick Pattawi, Danica Joy Tejano

Information Technology Education Program
School of Engineering, Architecture, and Information Technology Education, University of Saint Louis
Tuguegarao City, Cagayan

***Abstract*— Multi-label image classification is capable of providing multiple diagnoses for a single retinal fundus image. In this research, we used the Classification Transformer, a general framework that exploits transformers to learn the complex dependencies between visual features and category labels. A new multi-label retinal fundus image dataset, the Ocular Disease Intelligent Recognition ODIR-5K, was used. The transformer-based model for fundus multi-label disease classification was optimized through extensive experimentation for image analysis and disease classification. In this work, we also addressed the class imbalance of the dataset using the weighted loss function PolyLoss and the oversampling method Local Perturbation Random Over-Sampling algorithm which has a model score of 81.3% on 10% resampling. It is shown that the approach outperforms previous methods with an Area Under the Curve score of 90.2%.**

Keywords*— multi-label classification, PolyLoss, LP ROS, class imbalance, retinal fundus image

I. INTRODUCTION

Millions of individuals worldwide have been significantly affected by the global health issues of retinal disease and vision impairment, with uncorrected refractive errors and cataracts being the leading causes [1]. Retinal diseases refer to a group of eye conditions that affect the retina, a vital part of the eye responsible for vision. These diseases, such as glaucoma, diabetic retinopathy (DR), age-related macular degeneration (ARMD), and others, initially show symptoms that affect the retina and can eventually lead to blindness [2]. In India, approximately 75% of these cases could have been cured if they had been detected at an earlier stage [3]. However, many developing nations and underdeveloped regions face a lack of ophthalmologists specializing in diagnosing and treating fundus diseases, resulting in inadequate access to timely treatment for affected patients [4].

The rapid advancements in artificial intelligence (AI) have revolutionized the medical industry, particularly in diagnostic imaging, through the development of computer-aided diagnosis (CAD) systems [5]. These systems utilize machine learning techniques to analyze patient conditions, providing valuable insights to medical professionals for decision-making [6]. While numerous studies have been conducted to develop CAD systems for retinal disease classification using deep learning techniques, existing models often lack inclusivity in disease classifications and struggle to provide multiple diagnoses for a single image [7]. Addressing these limitations requires innovative approaches to overcome challenges such as class imbalance within datasets, which can significantly affect model performance [8].

In this study, the resampling method has offered a promising solution to this problem, specifically the use of two resampling techniques called LP ROS (Local Perturbation Random Over-Sampling) and PolyLoss. Over-sampling methods, which increased the representation of the minority class by duplicating existing samples or synthesizing new ones, helped to balance the class distribution and enhance the model's performance on underrepresented diseases [2]. Resampling techniques were applied to address the problem of dataset imbalance in the study of Obaid et al. [9]. The study used a dataset containing 400 observations and 4 variables. The data distribution was imbalanced, with 70% of the data belonging to one class and the remaining 30% belonging to the other class. The results showed that the classifiers' accuracy increased after treating the problem of imbalance, demonstrating the effectiveness of resampling techniques in improving the performance of classifiers on imbalanced datasets.

The application of resampling techniques in retinal disease classification holds promising potential to address the issue of class imbalance. By balancing the class distribution, these methods enhance the performance of the Classification Transformer (C-TRAN) Architecture model in accurately classifying multi-labeled retinal diseases. This approach paved the way for more accurate and reliable diagnosis of retinal diseases, ultimately contributing to improved patient outcomes, s it provides ophthalmologists with a valuable tool for diagnosing diseases through fundus image analysis, facilitating rapid and accurate assessments. Additionally, by mitigating the class imbalance problem and enhancing the balance between majority and minority classes, the study improves the robustness of classification models, thereby ensuring more precise diagnoses and better treatment decisions for patients.

II. METHODS

Throughout the entire project, the researchers used Python3. The researchers utilized several Python libraries such as PyTorch, sci-kit-learn, Tensorflow, and NumPy among others. To train the model with a higher RAM, video RAM, and stronger processors, the researchers made use of the Google Colab platform with a Pro subscription.

