Exploring Advances in Multi-Label Chest X-Ray Classification: A Comprehensive Literature Review

Subhash Vemula, Ram Pavan, Harshavardhan, Patanjali, Kesava Karthik
Department of Electronics and Electrical Engineering
Indian Institute of Technology
Guwahati, India

***Abstract*— This paper delves into the complex realm of classifying chest X-ray findings, where common cases coexist with rare conditions, leading to bias in traditional methods. It addresses the issue of handling long-tailed distributions in multi-label chest X-ray classification, aiming to improve accuracy across both prevalent and seldom-seen conditions, providing insights for future algorithm development in medical image classification.**

Keywords— Data Science, Deep Learning, X-Ray, Machine Learning, Artificial Intelligence, Health Care, CNN, Neural Network, Class imbalance

I. INTRODUCTION

A. Background

In the field of healthcare, the integration of deep learning has ushered in a new era of potential and innovation. This technology offers the capability to rapidly and accurately analyze extensive medical data, surpassing the limitations of rule-based algorithms by continuously learning from large datasets. One area of significant promise is the application of deep learning to chest X-ray analysis, where it has demonstrated its prowess in identifying a range of medical conditions. However, as with any technology, challenges persist, particularly in the realm of classifying long-tailed multilabel diseases. This challenge is particularly pronounced in clinical settings where the identification of rare conditions poses significant obstacles.

B. Challenges and Motivation

The focal point of our study addresses the inherent difficulty of classifying long-tailed multilabel diseases in chest X-rays. This issue is a common and critical challenge in clinical settings, where the identification of rare conditions is of paramount importance. The term "long-tailed" denotes an imbalanced distribution of diseases, where certain conditions are infrequent compared to others. Existing literature reviews conducted as part of our research have revealed that many image classification benchmarks primarily concentrate on datasets characterized by imbalanced label distributions. The papers we've reviewed consistently highlight the unique challenges associated with this imbalance, emphasizing the need for specialized methodologies to tackle the complexities of multilabel classification in the context of chest X-ray images. In the subsequent sections, we will delve into a comprehensive analysis of the methodologies employed, experimental outcomes, and the limitations of our proposed model in addressing these challenges.

II. RELATED WORKS

A. Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study

The paper addresses the challenge of long-tailed distribution in thorax disease classification using chest X-rays. It formalizes the long-tailed classification task, categorizing diseases into "head," "medium," and "tail" classes based on their frequency. The paper describes the dataset creation process for NIH-CXR-LT and MIMIC-CXR-LT, introducing rare diseases through text mining. It's important to note that these datasets primarily consist of single-labeled images rather than multi-disease labels. The datasets exhibit extreme class imbalance, mirroring clinical reality. The paper discusses methods for benchmarking long-tailed classification, including class re-balancing, information augmentation, and other techniques. Notably, standard approaches struggle with medium and tail classes, while re-weighting and Deferred Re-Weighting (DRW) show promise. Classifier Re-Training (cRT) decoupling emerges as a powerful technique for this task.

For a fair comparison, training utilized a ResNet50 model pre-trained on ImageNet, employing the Adam optimizer with a learning rate of 1×10^{-4}. Notably, the conventional approach of optimizing softmax cross-entropy with instance-balanced weights faced challenges in effectively classifying medium and tail classes across both the NIH-CXR-LT and MIMIC-CXR-LT datasets. Interestingly, despite the documented success of MixUp and Balanced-MixUp in various image-based tasks, our review of existing research revealed that these methods yielded performances akin to the baseline, prompting questions about their applicability in medical imaging contexts. Throughout our review, we consistently observed performance improvements with re-weighting strategies, such as Focal Loss [16], Label-Distribution-Aware Margin (LDAM) Loss [2], and Influence-Balanced Loss, though

the selection of a specific re-weighting method appeared to interact differently with the underlying loss function. Furthermore, the literature indicated that the incorporation of Deferred Re-Weighting (DRW), particularly in tandem with the LDAM loss, demonstrated promise in enhancing classification outcomes. Remarkably, the reviewed literature highlighted Classifier Re-Training (cRT) decoupling as the most effective approach on both datasets, emphasizing its simplicity and effectiveness for the challenging task of long-tailed disease classification in chest X-rays. Classifier re-training (cRT) achieving the highest group-wise average accuracy of 0.369 on the balanced test set. RW LDAM-DRW closely follows with 0.362 group-wise average accuracy.

