German Medical Data Sciences 2021: Digital Medicine: Recognize - Understand - Heal
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210537

Multi-Disease Detection in Retinal Imaging Based on Ensembling Heterogeneous Deep Learning Models

Dominik MÜLLER[a,1], Iñaki SOTO-REY[a,b], and Frank KRAMER[a]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Augsburg, Germany*
[b] *Medical Data Integration Center, Institute of Digital Medicine, University Hospital Augsburg, Augsburg, Germany*

Abstract. Preventable or undiagnosed visual impairment and blindness affect billion of people worldwide. Automated multi-disease detection models offer great potential to address this problem via clinical decision support in diagnosis. In this work, we proposed an innovative multi-disease detection pipeline for retinal imaging which utilizes ensemble learning to combine the predictive capabilities of several heterogeneous deep convolutional neural network models. Our pipeline includes state-of-the-art strategies like transfer learning, class weighting, real-time image augmentation and Focal loss utilization. Furthermore, we integrated ensemble learning techniques like heterogeneous deep learning models, bagging via 5-fold cross-validation and stacked logistic regression models. Through internal and external evaluation, we were able to validate and demonstrate high accuracy and reliability of our pipeline, as well as the comparability with other state-of-the-art pipelines for retinal disease prediction.

Keywords. Retinal Disease Detection, Ensemble Learning, Class Imbalance, Multi-label Image Classification, Deep Learning

1. Introduction

Even if the medical progress in the last 30 years made it possible to successfully treat the majority of diseases causing visual impairment, growing and aging populations lead to an increasing challenge in retinal disease diagnosis [1]. The World Health Organization (WHO) estimates the prevalence of blindness and visual impairment to 2.2 billion people worldwide, of whom at least 1 billion affections could have been prevented or is yet to be addressed [2]. Early detection and correct diagnosis are essential to forestall disease course and prevent blindness.

The use of clinical decision support (CDS) systems for diagnosis has been increasing over the past decade [3]. Recently, modern deep learning models allow automated and reliable classification of medical images with remarkable accuracy comparable to physicians [4]. Nevertheless, these models often lack capabilities to detect rare

[1] Corresponding Author, Dominik Müller, IT Infrastructure for Translational Medical Research, Alter Postweg 101, 86159 Augsburg, Germany; E-mail: dominik.mueller@informatik.uni-augsburg.de.

pathologies such as central retinal artery occlusion or anterior ischemic optic neuropathy [5,6].

In this study we push towards creating a highly accurate and reliable multi-disease detection pipeline based on ensemble, transfer and deep learning techniques. Furthermore, we utilize the new Retinal Fundus Multi-Disease Image Dataset (RFMiD) containing various rare and challenging conditions to demonstrate our detection capabilities for uncommon diseases.

2. Methods

The implemented medical image classification pipeline can be summarized in multiple core steps, which are illustrated in Figure 1.

2.1. Retinal Imaging Dataset

The RFMiD dataset consists of 3,200 retinal images for which 1,920 images were used as training dataset [7]. The fundus images were captured by three different fundus cameras having a resolution of 4288x2848 (277 images), 2048x1536 (150 images) and 2144x1424 (1,493 images), respectively.

The images were annotated with 46 conditions, including various rare and challenging diseases, through adjudicated consensus of two senior retinal experts. These 46 conditions are represented by the following classes, which are also listed in Table 1: An overall normal/abnormal class, 27 specific condition classes and 1 'OTHER' class consisting of the remaining extremely rare conditions. Besides the training dataset, the organizers of the RIADD challenge hold 1,280 images back for external validation and testing datasets to ensure robust evaluation [7,8].

2.2. Preprocessing and Image Augmentation

In order to simplify the pattern finding process of the deep learning model, as well as to increase data variability, we applied several preprocessing methods.