A. Dataset

The researchers used the Ocular Disease Intelligent Recognition (ODIR-5K) dataset, which is a benchmark collection of 5000 structured fundus images, for classifying multiple diseases in fundus images using a multi-label approach. The dataset came from patients who underwent ocular health examinations in hospitals and medical institutions, where eye disease diagnostic keywords are assigned by retinal specialists [10]. The images are grouped into eight disease classes, Normal (N), Diabetes (D), Glaucoma (G), Cataract (C), age-related macular degeneration (A), Hypertension (H), Myopia (M), and other abnormalities/diseases (O). The dataset is highly imbalanced, as it clearly shows when considering the number of images in each of the classes. Though more challenging and greatly decreases the accuracy and loss of the trained models, it is more applicable to real-life clinical situations. Figure 1 shows the distribution of the dataset represented in a bar chart. Figure 2 shows a sample of images of the dataset. Table I shows more detailed information on the dataset distribution.

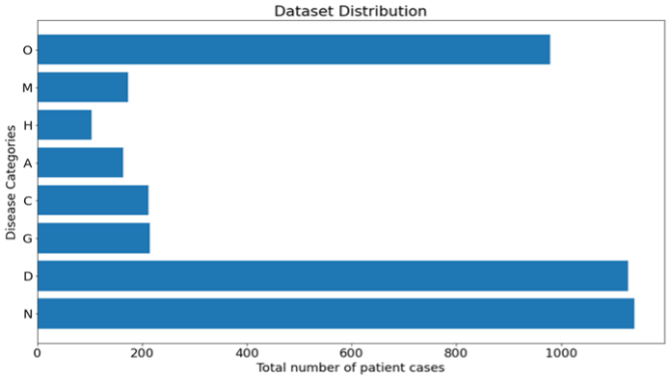

Fig. 1. Distribution of the ODIR-5K dataset

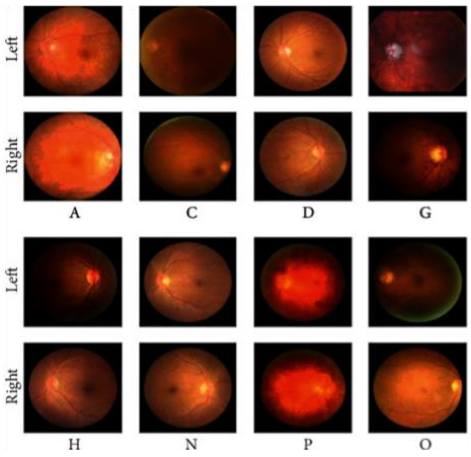

Fig. 2. Sample of left and right retinal images with their corresponding disease label.

TABLE I. DISTRIBUTION OF TRAINING IMAGE

Labels	Training Cases
Normal (N)	1135
Diabetes (D)	1131
Glaucoma (G)	207
Cataract (C)	211
Age-related macular degeneration (A)	171
Hypertension (H)	94
Pathological myopia (M)	177
Other diseases/abnormalities (O)	944

B. Data Preprocessing

Commonly, dimensions of input images in neural networks are in an aspect ratio of 1:1. Image cropping was utilized to transform the images for appropriate model training. To ensure compatibility with various DNN models, it is necessary to adjust the image size accordingly. As a widely accepted size by many DNN models, the cropped images were kept at 224 x 224 pixels.

C. Multi-label Classification

The C-Tran architecture, proposed by Lanchantin et al. (2021) [11], was selected as the classification model. This model is particularly designed for multi-label tasks and has shown impressive performance rates on widely-used multi-label datasets such as MS-COCO [12], which consists of 80 categories in its multi-label version, and Visual Genome [13], which involves the 500 most frequently occurring categories. In the study proposed by Lanchantin et al. (2021) [11], the best-performing backbone model was DenseNet 161, achieving an AUC score of 95.7% and a model score of 89.8% which was trained using the MuRed Dataset. Table II shows the results of using different backbone models as feature extractors taken from the study of Lanchantin et al. With DenseNet having the best results, the researchers made use of this backbone model for the following experimentations.

