Improving Disease Classification on Rare Class Distribution X-ray Images Using Supervised and Few-Shot Hybrid Learning

Jannatul Nayeem[1], Soumodeep Biswas[2], Md Ifty Rahman[3],
Sabiha Salam Ayesha[4], Tanim Ahmed Saad[5] and Md Tanzim Reza[6]
[1,3,4,5,6]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[1]jannatul.nayeem@bracu.ac.bd, [2]soumodeep.biswas@g.bracu.ac.bd, [3]md.ifty.rahman@g.bracu.ac.bd,
[4]sabiha.salam.ayesha@g.bracu.ac.bd, [5]tanim.ahmed.saad@g.bracu.ac.bd, [6]tanzim.reza@bracu.ac.bd

***Abstract*—Disease diagnosis through medical image analysis using various transfer learning models and neural networks have made significant progress in recent years. However, Medical Image datasets are highly imbalanced due to the minimal number of cases of rare diseases. As a result of this imbalance, pre-trained CNN models perform poorly in detecting rare diseases in supervised classification tasks. Classes with a high number of data samples dominate in conventional supervised learning setup. Therefore, our research focused on trying to minimize the effect of this class imbalance. We proposed a hybrid architecture which put together supervised learning and few-shot learning. For common class detection, we used a pre-trained MobileNet-V2 as the base model of the classical supervised learning. For Rare classes, a few shot learning model, Relation Network was responsible for detecting rare disease classes. Our proposed hybrid architecture achieved an average of 90% F1 score on the rare classes. In contrast, we experimented with 3 pre-trained CNN models for traditional supervised learning and observed that all of them had scored poor recall or precision value with an average of 45% F1 score on the rare classes. Therefore, our findings highlighted that our proposed hybrid architectural approach is more impactful and can achieve good results. For that reason, we believe our work opens the door for researchers for future study that a hybrid approach of combining supervised and few-shot learning can be effective instead of relying on only one.**

***Index Terms*—Supervised Learning, Few-shot Learning, Relation Network, Feature Embeddings, Support set, Query set**

I. Introduction

Deep neural networks (DNNs) have significantly advanced in the medical field. Deep learning techniques have enabled rapid disease detection, assisting healthcare providers in diagnosing conditions efficiently from medical images, including X-rays LeCun, Bengio, and Hinton, 2015 [1]. Medical image classification using deep learning has been pivotal in diagnosing diseases such as pneumonia, tuberculosis, and lung cancer from chest X-ray CXR images Rajpurkar et al. 2017 [2]. However, despite the powerful capabilities of DNNs in medical image analysis, several limitations arise when it comes to rare diseases. A critical challenge in these applications is the scarcity of labeled data for rare conditions, which limits the generalizability of supervised learning models Chen et al. 2019 [3]. Factors such as patient privacy concerns, the rarity of specific diseases, and high costs of collecting medical images contribute to the data limitation in medical imaging Kermany et al. 2018 [4]. These challenges lead to models overfitting to the few available training samples, thus performing poorly when exposed to unseen data Kang et al. 2020 [3]. This issue is particularly concerning in the diagnosis of rare diseases, where the availability of images is often insufficient to build effective supervised models. For these reasons, high class imbalance issues arise in medical CXR datasets and Deep neural networks in supervised learning setup performs poorly on rare diseases and common class disease detection dominates in performance. But detection of rare diseases in an efficient manner similar to common diseases is also important.

Hence, we propose a hybrid solution that combines the two popular learning methods: supervised learning and few-shot learning. In our proposed method, For common class detection we have used a supervised learning model and to detect rare diseases with high performance a few shot learning model was responsible.The main objective of our research work is to reduce the bias toward common classes in DNN's by developing a hybrid architecture that uses supervised and few-shot learning so that the performance of rare disease detection does not suffer from class imbalance issues.