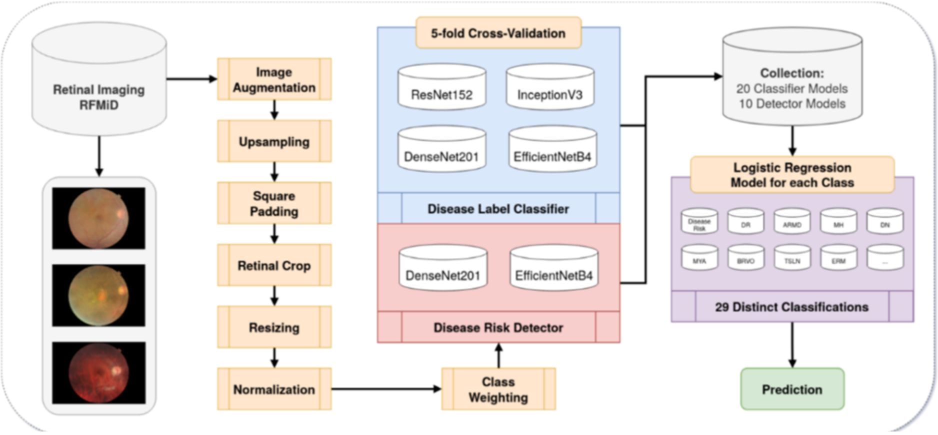

Figure 1. Flowchart diagram, which is of the implemented retinal disease detection pipeline starting with the retinal imaging dataset (RFMiD) and ends with computed predictions for novel images.

Table 1. Annotation frequency for each class in the dataset. Full disease names of all class acronyms in the RFMiD dataset can be found in the appendix.

Disease	Samples	Disease	Samples	Disease	Samples
D. Risk	1,519	DR	376	ARMD	100
MH	317	DN	138	MYA	101
BRVO	73	TSLN	186	ERM	14
LS	47	MS	15	CSR	37
ODC	282	CRVO	28	TV	6
AH	16	ODP	65	ST	5
AION	17	PT	11	RT	14
RS	43	CRS	32	EDN	15
RPEC	22	MHL	11	RP	6
OTHER	34				

We utilized extensive image augmentation for up-sampling to balance class distribution and real-time augmentation during training to obtain novel and unique images in each epoch. The augmentation techniques consisted of rotation, flipping, and altering in brightness, saturation, contrast and hue. Through the up-sampling, it was ensured that each label occurred at least 100 times in the dataset which increased the total number of training images from 1,920 to 3,354.

Afterwards, all images were square padded in order to avoid aspect ratio loss during posterior resizing. The retinal images were also cropped to ensure that the fundus is center located in the image. The cropping was performed individually for each microscope resolution and resulted in the following image shapes: 1424x1424, 1536x1536 and 3464x3464 pixels. The images were then resized to model input sizes according to the neural network architecture, which was 380x380 for EfficientNetB4, 299x299 for InceptionV3 and 244x244 for all remaining architectures [9-12].

Before feeding the image to the deep convolutional neural network, we applied value intensity normalization as last preprocessing step. The intensities were zero-centered via the Z-Score normalization approach based on the mean and standard deviation computed on the ImageNet dataset [13].

2.3. Deep Learning Models

The state-of-the-art for medical image classification are the unmatched deep convolutional neural network models [4,14]. Nevertheless, the hyper parameter configuration and architecture selection are highly dependent on the required computer vision task [4,15]. Thus, our pipeline combines two different types of image classification models: The disease risk detector for binary classifying normal/abnormal images and the disease label classifier for multi-label annotation of abnormal images.

Both model types were pretrained on the ImageNet dataset [13]. For the fitting process, we applied a transfer learning training, with frozen architecture layers except for the classification head, and a fine-tuning strategy with unfrozen layers. Whereas the transfer learning fitting was performed for 10 epochs using the Adam optimization with an initial learning rate of 1-E04, the fine-tuning had a maximal training time of 290 epochs and using a dynamic learning rate for the Adam optimization starting from 1-E05 to a maximum decrease to 1-E07 (decreasing factor of 0.1 after 8 epochs without improvement on the monitored validation loss) [16]. Furthermore, an early stopping and model checkpoint technique was utilized for the fine-tuning process, stopping after 20 epochs without improvement (after epoch 60) and saving the best model measured according to the validation loss. Instead of defining an epoch as a cycle through the full training dataset, we establish an epoch to have 250 iterations. The images for a batch were randomly drawn, considering that as many samples as possible are used based on the number of iterations. This allowed to increase the number of seen batches and, thus, to increase the information given to the model during the fitting process of an epoch. As training loss function, we utilized the weighted Focal loss from Lin et al. [17].