TABLE II. COMPARISON OF DIFFERENT BACKBONES

Backbone	ML F1	ML mAP	ML AUC	ML Score	Bin AUC	Bin F1	Model Score
Inception V3	0.469	0.569	0.933	0.751	0.951	0.755	0.851
EfficientNetB5	0.501	0.625	0.943	0.784	0.965	0.825	0.874
EfficientNetB6	0.504	0.627	0.946	0.787	0.964	0.789	0.875
WideResNet101	0.537	0.638	0.945	0.791	0.960	0.794	0.876
VGG16	0.508	0.622	0.940	0.781	**0.977**	**0.837**	0.879
EfficientNetV2-M	0.570	0.683	0.955	0.819	0.958	0.781	0.889
EfficientNetV2-L	0.585	0.680	0.954	0.817	0.961	0.806	0.889
ResNext101 32x4d	0.585	0.677	0.953	0.815	0.964	0.802	0.889
ResNext101 32x8d	**0.612**	0.683	0.947	0.815	0.966	0.785	0.890
ResNext101	**0.612**	**0.689**	0.955	0.822	0.970	0.808	0.896
DenseNet161	0.595	**0.689**	**0.957**	**0.823**	0.973	0.822	**0.898**

D. Diversity and Class Balancing

During model training, different random augmentations were used to increase the diversity of the samples on each batch. This approach follows the setup used in the study of Lanchantin et al. (2021). To implement this, the project made use of the albumentations Python library. Table III lists the different augmentation algorithms used.

TABLE III. AUGMENTATION USED DURING MODEL TRAINING

Augmentation	Description	Parameters	Probability
HorizontalFlip	Flips the image horizontally	-	0.5
VerticalFlip	Flips the image vertically	-	0.5
Rotate	Rotates the image around by an angle.	limit=30	0.5
MedianBlur	Applies a median filter to the image	blur_limit=7	0.3
GaussNoise	Applies Gaussian noise to the image.	var_limit=(0.38)	0.5
HueSaturationValue	Applies random changes to the hue, saturation, and value of the image.	hue shift limit=10, sat shift limit=10, val shift limit=10	0.3
RandomBrightness Contrast	Applies random brightness and contrast of the image.	brightness limit=(-0.2, 0.2), contrast limit=(-0.2, 0.2)	0.3
Cutout	Randomly crops square regions on the image	max h size=20, max w size=20, num cutout regions=5	0.5

As previously mentioned, multi-label classification models suffer from class imbalances of datasets. Therefore, as an intermediary step, the researchers implemented conventional strategies to address class imbalances. The ODIR-5K dataset is highly imbalanced, and as such, the researchers made use of two popular approaches, weighted loss functions and resampling methods.

Training detection models are consistently difficult due to class imbalances [14]. A common technique to address class imbalance is through the use of weighted loss functions. The goal of weighted loss functions is to help the model pay more attention to the less common group by making the cost of making mistakes about that group higher [15]. One of the most common weighted loss functions is the PolyLoss, proposed by Leng et al. (2023) [14]. The PolyLoss weighted loss function was utilized in the study of Lanchantin et al. (2021) [11]. In their experiments, the weighted loss function with the best model score was BCE (Binary Cross Entropy) and PolyLoss, which both achieved a model score of 89.8%. The researchers concluded to continue with the PolyLoss weighted loss function.

In this first experiment, the study made use of oversampling methods to improve the class distribution of the dataset. In the study of Lanchantin et al. (2021) [11], the best-performing oversampling method was LP ROS, with a model score of 89.9 on 10% resampling. The LP ROS algorithm over-samples data by taking the label set of the dataset and generates P% of the original number of images in the dataset [16]. For this reason, the study made use of the LP ROS algorithm for class imbalance. Then, the study made a comparison of results using different values for the parameter P to pass into the oversampling algorithm. For this part, the model used the DenseNet161 as the backbone model, along with a Learning Rate of 0.0001, the Adam optimizer, a batch size of 32, a maximum epoch of 40, and the Polynomial Loss as the weighted loss function. The result with the best model score was used for the following experiments.

The dataset is split into validation and training sets, with 80% of the dataset put into the validation set and 20% for the training set. Table IV shows the details of the classes and the number of samples per class for both the resampled training dataset and the validation set.