II. Background Study

A. Previous Works

The paper [5] addresses one of the most common and challenging problems in machine learning: class imbalance, which occurs when some classes in a dataset have significantly more samples than others. Traditional deep learning models show bias towards the majority classes and fail to perform good results on minority classes. The current methods can be grouped into two popular categories according to the survey:- data-level techniques, which involve increasing or decreasing the training data, and algorithm-level strategies, which focus on modifying the model's learning process from the data. The

widely used techniques of data level approaches are oversampling and under-sampling. Over sampling involves creating synthetic data instances of minority classes with SMOTE or GANs, whereas under sampling deletes some data samples from majority classes to form a uniform class distribution. But oversampling can cause over-fitting and under-sampling can be a reason for information loss. With a focus on the CXR14 dataset The study [6] analyzes the class imbalance in medical image classification using chest X-rays. The authors studied how deep learning models like ResNet and DenseNet, can be effective in diagnosis but struggle with imbalanced datasets. In the paper [7] which was published in Medical Image Analysis. The authors mention the limitation of conventional deep learning techniques, which depend on large annotated datasets which are in some of the situations unavailable for rare diseases because of their low ubiquity. To address this problem, they suggest a brand-new discriminative ensemble learning strategy that uses Few Shot Learning to categorize thoracic diseases using a small number of training examples. The paper [8] reviews few-shot learning (FSL) for COVID-19 classification in chest X-ray images under severe data imbalance, focusing on intra- and inter-domain scenarios. It proposes a Siamese neural network-based methodology that combines techniques like data augmentation, transfer learning, weighted loss, and balanced sampling to address data imbalance and scarcity. The study evaluates different approaches and compares their outcomes to a common CNN baseline using four publicly available chest X-ray datasets.

B. Algorithms

Relation Network: Relation Network is designed to build a classifier that will be able to classify only seeing a few images of each class. The Relation network was introduced by F Sung et al. in 2017. According to the author of the original paper the purpose of embedding module is to generate feature representation of the query images and the labeled support set images. The generated feature embeddings will then be concatenated and will be used as input of the relation module. In our study we have used MobileNet-V2 as the architecture of the embedding module. To reduce the output feature map of the MobileNet architecture we used a global average pooling. After that we flattened the pooled features and projected into a lower dimensional embedding of size 128. The author of the paper stated that relation module should be an artificial neural network hence we used a network of two fully connected layers of size 128 , to introduce non-linearity used ReLU activation and sigmoid function for classification.The purpose of this module is to learn the distance between support and query set images so that it can classify if the images belong to matching categories or not.

MobileNet-V2: It is an upgraded version of MobileNet-V1 architecture which was published by A. Howard et al. (2017). The architecture consist of 7 Bottleneck layer. Some of the important components of each Bottleneck layer are, it has one 1x1 Expansion layer, 3x3 depthwise convolution layer, two RELU6 activation function was used and a 1x1 projection layer. MobileNet-V2 was introduced by M. Sandler et al. (2018) [9] based on the principle of MobileNet-V1. Similar to MobileNet-V1 it also has depthwise separable convolution operation which is a key operation that reduces the computation complexity. Novel ideas that are introduced in MobileNet-V2 are Inverted residuals, Linear Bottlenecks and depthwise separable convolutions.Depthwise separable convolutional operation is divided into two parts:- depthwise and pointwise convolution. Standard convolution operations are done on the entire input channel which means if the input channel is 3 the convolution operation is performed into these 3 channels at a time. On the other hand, depthwise convolution operation is performed onto one single channel at a time which reduces the number of multiplications compared to a standard convolutional operation. By performing linear combination pointwise convolution binds all the outputs of depthwise convolution per channel and generates the final feature map. Pointwise convolution is a 1x1 kernel. M. Sandler et al. (2018) [9] claimed that expanding and compressing the feature maps using non-linear activation functions has its downsides. Exposure to non-linearity in low-dimensional spaces can lead to a good amount of information loss. Therefore, to minimize the information loss they introduced linear bottlenecks. Linear bottleneck layers reduces the feature map channels without using the non-linear activation functions like ReLU.Inverted residuals are different from Traditional residual blocks hence the name inverted. In traditional residual block expansion is done first and then compression is performed. On the other hand, inverted residuals do the opposite, it is first compressed and following the compressed operation, expansion is done.

