CS633 Final Project: Improving Generalization of Deep AUC Maximization for Medical Image Classification

Ada Huang
Department of Computer Science & Engineering
Texas A&M University University
College Station, Texas 77843
adahuang@tamu.edu

Abstract

Deep AUC Maximization (DAM) has shown promising results in various machine learning applications, including medical image classification tasks. However, its generalization capability is limited when applied to small datasets, often leading to overfitting. In this study, I aimed to improve the generalization of DAM for medical image classification tasks on small datasets. Using the LibAUC library, I conducted experiments on seven medical image classification tasks from the MedMNIST dataset. I investigated different network structures to enhance the performance of DAM on these tasks, with a focus on improving benchmark results reported in the MedMNIST paper. My comprehensive approach involved exploring multiple directions for improvement and demonstrating innovative ideas to enhance the generalization of LibAUC. The findings of this study contribute to the development of more robust and generalizable DAM-based models for medical image classification tasks, particularly when dealing with small datasets.

1 Introduction

Deep AUC Maximization (DAM) has emerged as a powerful technique for learning deep neural networks by maximizing the AUC score of a model on a dataset. Despite its successful applications in various machine learning tasks, its generalization ability is often hindered when applied to small datasets, resulting in overfitting. This study aims to address this limitation and enhance the generalization of DAM for medical image classification tasks, specifically when working with small datasets.

To achieve this goal, I conducted experiments on seven medical image classification tasks from the MedMNIST dataset, utilizing the LibAUC library. In order to improve the performance of DAM on these tasks, I explored different network structures while focusing on surpassing the benchmark results reported in the MedMNIST paper(1). For 2-D datasets, such as BreastMNIST, PneumoniaMNIST, and ChestMNIST, I employed the ResNet-18 architecture(2), while for 3-D datasets, including NoduleMNIST3D, AdrenalMNIST3D, VesselMNIST3D, and SynapseMNIST3D, I utilized ResNet-18 combined with a 3D extension(3).

Through a comprehensive approach involving multiple directions of improvement, I demonstrated innovative ideas to enhance the generalization of LibAUC(4). The findings of this study contribute to the development of more robust and generalizable DAM-based models for medical image classification tasks, with a particular emphasis on addressing the challenges posed by small datasets.

2 2-D Data Analysis

In this section, I analyze three 2-D medical image datasets, namely BreastMNIST, PneumoniaMNIST, and ChestMNIST. Each of these datasets comprises 2-D images representing a range of medical conditions. The objective is to accurately classify these images based on their respective labels.

To ensure consistency and enhance the performance of my models, I apply the same data augmentation techniques to all three datasets. These techniques include:

- Random rotation with a maximum angle of 15 degrees,
- Random horizontal flipping,
- Random resized cropping to a size of 28x28 pixels, with a scale factor between 0.8 and 1.2,
- Converting images to tensors, and
- Normalizing the tensor values with a mean of 0.1307 and a standard deviation of 0.3081.

Applying these data augmentation methods helps improve the robustness and generalization capabilities of my models, allowing them to better recognize and classify the medical images in each dataset.

For this study, I use the ResNet-18 (28) architecture on all three datasets. The ResNet-18 (28) model is first pre-trained using Deep AUC Maximization (DAM), an optimization technique for the Area Under the Receiver Operating Characteristic Curve (AUROC), which is especially relevant for imbalanced datasets and medical applications.

After pretraining with DAM, I fine-tune the model using LibAUC, a library specializing in AUC-based learning objectives optimization. I explore two ResNet-18 configurations: one with dropout layers and one without. Dropout serves as a regularization technique to prevent overfitting by randomly dropping neurons during training.

By utilizing the pre-trained ResNet-18 (28) model and fine-tuning it with LibAUC, I aim to enhance the classification performance of BreastMNIST, PneumoniaMNIST, and ChestMNIST datasets. Comparing models with and without dropout layers allows me to evaluate the impact of regularization on classification accuracy and generalization capabilities.

2.1 BreastMNIST

2.1.1 Dataset and Proprocessing

BreastMNIST is a dataset derived from a collection of 780 breast ultrasound images, which are divided into three categories: normal, benign, and malignant. Given that low-resolution images are used in this dataset, the classification task is simplified to a binary classification problem. To achieve this, the normal and benign images are combined into one class (positive) and contrasted against malignant images (negative). The dataset is split with a $7:1:2$ ratio for training, validation, and test sets, resulting in 546 training samples, 78 validation samples, and 156 test samples. Originally sized at $1 \times 500 \times 500$, the source images have been resized to $1 \times 28 \times 28$.