TABLE IV. ODIR-5K DATASET

Augmentation	Full Name	Training	Validation	Total
N	Normal	2299	574	2873
D	Diabetes	1287	321	1608
G	Glaucoma	315	56	371
C	Cataract	321	59	380
A	Age-related macular degeneration	299	54	353
H	Hypertension	189	26	215
M	Pathological myopia	273	46	319
O	Other diseases/ abnormalities	640	142	602

E. Performance Metrics

The chosen model's performance was assessed using the scoring metric recommended in the RIADD challenge [17], which places equal emphasis on accurately detecting the existence of the disease and correctly categorizing it. The proposed method similarly follows the sets of experiments and performance evaluation in the study of Rodriguez, et al. [2]. The F1, mAP, and AUC scores are calculated for all labels in the dataset, given by the equations:

$$AP = \sum_{i=0}^{|T|-1} [recall_i - recall_{i+1}] \times precision_i$$

$$ML\_mAP = \frac{1}{|T|} \sum_{i=1}^{|T|} AP_i$$

$$ML\_F1 = \frac{1}{|T|} \sum_{i=1}^{|T|} F1_i$$

$$ML\_AUC = \frac{1}{|T|} \sum_{i=1}^{|T|} AUC_i$$

Using the scores from set T, defined as the set of labels representing a single retinal disease label, the average score for each metric of the disease classes is calculated and named: ML_mAP, ML_F1, and ML_AUC.

These metrics are used to calculate the two most important metrics for performance evaluation: ML_SCORE and MODEL_SCORE. ML_SCORE is the average score of ML_mAP and ml_AUC. MODEL_SCORE is the average of the ML_SCORE and the AUC score of the Normal class (termed Bin_AUC). The formula for these two metrics is shown below:

$$ML\_Score = \frac{ML\_mAP + ML\_AUC}{2}$$

$$Model\_Score = \frac{ML\_Score + Bin\_AUC}{2}$$

The final metric, bin_f1, represents the f1-score of the normal label.

These metrics will be used to determine the best-performing backbone model as feature extractors for the c-tran architecture. Moreover, these metrics will be used to compare the results of different resampling algorithms to be used, as well as different image sizes.

F. Optimal Model Configuration

After determining the best value for the parameter p for the lp ros resampling algorithm, the final experiment focused on finding the optimal model configuration by testing different image sizes and batch sizes.

The image sizes tested out were 224 x 224, 384 x 384, and 448 x 448. For the batch size, three sets of model training were performed with batch size values of 16, 32, and 64.

G. Comparison of results from different approaches

Upon determining the optimal model configuration, a comparison among different proposed approaches for multi-label classification was implemented to analyze performance differences.

The researchers have identified other research that can be fairly compared with the results of the proposed model. The researchers selected different studies that performed multi-label classification problems on the odir-5k dataset. Because the source code for these chosen studies was not readily available online, the researchers were unable to replicate the scoring measures that were used in this study. Hence, the results were compared according to their AUC score, which was available in all the studies.

III. Results and Discussion

A. Class Imbalance

Using the base model configuration discussed in the methodology section of the study, the researchers were able to determine the best percentage value for the LP ROS resampling algorithm. Table V shows the comparison of different results from using different percentage values of the parameter P for the LP ROS resampling algorithm.

TABLE V. LP ROS RESEAMPLING ALGORITHM RESULTS

Algorithm	ML F1	ML mAP	ML AUC	ML Score	Bin AUC	Bin F1	Model Score
LP ROS 10%	0.597	0.692	0.902	0.797	0.829	0.5741	0.813
LP ROS 20%	0.597	0.642	0.892	0.767	0.829	0.5741	0.7980
LP ROS 30%	0.623	0.653	0.886	0.770	0.829	0.585	0.7990
LP ROS 40%	0.597	0.595	0.872	0.734	0.829	0.571	0.7815

In this first set of experiments, the results show that increasing the resampling percentage does not yield a higher model score. Thus, it was concluded that LP ROS with 10% resampling yielded better results than those with higher percentage values. The following experiments will employ LP ROS 10% in the following experiments.

B. Optimal Model Configuration

In this section, two sets of experiments were conducted to determine the optimal model configuration for hyperparameters. First, the researchers determined the best-performing image size. Table VI shows the comparison of results for different image sizes.

TABLE VI. RESULTS FOR DIFFERENT IMAGE SIZES

Image Size	ML F1	ML mAP	ML AUC	ML Score	Bin AUC	Bin F1	Model Score
224x224	0.597	0.692	0.902	0.797	0.829	0.5741	0.813
384x384	0.581	0.681	0.881	0.781	0.811	0.5601	0.796
448x448	0.579	0.684	0.885	0.7845	0.815	0.5632	0.800

From the table above, it can be seen that increasing the image size does not improve the overall performance of the model. Thus, training with an image size of 224 x 224 pixels was used for the following experiment.