While the paper followed standard practice of utilizing ImageNet pre-trained weights, this choice constrained our selection of potential long-tailed learning methods. For instance, certain long-tailed methods employing specialized architectures or investigating self-supervised learning on other datasets may not align with ImageNet pretraining. There is room for improvement in addressing long-tailed data through the exploration of alternative weight initialization techniques. Furthermore, we can consider adapting multi-label long-tailed learning methods to these datasets, recognizing the clinical reality that patients often present with multiple pathologies simultaneously.

B. SwinCheX: Multi-label classification on chest X-ray images with transformers.

The paper addresses multi-label chest X-ray image classification, proposing a model based on the Swin Transformer as the core architecture with MLP for the head architecture. Computer vision has traditionally relied on Convolutional Neural Networks (CNNs) for high-accuracy classification tasks. However, emerging research demonstrates that Transformers, initially prominent in Natural Language Processing (NLP), can surpass many CNN-based models in vision tasks. This paper evaluates a model on the "Chest X-ray14" dataset, employing Vision Transformers (ViT) for multi-label classification. The model adapts a pre-trained Swin Transformer and integrates multi-head feedforward neural network layers for each class, considering different configurations. Additionally, Grad-CAM saliency heatmaps are used to verify that the model attends to clinically relevant chest areas. Attention mechanisms, a pivotal development, enable models to selectively focus on specific input aspects. Transformers, which are self-attention-based architectures, expedite sequence translation, although challenges persist in transferring their language domain success to computer vision. Swin Transformers address these challenges, mitigating computational complexity and enabling their practical use in vision tasks.

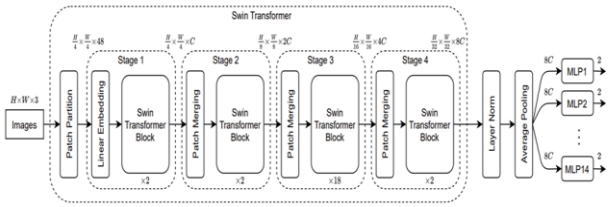

Figure 1. The overall architecture of our method. Chest X-ray images after being converted to RGB and resized to 224 are passed through a Swin-L17 transformer. This follows a Layer-Norm, and a 7×7 average pooling, before the shared section is finished. 14 MLP heads are branched from the shared section.

The model employs the Swin Transformer as the shared component for predicting each label. The Swin Transformer initially divides the input RGB image into non-overlapping patches, treating each patch as a token. A linear embedding layer is then applied to these patches to project them into a dimension denoted as C. Several Swin Transformer blocks are subsequently applied, and as the network deepens, the number of tokens decreases through patch merging layers. This process is repeated, featuring a shared architecture at the beginning and one head for each pathology after the initial section. The shared section utilizes the Swin-L17 transformer with weights initialized from ImageNet22k pre-trained weights. To reduce the output dimension, Layer-Norm average pooling is applied. The head sections employ a 3-Layer MLP with 384, 48, and 48 neurons in each layer, respectively. The last layer is connected to a single output, normalized by a sigmoid function. Each head provides a probability indicating the presence of the corresponding disease. During training, binary cross-entropy loss is used as the loss function, with a batch size of 32 and a learning rate of $3\times10-5$.

The evaluation protocol involves splitting the training data into training and validation sets, with 80% for training. Patients don't appear in both splits to prevent bias. We use AUROC as the primary evaluation metric. During training, we monitor AUROC on the validation set after each epoch and select the best-performing model. We then report the selected model's AUC on the official test set. This approach simplifies hyperparameter use and is easily applicable to other datasets.