III. Dataset Details and Processing

A. Dataset Description

We have sourced our dataset from multiple open source dataset. Since we are working with a problem that requires an imbalanced dataset that is why we were looking for rare diseases among these open sourced available CXR image datasets and we have found that Edema Fibrosis and Hernia has very few data samples available compared to other CXR images. In our dataset we have a total of 9 different chest X-ray diseases and a total of 17,615 frontal Chest X-ray radiology images.

B. Data Analysis and Pre-processing

The dataset we created by sourcing images from multiple datasets is highly imbalanced. Edema, Fibrosis, Hernia these three classes have very few data samples compared to the other 6 classes. Therefore, the dataset is perfectly suitable for our research problem because we needed a dataset where few classes will suffer from data scarcity and will represent as rare classes. From the frequency bar chart given below it is clearly visible that the class distribution is not uniform and the height of the aforementioned three classes Edema, Fibrosis, Hernia is much smaller than the other classes.

Also it is worth mentioning that since CXR image datasets are mostly multi-labeled for that reason we first filtered out

TABLE I
DISEASE CLASSES WITH SAMPLE COUNTS AND DATASET SOURCES

Dataset Summary	
Disease	**Count & Source**
COPD	3510 (PadChest)
Covid-19	3017 (Kaggle)
Edema	100 (NIH CXR-14)
Emphysema	2550 (NIH CXR-14)
Fibrosis	105 (NIH CXR-14)
Hernia	106 (NIH CXR-14)
Normal	3271 (Guangzhou Women and Children's Medical Center)
Pneumonia	4273 (Guangzhou Women and Children's Medical Center)
Tuberculosis	1683 (Kaggle)

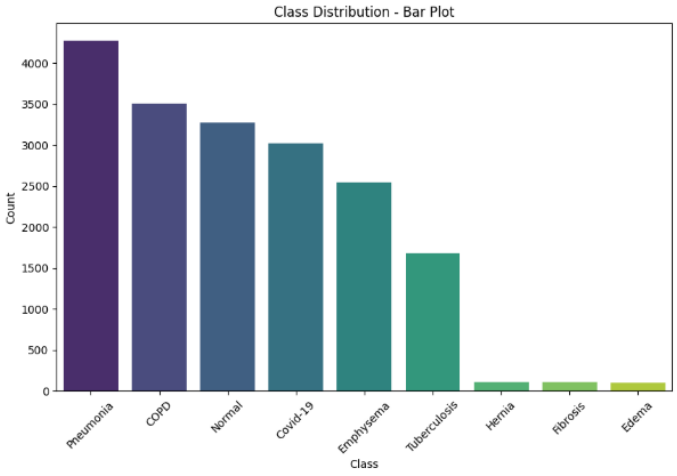

Fig. 1. Class Distribution

the single labeled images first and selected the desired images from the chosen class.Some of the images were blurry or noisy in our dataset that we collected. Hence we applied a standard procedure to enhance the quality of those images using a widely used technique called CLAHE. The following figure depicts the comparison of before and after applying CLAHE. Before applying CLAHE the original image looks blurry specially in the chest region and the edges inside the rib cage do not look much visible. After applying CLAHE the contrast of the image improved and the edges inside the rib cages looked visible better than the original image

C. Image Transformations

To introduce diversity and reduce any possible overfitting issue and models to generalize better we have used several image transformations techniques in our training data pipeline. The followings are the techniques we used in our experiments-

• Image Resize and Center Crop: We resized the CXR images by 256*256 pixel values from original image size then

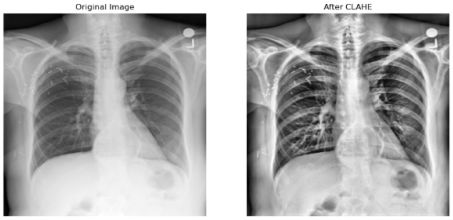

Fig. 2. Before and After CLAHE

did a center cropping of size 224*224.