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

In the above formula, pt is the probability for the correct ground truth class t, γ a tunable focusing parameter (which we set to 2.0) and αt the associated weight for class t.

2.3.1. *Disease Risk Detector*

The disease risk detector was established as a binary classifier of the disease risk class for general categorizing between normal and abnormal retinal images. Thus, this model type was trained using only the disease risk class and ignoring all multi-label annotations. Rather than using a single model architecture, we trained multiple models based on the DenseNet201 and EfficientNetB4 architecture [9,10]. For class weight computation, we divided the number of samples by the multiplication of the number of classes (2 for a binary classification) with the number of class occurrences in the dataset.

2.3.2. *Disease Label Classifier*

In contrast, the disease label classifier was established as multi-label classifier of all 28 remaining classes (excluding disease risk) and was trained on the one hot encoded array of the disease labels. Furthermore, we utilized four different architectures for this model type: ResNet152, InceptionV3, DenseNet201 and EfficientNetB4 [9-12]. Identical to class weight computation of the disease risk detector, we computed the weights individually as binary classification for each class. Even if this classifier is provided with all classes, the binary weights balance the decision for each label individually.

2.4. *Ensemble Learning Strategy*

2.4.1. *Bagging*

Next to the utilization of multiple architecture, we also applied a 5-fold cross-validation based as a bagging approach for ensemble learning. Our aim was to create a large variety of models which were trained on different subsets of the training data. This approach not only allowed a more efficient usage of the available training data, but also increased the reliability of a prediction. This strategy resulted in an ensemble of 10 disease risk detector models (2 architectures with each 5 folds) and 20 disease label classifier models (4 architectures with each 5 folds).

2.4.2. *Stacking*

For combining the predictions of our, in total, 30 models, we integrated a stacking setup. On top of all deep convolutional neural networks, we applied a binary logistic regression algorithm for each class, individually. Thus, the predictions of all models were utilized as input for computing the classification of a single class. This approach allowed combining the information of all other class predictions to derive an inference for one single class. Overall, this strategy resulted in 29 distinct logistic regression models (1 for disease risk and 28 for each disease-label including the ‘other’ class). The individual predicted class probabilities are then concatenated to the final prediction.

The logistic regression models were also trained with the same 5-fold cross-validation sampling on a heavily augmented version of the training dataset to avoid overfitting as well as avoiding training the logistic regression models on already seen images from the neural network models. As logistic regression solver, we utilized the large-scale bound-constrained optimization (short: ‘LBFGS’) from Zhu et al. [18].

3. Results and Discussion

The sequential training took 13.5 hours with 63 epochs on average for each deep convolutional neural network model. Logistic Regression training required less than 30 minutes for all class models combined. No signs of overfitting were observed for the disease label classifiers through validation monitoring, as it can be seen in Figure 2. However, the disease risk detectors showed a strong trend to overfit. A reason for this is that the binary classification results into a too low inductive bias of the model. Due to the transfer learning, high correlation of ImageNet features with strong visual disease features are plausible resulting in neglecting minor disease features. Especially, the DenseNet architecture reveals a high risk of re-using these starting features resulting in distinct overfitting. However, through our strategy to use the earlier models based on validation loss monitoring, it was still possible to obtain powerful models for detection.

3.1. Internal Performance Evaluation

For estimating the performance of our pipeline, we utilized the validation subsets of the 5-fold cross-validation models from the heavily augmented version of our dataset. This approach allowed to obtain testing samples which were never seen in the training process for reliable performance evaluation. For the complex multi-label evaluation, we computed the popular area under the receiver operating characteristic (AUROC) curve, as well as the mean average precision (mAP). Both scores were macro-averaged over classes and cross-validation folds to reduce complexity.