Table 1. Comparison of the DNet and proposed method based on the AUCs

pathology	DNet[14]	DNet with the proposed evaluation method	3-layer head SwinCheX
Cardiomegaly	**0.883**	0.878	0.875
Emphysema	0.895	0.886	**0.914**
Edema	0.835	0.828	**0.848**
Hernia	0.896	**0.9**	0.855
Pneumothorax	0.846	0.842	**0.871**
Effusion	**0.828**	0.826	0.824
Mass	0.821	0.803	**0.822**
Fibrosis	0.818	0.823	**0.826**
Atelectasis	0.767	0.763	**0.781**
Consolidation	0.745	0.744	**0.748**
Pleural Thicken	0.761	0.768	**0.778**
Nodule	0.758	0.742	**0.78**
Pneumonia	**0.731**	0.7	0.713
Infiltration	**0.709**	0.686	0.701
Mean	0.807	0.799	**0.81**

3-layer headed SwinCheX model achieves an average AUC value of 0.81 for 14 pathologies, surpassing the previous state-of-the-art DNet model's AUC of 0.807. Notably, the model's advantage lies in its simplicity, as it does not employ adaptive learning rates, which introduce additional hyper-parameters and potential overfitting challenges.

Although our proposed model demonstrated certain improvements, a broader examination of various vision transformers extending beyond ViT16 and Swin17 can give deeper insights into their suitability for multi-label chest X-ray classification. Additionally, development of proficient preprocessing methods tailored to address issues like scattered lung positions, watermarks, and contrast variations within the dataset can enhance the model's robustness

C. Multi-Label Chest X-Ray Classification via Deep Learning.

This paper aims to create a lightweight solution for detecting 14 chest conditions from X-ray images, expanding upon the original study by the Stanford ML Group, which focused on predicting 5 diseases. It enhancd previous work and offered insights for future chest radiography research.

This paper proposes a model leveraging CNN architecture, known for its parameter efficiency while retaining critical features, for multi-class, multi-label image classification, to detect 14 chest conditions form chest X-ray image. It experiments with various models, including custom CNNs, DenseNet121, ResNet-50, Inception_V3, and VGG. These models are trained using a batch size of 96, binary cross-entropy loss, and the Adam optimizer with an initial learning rate of 0.001. It employs weight fine-tuning after validation loss plateaus and utilizes sigmoid functions to convert outputs into probabilities, considering values above 0.50 as positive detections. DenseNet121 and ResNet-50 utilized pre-trained weights from ImageNet data, with selective adaptation of last layer weights. For Inception_V3, it initializes with pre-trained ImageNet weights and freezes the first 8 layers. Meanwhile, VGG16, with its 134 million trainable parameters, has its first 6 layers frozen to limit parameters to 57k.

The Chexpert dataset is split into train and validation sets with an 80:20 proportion. An additional unseen test dataset is used to evaluate model predictions. Some images in the Chest X-ray dataset have uncertain labels, denoted as value = −1, which are categorized into u-zero (negative) and u-one (treated as positive) based on prior research. For instance, Atelectasis and Edema labels are considered u-one, while the rest are u-zero. Augmentation techniques were applied to both the training and test datasets to increase size and quality, mitigate overfitting, and enhance model generalization during training and class label prediction.

	Model Name	AUROC	Accuracy	f1 Score
0	CustomNet	0.740989	0.862637	0.259044
1	DenseNet121	0.779881	0.866486	0.271429
2	ResNet50	0.716902	0.854088	0.219912
3	Inception	0.650021	0.846562	0.151924
4	Vgg16	0.671470	0.848543	0.164998

Figure 2. Test metrics overall summary.

The multi-label chest X-ray classification models demonstrated notable success, achieving an ROC of approximately 0.78 and an overall accuracy of around 87 percent. Among the tested models, DenseNet121 exhibited the highest proficiency in predicting test data, achieving accuracy exceeding 95 percent for various labels, including Fracture, Lung Lesion, Pleural Other, Pneumonia, and Pneumothorax.

Despite overall accuracy, models struggled with specific disease predictions due to training data imbalance, highlighting the need for better-balanced datasets to improve positive case predictions. Over sampling and Under sampling techniques and better addressing of uncertain labels for rare diseases can enhance model prediction capabilities.

III. Proposed Work

A. Notations.

In this section, we propose a multilabel chest xray classification model based on MobileNet. The following notations are used throughout this section:

X: The set of chest xray images.

Y: The set of chest xray labels.

x: A chest xray image.

y: A chest xray label vector.

y^ : The predicted label vector for a chest xray image.

f: The proposed multilabel chest xray classification model.