• Random Horizontal Flip: This transformation technique does a random horizontal flip of the input image.

• Random Rotation: This technique rotates any given image to a certain degree. We provided the degree to be 10.

• Random Affine: This transformation does multiple random operations on an image object including translation, rotation, shearing and scaling. It is a kind of geometric transformation that maintains the lines and parallelism

• Color Jitter: With the help of pytorch's color jitter transformation we can tweak an image's color properties such as brightness,contrast,hue and saturation.

IV. PROPOSED METHOD

Because of the nature of our dataset which is imbalanced, a supervised model suffers to identify every class with the same good performance.Therefore, in order to build a reliable architecture that will have good efficient performance across all the classes we propose a hybrid architecture by combining two different learning methods, Supervised and Few-shot learning. Since we can divide our dataset into two subsets. The elements of subset one can be the classes that have vast amounts of data samples. We call this subset "Common class" and we named the other subset as a "Rare class" because the number data samples of these classes are inadequate. The classes in the Common diseases are as follows: COPD,Covid-19,Emphysema,Normal,Pneumonia,Tuberculosis. Rare class diseases are Edema, Fibrosis, Hernia.

We divided our entire problem into two sections. The first task is for identifying the Common class elements and we chose to use a supervised learning model for this task because this method is known for performing at a high level when there is enough data sample. On the other hand, since the other subset Rare class elements have less amount of data the second task was to find a learning method that is not dependent on the number of data samples. For that reason, we selected Few-shot learning since this learning method is designed with the purpose of performing at a high level for scenarios when there is data scarcity. To achieve this hybrid architecture we divided the entire architecture into phase 1 and phase 2.

Phase 1 has two goals-

- Classifying the elements of the Common class subset with high precision and recall value.
- Routing any data sample belonging to the Rare Class subset to the Phase 2 Few-shot learning method.

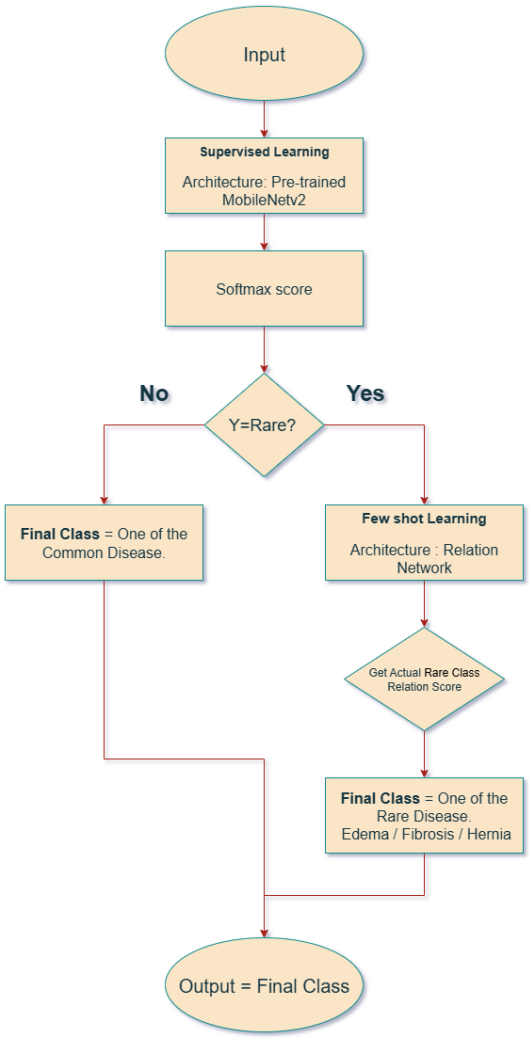

Fig. 3. Proposed Hybrid architecture

In Phase 2 we used a Few shot learning method so that the data sample rerouted from Phase 1 can be classified into the specific element of the Rare class subset.