2.1.2 Hyperparameters

To optimize the performance of the DAM and LibAUC models, I carefully selected appropriate hyperparameters for the dataset. The chosen hyperparameters for DAM and LibAUC are in Table 1 and Table 2, respectively. The optimizer here for *BreastMNIST* is stochastic gradient descent (SGD) since SGD performs well on small datasets.

2.1.3 Experiments and Results

In this experiment and results section, I summarize the performance of my models trained using DAM and LibAUC. The best test AUC obtained with DAM was 0.8835, with an accuracy of 67.7308% at epoch 89 (Figure 1).

I leveraged the pretrained ResNet-18 (28) model and fine-tuned it using LibAUC, I experimented with ResNet-18 models with and without dropout. With a dropout rate of 0.5, the ResNet-18 model

Table 1: BreastMNIST Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	200
Batch Size	64
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	SGD
Momentum	0.9
Weight Decay	0.0001

Table 2: BreastMNIST Hyperparameters for LibAUC ResNet-18 with and without Dropout

Hyperparameter	Value
Number of Epochs	200
Batch Size	128
Learning Rate	0.001
Epoch Decay	0.002
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG
Dropout Rate	0.7 (not dropout rate for without dropout ResNet-18)

achieved a test AUC of 0.84. When the dropout rate was increased to 0.7, the model achieved a test AUC of 0.8350 and an accuracy of 55.3974% at epoch 108 (Figure 2). For the ResNet-18 model without dropout, the best test AUC was 0.8499, with an accuracy of 57.1667% at epoch 118 (Figure 3).

The DAM approach did not result in a significant improvement in the benchmark AUC, but it did result in higher test AUC scores compared to the LibAUC-based ResNet-18 models. However, the impact of the dropout regularization technique can still be observed in the results, with models using dropout showing improved performance. The difference in performance between models with and