B. Network Architecture.

The proposed model is a sequential model that consists of the following layers:

The base model is MobileNet, configured with no top classification layer and initialized with no pre-trained weights

Global Average Pooling 2D Layer: After the base MobileNet model, a global average pooling layer is added. This layer reduces the spatial dimensions of the data while retaining important features.

Dropout Layer (0.5): A dropout layer is included with a dropout rate of 0.5 for regularization. This helps prevent overfitting by randomly dropping out a fraction of input units during training.

Dense Layer (512): A dense layer with 512 units is added, contributing to the model's capacity to capture intricate patterns in the data.

Another Dropout Layer (0.5): Another dropout layer is added for further regularization.

Dense Layer (Output Layer): The final dense layer has a number of units equal to the number of output classes , and it uses a sigmoid activation function.

C. Cost Function.

We use the binary cross-entropy loss function to train our multilabel image classification model. The binary cross-entropy loss function is a good choice for multilabel image classification problems because it is robust to class imbalance and can be used to train classifiers with different outputs. The binary cross-entropy loss function is able to handle class imbalance by weighting the loss for each label according to its frequency. This ensures that the model does not learn to ignore the less frequent labels. the binary cross-entropy loss function is also relatively easy to optimize. This is important because training multilabel image classification models can be computationally expensive.

$$\text{Loss} = -\frac{1}{\text{output size}} \sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i$$

y: is the ground truth label vector.

y^: is the predicted label vector.

IV. Experimental Details

A. Datasets

In the development of our model, we opted for a more manageable subset of the NIH Chest X-ray Dataset available on Kaggle. This curated subset consists of 5,606 images accompanied by corresponding labels, drawn from the larger dataset of 112,120 X-ray images featuring disease classifications from 30,805 unique patients. The challenge in leveraging chest X-ray examinations for computer-aided detection and diagnosis (CAD) lies in the scarcity of large, publicly available datasets with comprehensive annotations. The original dataset's labels were meticulously generated using Natural Language Processing techniques to extract disease information from associated radiological reports. Despite the unavailability of the original reports, the labels are asserted to be over 90% accurate, making them suitable for weakly-supervised learning. The decision to work with this subset was motivated by the practical constraints of computational resources, enabling efficient model development and training on a more manageable scale without compromising the integrity of the data.

In our study, we initially partitioned the dataset into training and validation sets through a random split. Addressing the challenge of imbalanced classes inherent in medical datasets, particularly in the context of chest X-ray images, we implemented a two-step approach. Firstly, we loaded the dataset into a dataframe and performed the random split to establish distinct training and validation sets. Subsequently, to mitigate the effects of class imbalance, we adopted a sampling strategy based on the number of classes represented in the dataset. This method, executed through the sample function, facilitated strategic sampling of datapoints, ensuring a more equitable representation of each class. By removing the bias associated with class imbalance, our approach aimed to enhance the robustness of our model and provide a more balanced and representative training dataset. This thoughtful two-step process in dataset preparation contributes to the overall effectiveness and reliability of model.

B. Training Details

During the training phase of our multi-disease classification model utilizing the MobileNet architecture, we meticulously partitioned the dataset into distinct training and validation sets, strategically prioritizing the evaluation of the model's capacity to generalize to unseen data. The model underwent training for a predefined number of epochs, precisely established at 10 iterations. Throughout this iterative process, we employed two pivotal callback functions

to bolster training efficiency and counteract overfitting.