To achieve the goal of phase 1, first we introduced a new class called "Rare" into our dataset and this class has all the images belonging to any of the following elements of the Rare class subset which are Edema, Fibrosis and Hernia. All these three classes data samples are merged into one single class called "Rare". Therefore, with this modified dataset we trained several pre-trained CNN models to find the best performing model. After conducting several experiments with different pre-trained models we observed that MobileNet-V2 architecture showed the best results compared to the other models by achieving an average value of 95.29% F1 score which indicates the high precision and recall value across all the classes. Hence, we selected the MobileNet-V2 architecture as the base model for classifying the common classes and rerouting the newly merged "Rare" class to the Few-shot model.

To further classify the re-routed images from Phase 1 we have used a Few-shot learning model called "Relation Network". The sole purpose of Phase 2 is finding the best FSL model that performs much better than plain supervised models with limited data. We experimented with several FSL models and found that Relation Network achieved higher performance in our three Rare classes. We have used 3 way 5 shot training setup with 100 episodes and the size of input images were 224*224 since we have selected MobileNet V2 for extracting features. The Relation Network is divided into two parts. The first part is responsible for creating feature embeddings of support and query set images known as embedding module. The second part is known as the relation module which takes the support and query set images feature embeddings as inputs, analyzes them and generates the relation score between the support and query set images.

V. Result and Analysis

Phase 1 Supervised Model: To find the best model suited for Phase 1 supervised learning we have experimented with three State of the Art pretrained CNN architecture- MobileNet-V2, EfficientNet-B1 and DenseNet121. All the three models achieved more or less 91% overall accuracy. Along with other class performance we mostly prioritized the F1-score of the class that was formulated combining the three rare disease classes from the actual dataset, named "RARE". In order to get the most benefit from the few-shot model in Phase 2, it is significantly important that Phase 1 model successfully captures almost all of the images belonging to the Rare class and hardly gets confused with other common classes.

TABLE II
F1 Score Comparison Across Different Models

F1 Score (%)			
Disease Name	**MobileNet-V2**	**EfficientNet-B1**	**DenseNet-121**
COPD	100%	100%	100%
Covid-19	89%	91%	85%
Emphysema	89%	90%	83%
Normal	99%	97%	92%
Pneumonia	100%	98%	96%
Tuberculosis	96%	97%	88%
RARE	94%	85%	63%

For all of the three models batch size was 32, total epoch was 25 and 0.001 learning rate. The best model was saved when the model scored lowest validation loss on a held out validation set. The selected pretrained MobileNet-V2 model was trained for 25 epochs. The best model was saved at epoch 19 with 8% validation loss. Training parameters like, batch size was 32 , learning rate was 0.001. For regularization purposes the dropout rate in the fully connected layer was 0.5

in order to prevent the overfitting. For all of the models, we used Adam optimizer and the loss function was cross entropy loss.

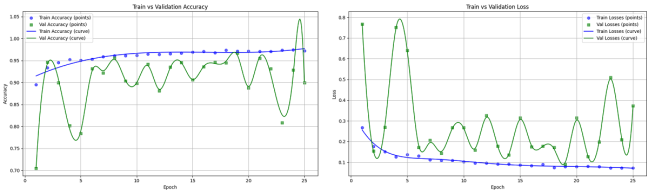

Fig. 4. Training History of Supervised Model

Phase 2 Few Shot Model: In order to select the best Few-shot learning algorithms we implemented two of the famous Few-shot learning methods Siamese Network and Relation Network. Between these two methods **Relation Network outperformed Siamese Network [10] in terms of overall performance**. It has achieved overall 96% and 95% precision, recall value on Test dataset. For that reason, we have selected Relation Network as the best candidate model for the Phase 2 of our proposed hybrid architecture.The Relation Network model was trained for a total 100 epochs and each epoch there were 50 episodes. Learning rate was 0.001 and Image size was 224*224. The training and testing of the model was done in a 3 way 5 shot manner. Which means, in each episode 3 classes were chosen with 5 randomly supported images. At the same time, from each 3 class 5 query images were randomly chosen as well. Hence, the length of the query set images was 15.