The second set of experiments aimed to find the optimal batch size, starting with a batch size of 16 until 64. In this incremental approach, batch size is doubled in every experiment. Table VII shows the comparison of results for different batch sizes.

TABLE VII. COMPARISON OF RESULTS FOR DIFFERENT BATCH SIZES

Batch Size	ML F1	ML mAP	ML AUC	ML Score	Bin AUC	Bin F1	Model Score
16	0.597	0.680	0.888	0.784	0.822	0.5121	0.803
32	0.597	0.692	0.902	0.797	0.829	0.5741	0.813
64	0.610	0.671	0.881	0.776	0.817	0.5510	0.7965

The results shown in Table VII indicate that a batch size of 32 still performs the best. It was analyzed that configuring the batch size does not largely impact the model's performance. Thus, it was concluded that the batch size of 32 is the optimal batch size.

C. Comparison of Results from Different Approaches

The final experiment aimed to compare the different approaches for multi-label classification. The comparison of results is shown in the table below.

TABLE VIII. RESULT COMPARISON FOR THE DIFFERENT APPROACHES FOR MULTI-LABEL CLASSIFICATION

Author	Area Under the Curve (AUC)
Islam et al. (2019) [17]	80.5
Wang et al. (2020) [13]	73
Gour and Khanna (2020) [18]	84.93
Li et al. (2021) [19]	88
Lin et al. (2021) [20]	78.16
Proposed Method	**90.2**

The results show that the C-Tran method does better than earlier methods that use cnn architectures by a considerable amount. This indicates that the transformer-based approach is superior to the odir-5k dataset.

To better understand how well the C-Tran model performs, the researchers calculated various measures for each label class. Table IX displays the metric for each class.

TABLE IX. METRICS FOR EACH CLASS

Class	Precision	Recall	F1	AUC
N	0.680	0.789	0.731	0.826
D	0.722	0.477	0.574	0.829
G	0.676	0.446	0.538	0.947
C	0.867	0.881	0.874	0.988
A	0.683	0.519	0.589	0.951
H	0.500	0.154	0.235	0.735
M	0.927	0.826	0.874	0.995
O	0.995	0.239	0.338	0.800

From Table IX, it can be seen that the AUC scores for most classes performed very well. However, the label H (Hypertension), only achieved an AUC score of 0.735. This can be attributed to the class imbalance which remains a problem in the study. Other approaches are recommended to address the class imbalance problem. The AUC score for the label "Others" (O) is only 0.800. It is challenging to make accurate predictions for this class since it is an "umbrella" class that covers diseases not listed or not in their labels in the dataset.

IV. CONCLUSION

The C-Tran architecture proposed in this work was used for multi-label classification, trained on an imbalanced retinal fundus dataset. The ODIR-5K dataset has a total of 6392 images, with left and right eye pairing, for eight common retinal diseases. Before training, preprocessing was necessary for uniformity of data and to ensure a better model performance upon training. The study also proposed to use two common approaches to address the class imbalance problem, namely weighted loss functions and resampling algorithms. In this first experiment, it was found that the resampling algorithm, LP ROS, with a resampling percentage of 10% performed the best on all scoring metrics, achieving a model score of 81.3%. In determining the optimal model configuration, increasing the image size or the batch size did not yield better results in terms of the model score metric.

In terms of future research, the researchers will aim to find more novel approaches to dealing with the class imbalance problem. One suggestion discussed by the researchers is to use alternative models for data generation from the given training dataset. Further research and experiments are indeed necessary to obtain better model performance using the C-Tran architecture.