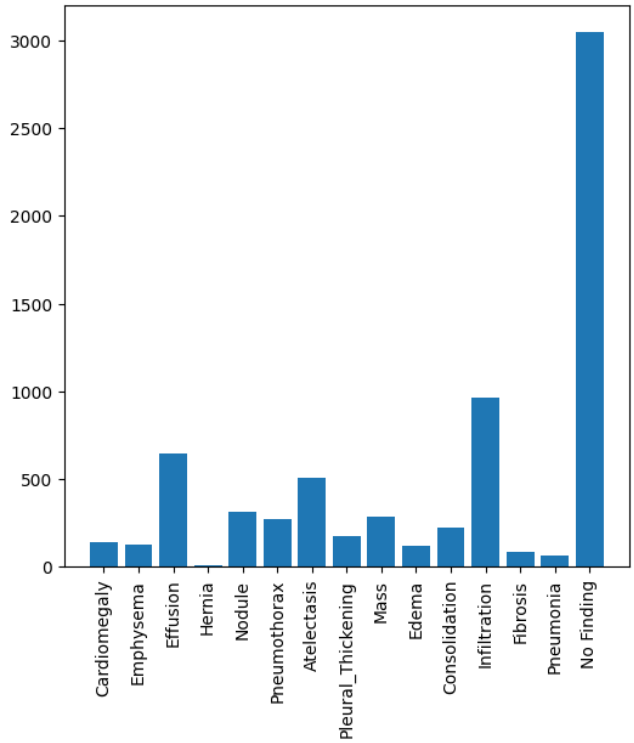

Figure 3. Different classes and their distribution.

The EarlyStopping callback, with a patience parameter set to five epochs, dictated that the training process would cease if there were no discernible improvements in validation performance over five consecutive epochs. This strategic decision aimed to mitigate overfitting, preventing the model from excessively adapting to noise within the training data. Concurrently, the ReduceLROnPlateau callback, configured with a reduction factor of 0.1 and a patience of 2, dynamically adjusted the learning rate during training. This adaptive mechanism facilitated adjustments to the model's weights, reducing the learning rate by one-tenth when validation performance plateaued for two epochs. By combining EarlyStopping and ReduceLROnPlateau, our approach not only guarded against overfitting but also optimized the training process, culminating in a more robust and widely applicable multi-disease classification model tailored for chest X-ray images.

C. Baseline Methods

The multi-disease classification model, utilizing the MobileNet architecture, comprises a MobileNet base model, which is initially pre-trained for visual recognition tasks and is adept at extracting hierarchical features from images. This base model is succeeded by a Global Average Pooling 2D layer, strategically applied for dimensionality reduction, facilitating the retention of essential features. To prevent overfitting, two Dropout layers, each with a 50% dropout rate, are integrated on either side of a densely connected layer with 512 units. This densely connected layer serves as a feature extractor, specializing in capturing intricate patterns within the data. Another Dropout layer follows to enhance regularization. The ultimate dense layer, acting as the output layer, is tailored for binary classification through a sigmoid activation function. This final layer predicts the probability of each label independently. The model's compilation involves the use of the Adam optimizer, binary cross-entropy loss function, and performance metrics such as binary accuracy and mean absolute error. This comprehensive architecture aims to optimize feature extraction, minimize overfitting, and yield accurate predictions in the context of multi-disease classification.

Layer (type)	Output Shape	Param #
mobilenet_1.00_128 (Functional)	(None, 4, 4, 1024)	3228864
global_average_pooling2d_1 (GlobalAveragePooling2D)	(None, 1024)	0
dropout_2 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 512)	524800
dropout_3 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 15)	7695

Total params: 3761359 (14.35 MB)
Trainable params: 3739471 (14.26 MB)
Non-trainable params: 21888 (85.50 KB)

Figure 4. Proposed model layout.

D. Evaluation Metrics.

The Area Under the Receiver Operating Characteristic curve (AUC-ROC) is a widely used metric in machine learning and binary classification tasks to evaluate the performance of a predictive model. The ROC curve is a graphical representation that illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across various thresholds for a classifier. AUC-ROC quantifies the discriminatory power of a model by calculating the area under this curve. In the context of the ROC curve, true positives (TP) represent instances where the model correctly predicts the positive class, and false positives (FP) indicate instances incorrectly classified as positive. The true positive rate (Sensitivity or Recall) is the ratio of true positives to the total actual positive instances, while the false positive rate is the ratio of false positives to the total actual negative instances. A higher AUC-ROC score, closer to 1,

signifies superior model performance, indicating a better balance between true positive rate and false positive rate across different classification thresholds. Understanding the nuances of true positives and false positives within the ROC curve is crucial for comprehensively evaluating and interpreting the effectiveness of classification models.