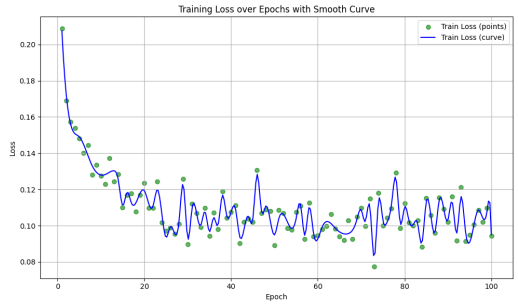

Fig. 5. Training History of Few-Shot model

Proposed Hybrid Pipeline Result: The performance of our proposed methodology is shown in the table below:

From the table above, it is clear that our hybrid architecture has significantly improved the performance of the three rare classes without degrading the performance of other common classes. If we observe the recall value of Edema we can see that it achieved 100% score with 90% precision which means the proposed pipeline successfully captured all the positive cases of Edema with high precision. The same can be said with other two rare classes that achieved close to 90% recall and precision value.

Comparative Study: In order to understand the effectiveness of our proposed methodology we have compared the rare class result of our proposed solution with three SOTA pre-trained CNN model's, MobileNet-V2, Efficient-B1 and DenseNet-121.

TABLE III
CLASSIFICATION PERFORMANCE OF THE PROPOSED HYBRID ARCHITECTURE

Performance Metrics (%)			
Disease Name	**Precision**	**Recall**	**F1-Score**
COPD	100%	100%	100%
Covid-19	96%	87%	91%
Edema	90%	100%	94.7%
Emphysema	88%	96%	92%
Fibrosis	88.9%	88.9%	88.9%
Hernia	88.9%	88.9%	88.9%
Normal	99%	97%	98%
Pneumonia	98%	100%	99%
Tuberculosis	82%	100%	90%

TABLE IV
F1 SCORE ACROSS DIFFERENT PRE-TRAINED CNN MODELS VS PROPOSED HYBRID ARCHITECTURE

F1 Score (%)				
Rare Class	**MobileNet-V2**	**EfficientNet-B1**	**DenseNet-121**	**Our Model**
Edema	64%	57%	75%	94.7%
Fibrosis	31%	79%	0%	88.9%
Hernia	14%	67%	25%	88.9%
Average Score	**36.33%**	**68.33%**	**33.33%**	**90.83%**

From the table it can be inferred that our proposed hybrid architecture achieved an average of 90% F1 score on the rare classes. In contrast, all 3 SOTA pre-trained CNN models performed poorly with an average of 45.98% F1 score on the rare classes. From the below two bar charts which represents the comparison of Recall and F1 Score of the Hybrid pipeline vs a plain supervised model across the all classes we can observe that significant improvement in both recall and F1 score value in the proposed hybrid pipeline.

We can see that among the three rare classes Fibrosis and Hernia performed awfully in terms of recall value whereas in the proposed hybrid architecture the recall value of Fibrosis and Hernia rose exceptionally high which indicates that compared to the plain supervised pretrained models proposed hybrid architecture successfully captured all the positive cases of Fibrosis and Hernia.On the other hand, even though in terms of supervised models, Edema has a good height in the recall bar chart but its height is significantly low in the F1-score bar chart slightly above 50% which indicates that