REFERENCES

[1] World Health Organization. (2022, October 13). Blindness and vision impairment. Retrieved from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[2] Rodriguez, M. A., AlMarzouqi, H., & Liatsis, P. (2022). Multi-label Retinal Disease Classification Using Transformers. IEEE Journal of Biomedical and Health Informatics, 1–13. Retrieved from https://doi.org/10.1109/jbhi.2022.3214086

[3] Sattigeri, S. K., Harshith N., Gowda, D. N., Ullas, K. A., & Aditya M. S. (2022). Eye Disease Identification Using Deep Learning. International Research Journal of Engineering and Technology (IRJET), 9(7), pp. 974-978. Retrieved from https://www.irjet.net/archives/V9/i7/IRJET-V9I7185.pdf

[4] Sun, G., Wang, X., Xu, L., Li, C., Wang, W., Yi, Z., Luo, H., Su, Y., Zheng, J., Li, Z., Chen, Z., Zheng, H., & Chen, C. (2022). Deep Learning for the Detection of Multiple Fundus Diseases Using Ultra-widefield Images. Ophthalmology and Therapy, 12(2), 895–907. Retrieved from https://doi.org/10.1007/s40123-022-00627-3

[5] Fujita H. (2020). AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiological physics and technology, 13(1), 6–19. Retrieved from https://doi.org/10.1007/s12194-019-00552-4

[6] Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. Expert Systems with Applications, 138, 112821. Retrieved from https://doi.org/10.1016/j.eswa.2019.112821

[7] Asiri, N., Hussain, M., Al Adel, F., & Alzaidi, N. (2019). Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. Artificial Intelligence in Medicine, 99, 101701. Retrieved from https://doi.org/10.1016/j.artmed.2019.07.009

[8] Guo, C., Yu, M., & Li, J. (2021). Prediction of Different Eye Diseases Based on Fundus Photography via Deep Transfer Learning. Journal of Clinical Medicine, 10(23), 5481. Retrieved from https://doi.org/10.3390/jcm10235481

[9] W. Obaid and A. B. Nassif, "The Effects of Resampling on Classifying Imbalanced Datasets," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9735021.

[10] ODIR-2019 - Grand Challenge. (2019). Grand-Challenge.org. Retrieved from https://odir2019.grand-challenge.org/, accessed: April 12, 2023.

[11] Lanchantin, J., Wang, T., Ordonez, V., Qi, Y. (2021). General multi-label image classification with transformers. Institute of Electrical and Electronics Engineers. DOI: 10.1109/CVPR46437.2021.01621

[12] Zhou, W., Dou, P., Su, T., Hu, H., Zheng, Z. (2023). Feature learning network with transformer for multi-label image classification. Elsevier Science Inc., 136(C). Retrieved from https://doi.org/10.1016/j.patcog.2022.109203

[13] Wang, J., Liu, Y., Huo, Z., He, W., & Liu, J. (2020). Multi-Label Classification of Fundus Images With EfficientNet. IEEE Access, 8, 212499–212508. Retrieved from https://doi.org/10.1109/access.2020.3040275

[14] Leng, Z., Tan, M., Liu, C., Dogus Cubuk, E., Shi, X., Cheng, S., & Anguelov, D. (n.d.). POLYLOSS: A POLYNOMIAL EXPANSION PERSPEC- TIVE OF CLASSIFICATION LOSS FUNCTIONS. Retrieved November 18, 2023, from https://arxiv.org/pdf/2204.12511.pdf

[15] Johnson, J.M., Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. J Big Data 6, 27. Retrieved from https://doi.org/10.1186/s40537-019-0192-5

[16] Retinal image analysis for multi-disease detection. Retrieved from https://riadd.grand-challenge.org/Home/. Accessed May 3, 2023.

[17] Md. Tariqul Islam, Sheikh Asif Imran, Asiful Arefeen, Hasan, M., & Shahnaz, C. (2019). Source and Camera Independent Ophthalmic Disease Recognition from Fundus Image Using Neural Network. IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON). Retrieved from https://doi.org/10.1109/spicscon48833.2019.9065162

[18] Gour, N., & Khanna, P. (2020). Multi-class multi-label ophthalmological disease detection using transfer learning-based convolutional neural network. Biomedical Signal Processing and Control 66(3), 102329. Retrieved from https://www.researchgate.net/publication/347480372_Multi-class_multi-label_ophthalmological_disease_detection_using_transfer_learning_based_convolutional_neural_network

[19] Li, N., Li, T., Hu, C., Wang, K., & Kang, H. (2021). A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection. Lecture Notes in Computer Science, 177–193. https://doi.org/10.1007/978-3-030-71058-3_11

[20] Lin, J., Cai, Q., & Lin, M. (2021). Multi-Label Classification of Fundus Images With Graph Convolutional Network and Self-Supervised Learning. IEEE Signal Processing Letters, 28, 454–458. https://doi.org/10.1109/lsp.2021.3057548