V. Results

In this section, we present the results of our experiments on the multilabel image classification model. We evaluated the model on a random sampled dataset and a balanced dataset. The following pictures shows the model accuracy, model loss, and AUROC results for both datasets:

Imbalanced Dataset:

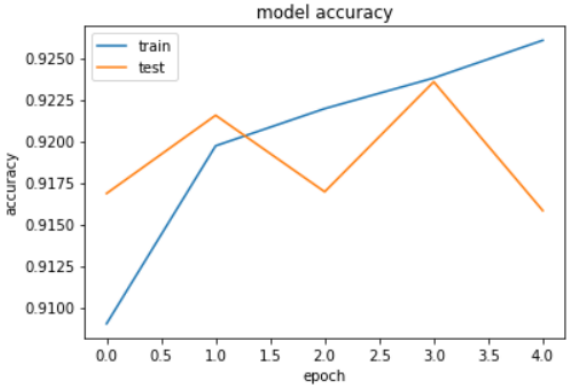

Figure 5. Training accuracy at different epochs.

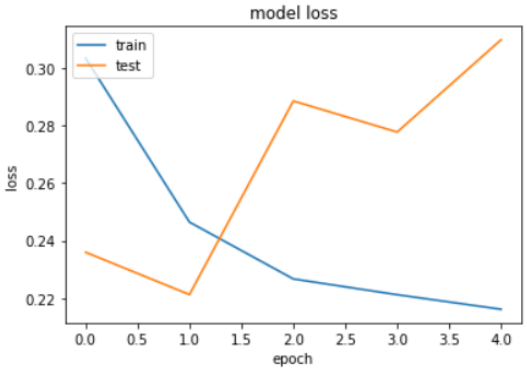

Figure 6. Loss observed at different epochs.

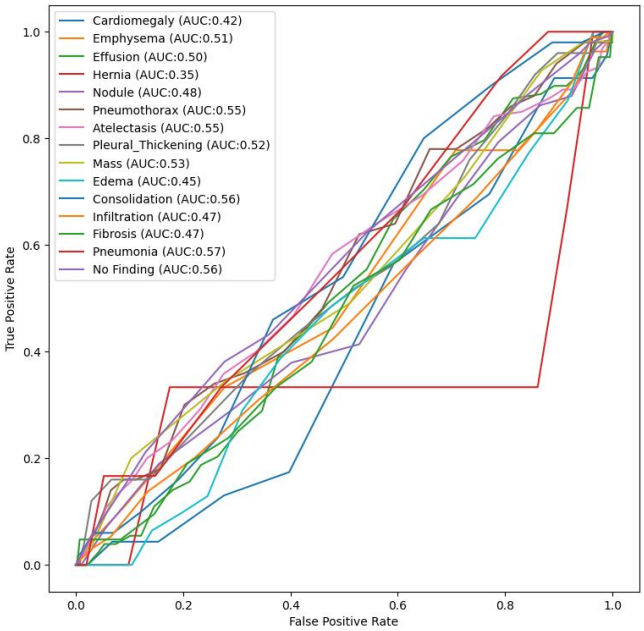

Figure 7. AUC of different classes from ROC curve.

Balanced Dataset:

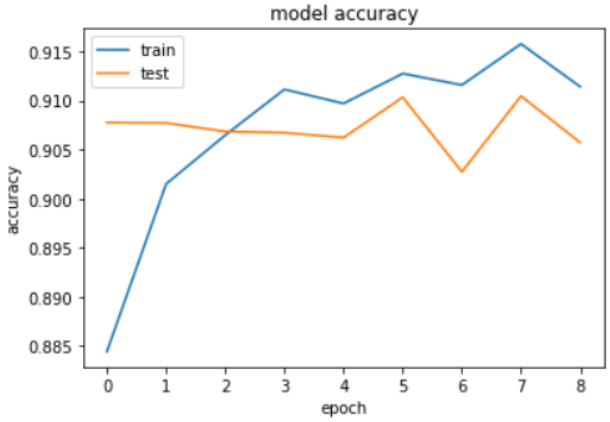

Figure 8. Training accuracy at different epochs.