Our multi-disease detection pipeline revealed a strong and robust classification performance with the capability to also detect rare conditions accurately in retinal

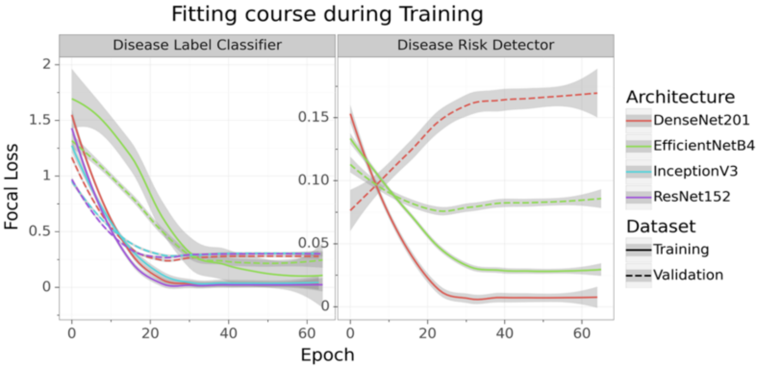

Figure 2. Loss course during the training process for training and validation data. The lines were computed via locally estimated scatterplot smoothing and represent the average loss across all folds. The gray areas around the lines represent the confidence intervals.

Table 2. Achieved results of the internal performance evaluation showing the average AUROC and mAP score for each model utilized in our pipeline. The scores were macro-averaged across all cross-validation folds and classes.

Model Type	Architecture	AUROC	mAP
Classifier	DenseNet201	0.973	0.931
Classifier	EfficientNetB4	0.969	0.929
Classifier	ResNet151	0.970	0.930
Classifier	InceptionV3	0.932	0.663
Detector	DenseNet201	0.980	0.997
Detector	EfficientNetB4	0.993	0.999
Ensembler	Logistic Regression	0.999	0.999

images. Whereas the disease label classifier models separately only achieved an AUROC of around 0.97 and a mAP of 0.93, the disease risk detectors demonstrated to have a really strong predictive power of 0.98 up to 0.99 AUROC and mAP. However, for the classifiers the InceptionV3 architecture indicated to have the worst performance compared to the other architectures with only 0.93 AUROC and 0.66 mAP. The associated receiver operating characteristics of the models are illustrated in Figure 3.

Training a strong multi-label classifier is in general a complex task, however, the extreme class imbalance between the conditions revealed a hard challenge for building a reliable model [19,20]. Our applied up-sampling and class weighting technique demonstrated to have a critical boost on the predictive capabilities of the classifier models. We base this critical boost on a synergy effect between the weighted focal loss, by handicapping samples with very high model confidence or high class frequency, and the up-sampling augmentation, by increasing the probability of a minority class to be present in a randomly drawn batch from the dataset.

Nearly all labels were able to be accurately detected, including the 'OTHER' class consisting of various extremely rare conditions. Nevertheless, the two classes 'EDN' and 'CRS' were the most challenging conditions for all classifier models. Both classes belong to very rare conditions, combined with 47 occurrences (1.2%) in the original and 209 occurrences (2.5%) in the up-sampled dataset. Still, our stacked logistic regression algorithm was able to balance this issue and infer the correct 'EDN' and 'CRS' classifications through context. Overall, our applied ensemble learning strategies resulted in a significant performance improvement compared to the individual deep