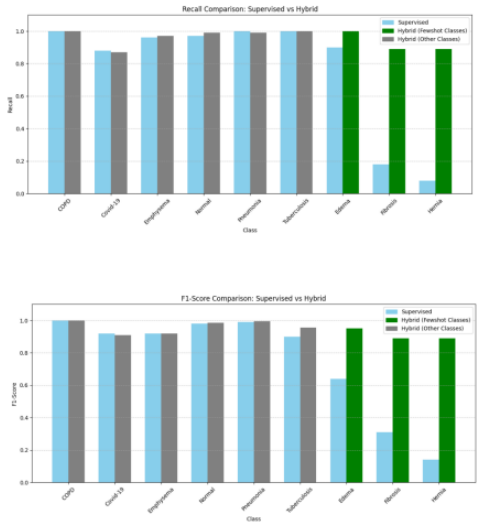

Fig. 6. Recall and F1 score comparison with Pre-trained Model

the precision of Edema was low. Lastly, if we observe the bar chart height of other two rare class Fibrosis and Hernia in F1-score bar chart it is clearly understandable that not only the recall but also the precision value of the proposed hybrid model significantly improved compared to the pre-trained supervised model because the F1-score of these two classes in the supervised model was extremely low which clearly reflects in the bar chart as well.

Therefore, we can conclude that a supervised model tends to suffer from identifying the classes with less amount of data and will always excel at classifying categories with vast amounts of data. Hence, it is feasible to use a specialized architecture to categorize the classes with scarce data. And our experiment and results of the proposed hybrid pipeline supports the above mentioned approach.

VI. Conclusion and Future Works

In the realm of medical diagnosis leveraging the most of Artificial Intelligence will be a game changing factor. But most of the CXR image dataset are highly imbalanced and in image classification tasks AI becomes ineffective to determine class instances with less amount of data. For that reason, we were motivated to study a solution that is both effective and robust with the data already available. Since most of the CXR image dataset has both common and rare diseases, only using a pre-trained supervised model will not give expected results due to the class imbalance problem. For that reason, we explored the idea of combining the two learning methods, supervised learning in order to classify common diseases and few-shot learning for rare diseases. Our experiment results have shown that our proposed method achieved almost double the F1 score on an average compared to the supervised models. We believe our findings open the door for researchers for future work using a hybrid architectural approach of combining supervised and few shot learning instead of relying on only one. Even though we achieved a good result there are certain limitations in our proposed methodology, for instance since we have merged the three rare diseases into a single unified class called ”RARE” in Phase-1 hence, if these rare disease images are not highly correlated there is a chance that the Phase 1 Supervised Model may fail to reroute the rare disease class images effectively to the Phase 2 Few-Shot model.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[2] P. Rajpurkar, J. Irvin, and R. Ball. Deep learning for chest x-ray analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 64(12):2862–2873, 2017.

[3] J. Chen, S. Zhang, and Z. Liu. Exploring the impact of data scarcity on deep learning models for rare disease classification. *Journal of Medical Imaging*, 6(2):010501, 2019.

[4] D. S. Kermany, K. Zhang, and M. Goldbaum. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[5] J.M. Johnson and T.M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 2019.

[6] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, and A. Tayebi. Unraveling the impact of class imbalance on deep-learning models for medical image classification. *Applied Sciences*, 14(8), 2024.

[7] Angshuman Paul, Yu-Xing Tang, Tony C. Shen, and Ronald M. Summers. Discriminative ensemble learning for few-shot chest x-ray diagnosis. *Medical Image Analysis*, 68, 2021.

[8] A. Galán-Cuenca, A. J. Gallego, M. Saval-Calvo, and A. Pertusa. Few-shot learning for covid-19 chest x-ray classification with imbalanced data: An inter vs. intra domain study, 2024.

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *arXiv preprint arXiv:1801.04381*, 2018.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[12] J. D. Kang, S. Lee, D. Kim, and S. Kim. Overfitting in medical image classification: How deep learning struggles with small datasets. *Journal of Computational Medicine*, 12(4):287–296, 2020.

[13] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[14] Antonio Bustos, Antonio Pertusa, José M. Salinas, and María de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66, 2020.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *CoRR*, 2017.