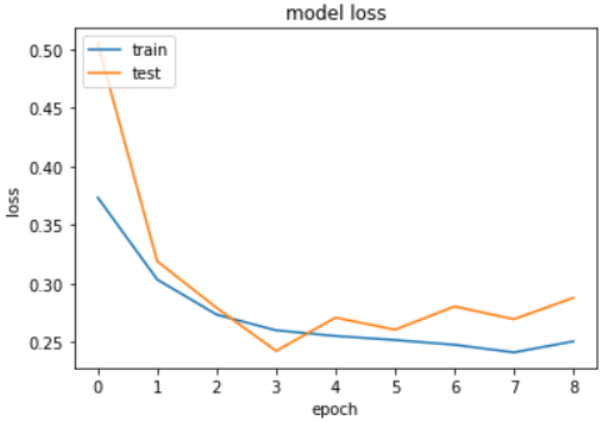

Figure 9. Loss observed at different epochs.

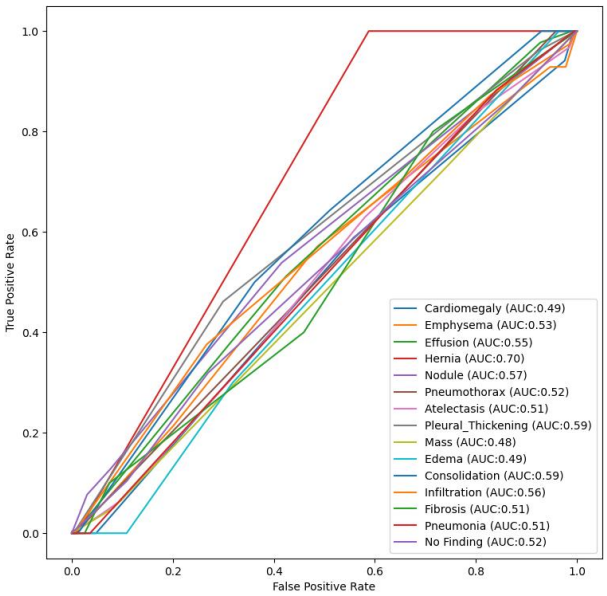

Figure 10. AUC of different classes from ROC curve

In our effort to boost the model's effectiveness, we applied a targeted sampling strategy to the dataset. This approach, implemented through the sample function, aimed to address imbalances in class representation during training. The outcome was notable improvement in the model's performance for less common classes, such as Hernia. However, this improvement came at the expense of reduced accuracy for other classes, as the downsized dataset impacted their predictive capabilities. This trade-off is evident in the comparison of Area Under the Receiver Operating Characteristic (AUC-ROC) curves between the two sampling methods. The tailored sampling approach, while enhancing sensitivity to minority classes, underscores the need for a nuanced balance between rectifying class imbalances and sustaining overall model accuracy

VI. Conclusion

In summary, our study introduces a straightforward and easily trainable model tailored to resource limitations and time constraints. While the model's average AUC score of 0.58 falls below the state-of-the-art model's 0.78, it underscores the importance of balancing simplicity with resource-intensive alternatives. Despite its reduced scale, our model consistently aligns with the outcomes of more complex counterparts. The introduced weight-based sampling technique successfully improved the AUC for less represented classes, albeit at the expense of overall model accuracy due to the scarcity of images in these classes. A more extensive and diverse dataset may yield enhanced results, but given our resource constraints, this dataset forms the basis of our study.

VII. Refferences

- Holste, G., Wang, S., Jiang, Z., Shen, T. C., Shih, G., Summers, R. M., Peng, Y., & Wang, Z. (2022). Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study. Data Augmentation, Labelling, and Imperfections : Second MICCAI Workshop, DALI 2022, Held in Conjunction With MICCAI 2022, Singapore, September 22, 2022, Proceedings. DALI (Workshop) (2nd : 2022 : Singapore), 13567, 22. https://doi.org/10.1007/978-3-031-17027-0_3
- Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., & Rohban, M. H. (2022). SwinCheX: Multi-label classification on chest X-ray images with transformers. arxiv.org/abs/2206.04246
- Multi-Label Chest X-Ray Classification via Deep Learning. https://paperswithcode.com/paper/multi-label-chest-x-ray-classification-via
- Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1), 1-8. https://doi.org/10.1038/s41597-019-0322-0
- Random Sample of NIH Chest X-ray Dataset(kaggle). https://www.kaggle.com/datasets/nih-chest-xrays/sample