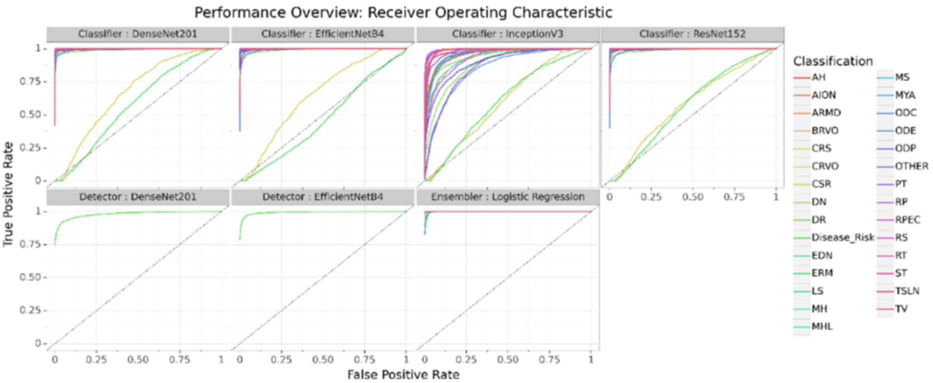

Figure 3. Receiver operating characteristic (ROC) curves for each model type applied in our pipeline. The ROC curves showing the individual model performance measured by the true positive and false positive rate. The cross-validation models were macro-averaged for each model type to reduce illustration complexity.

convolutional neural network models. More details on the internal performance evaluation are listed in Table 2.

3.2. External Evaluation through the RIADD Challenge

Furthermore, we participated at the RIADD challenge which was organized by the authors of the RFMiD dataset [7,8]. The challenge participation allowed not only an independent evaluation of the predictive power of our pipeline on an unseen and unpublished testing set, but also the comparison with the currently best retinal disease classifiers in the world.

In our participation, we were able to reach rank 19 from a total of 59 teams in the first evaluation phase and rank 8 in the final phase. In the independent evaluation from the challenge organizers, we achieved an AUROC of 0.95 for the disease risk classification. For multi-label scoring, they computed the average between the macro-averaged AUROC and the mAP, for which we reached the score 0.70. The top performing ranks shared only a marginal scoring difference which is why we had only a final score difference of 0.05 to the first ranked team.

4. Conclusions

In this study, we introduced a powerful multi-disease detection pipeline for retinal imaging which exploits ensemble learning techniques to combine the predictions of various deep convolutional neural network models. Next to state-of-the-art strategies, such as transfer learning, class weighting, extensive real-time image augmentation and Focal loss utilization, we applied 5-fold cross-validation as bagging technique and used multiple convolutional neural network architectures to create an ensemble of models. With a stacking approach of class-wise distinct logistic regression models, we combined the knowledge of all neural network models to compute highly accurate and reliable retinal condition predictions. Next to an internal performance evaluation, we also proved the precision and comparability of our pipeline through the participation at the RIADD challenge. As future work, we are interested in validating the medical gain of our pipeline for automated multi-disease detection in retinal imaging as clinical decision support through a clinical study.

Appendix

In order to ensure full reproducibility and to create a base for further research, the complete code of this study, including extensive documentation, is available in the following public Git repository: https://github.com/frankkramer-lab/riadd.aucmedi
Furthermore, the trained models, evaluation results and metadata are available in the following public Zenodo repository: https://doi.org/10.5281/zenodo.4573990

Acronym list of class names in the RFMiD dataset (all classes which are not represented in table 1 had less than 10 samples in the public as well as hidden/hold-out dataset and were merged as “OTHER”):
Disease Risk (D. Risk), diabetic retinopathy (DR), age-related macular degeneration (ARMD), media haze (MZ), drusen (DN), myopia (MYA), branch retinal vein occlusion (BRVO), tessellation (TSLN), epiretinal membrane (ERM), laser scar (LS), macular scar (MS), central serous retinopathy (CSR), optic disc cupping (ODC), central retinal vein

occlusion (CRVO), tortuous vessels (TV), asteroid hyalosis (AH), optic disc pallor (ODP), optic disc edema (ODE), shunt (ST), anterior ischemic optic neuropathy (AION), parafoveal telangiectasia (PT), retinal traction (RT), retinitis (RS), chorioretinitis (CRS), exudation (EDN), retinal pigment epithelium changes (RPEC), macular hole (MHL), retinitis pigmentosa (RP), cotton wool spots (CWS), coloboma (CB), optic disc pit maculopathy (ODPM), preretinal hemorrhage (PRH), myelinated nerve fibers (MNF), hemorrhagic retinopathy (HR), central retinal artery occlusion (CRAO), tilted disc (TD), cystoid macular edema (CME), post traumatic choroidal rupture (PTCR), choroidal folds (CF), vitreous hemorrhage (VH), macroaneurysm (MCA), vasculitis (VS), branch retinal artery occlusion (BRAO), plaque (PLQ), hemorrhagic pigment epithelial detachment (HPED) and collateral (CL)
For more information and details on the dataset, we refer to Pachade et al. [7,8].

Acknowledgements
We want to thank Dennis Klonnek, Edmund Müller and Johann Frei for their useful comments and support.

Compliance with Ethical Standards
This research study was conducted retrospectively using human subject data made available in open access by Pachade et al. [7,8]. Ethical approval was not required as confirmed by the license attached with the open access data.

Contributions of the Authors
Dr. Frank Kramer and Dr. Iñaki Soto-Rey were in charge of coordination, review, and correction of the manuscript. Dominik Müller contributed to the conception and design of this work, its data analysis and interpretation, and was in charge for draft and revision of the manuscript.

Conflict of Interest
None declared.

Funding
This work is a part of the DIFUTURE project funded by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) grant FKZ01ZZ1804E.

References

[1] J. D. Adelson *et al.*, “Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study,” *Lancet Glob. Heal.*, vol. 9, no. 2, pp. e144–e160, Feb. 2021, doi: 10.1016/S2214-109X(20)30489-7.

[2] World Health Organization, “Blindness and vision impairment.” https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed Feb. 27, 2021).

[3] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: benefits, risks, and strategies for success,” *npj Digital Medicine*, vol. 3, no. 1. Nature Research, pp. 1–10, Dec. 01, 2020, doi: 10.1038/s41746-020-0221-y.

[4] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, no. December 2017, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[5] J. Y. Choi, T. K. Yoo, J. G. Seo, J. Kwak, T. T. Um, and T. H. Rim, “Multi-categorical deep learning

neural network to classify retinal images: A pilot study employing small database," *PLoS One*, vol. 12, no. 11, p. e0187336, Nov. 2017, doi: 10.1371/journal.pone.0187336.

[6] G. Quellec, M. Lamard, P. H. Conze, P. Massin, and B. Cochener, "Automatic detection of rare pathologies in fundus photographs using few-shot learning," *Med. Image Anal.*, vol. 61, p. 101660, Apr. 2020, doi: 10.1016/j.media.2020.101660.

[7] S. Pachade *et al.*, "Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research," *Data*, vol. 6, no. 2, p. 14, Feb. 2021, doi: 10.3390/data6020014.

[8] "Home - RIADD (ISBI-2021) - Grand Challenge." https://riadd.grand-challenge.org/Home/ (accessed Feb. 27, 2021).

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, Accessed: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1608.06993.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1905.11946.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, vol. 2016-December, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, vol. 2016-December, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[13] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[14] S. Muhammad *et al.*, "Medical Image Analysis using Convolutional Neural Networks A Review," *J. Med. Syst.*, vol. 42, no. 11, pp. 1–13, Nov. 2018, doi: 10.1007/s10916-018-1088-1.

[15] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9379, 2017, doi: 10.1109/ACCESS.2017.2788044.

[16] D. P. Kingma and J. Lei Ba, "Adam: A Method for Stochastic Optimization," 2014. https://arxiv.org/abs/1412.6980.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Aug. 2017, Accessed: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1708.02002.

[18] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, Dec. 1997, doi: 10.1145/279232.279236.

[19] P. Kaur and A. Gosain, "Issues and challenges of class imbalance problem in classification," *Int. J. Inf. Technol.*, pp. 1–7, Oct. 2020, doi: 10.1007/s41870-018-0251-8.

[20] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling imbalanced medical image data: A deep-learning-based one-class classification approach," *Artif. Intell. Med.*, vol. 108, p. 101935, Aug. 2020, doi: 10.1016/j.artmed.2020.101935.