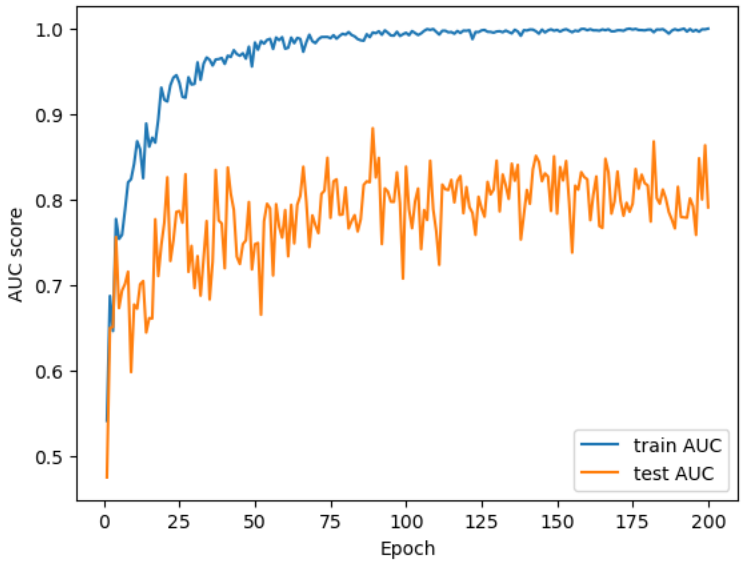

Figure 1: BreastMNIST Train and Test AUC Scores vs Epoch using DAM

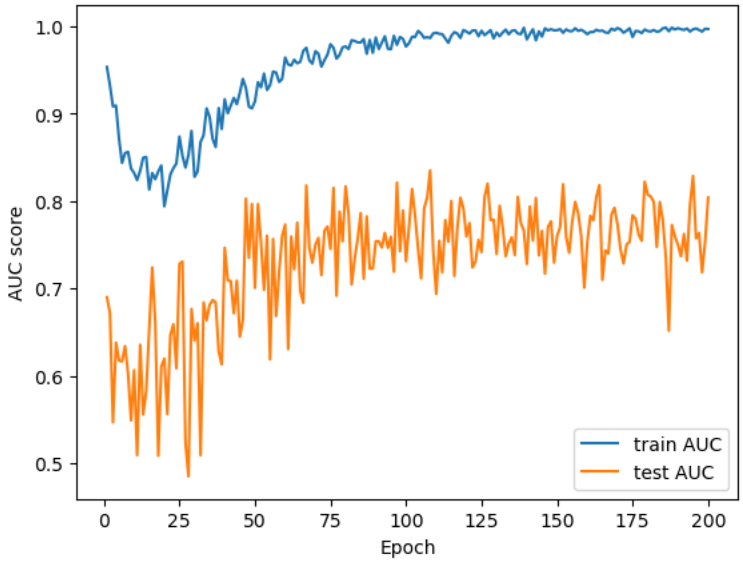

Figure 2: BreastMNIST Train and Test AUC Scores vs Epoch using LibAUC with Dropout

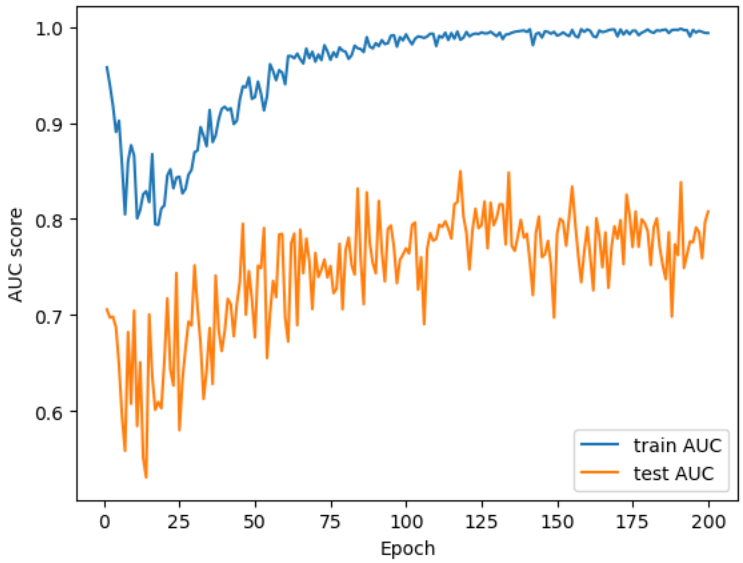

Figure 3: BreastMNIST Train and Test AUC Scores vs Epoch using LibAUC without Dropout

without dropout underscores the importance of regularization in enhancing model generalization capabilities.

2.2 PneumoniaMNIST

2.2.1 Dataset and Preprocessing

The *PneumoniaMNIST* dataset is based on 5,856 pediatric chest X-Ray images and is used for binary-class classification of pneumonia against normal. The source training set was split with a 9:1 ratio into a training set and a validation set, and the source validation set was used as the test set. The images are grayscale and had sizes of $(384 - 2,916) \times (127 - 2,713)$, but were center-cropped and resized to $1 \times 28 \times 28$. The training set consists of 4,708 samples, the validation set consists of 524 samples, and the test set consists of 624 samples.

2.2.2 Hyperparameters

To enhance the performance of the DAM and LibAUC models, I selected suitable hyperparameters tailored to the dataset. The specific hyperparameters chosen for DAM can be found in Table 3, while those for LibAUC are presented in Table 4.

Table 3: PneumoniaMNIST Hyperparameters for DAM

Hyperparameter	**Value**
Number of Epochs	50
Batch Size	128
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	SGD
Momentum	0.9
Weight Decay	0.0001

Table 4: PneumoniaMNIST Hyperparameters for LibAUC ResNet-18 with and without Dropout

Hyperparameter	**Value**
Number of Epochs	50
Batch Size	128
Learning Rate	0.001
Epoch Decay	0.002
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG
Dropout Rate	0.7 (not dropout rate for without dropout ResNet-18)

2.2.3 Experiments and Results

When comparing the results of my experiments with the benchmark performance, which has an AUC of 0.944 and an accuracy of 85.4%, it becomes evident that my models with DAM surpass the benchmark AUC while falling short in terms of accuracy.

I evaluated the performance of the Adam and SGD optimizers along with the LibAUC model using various configurations. The results obtained from these experiments are presented in the following:

For the Adam optimizer, the best test AUC obtained with DAM was 0.9686, with an accuracy of 72.4167% at epoch 42. On the other hand, for the SGD optimizer, the best test AUC obtained with DAM was 0.9722, with an accuracy of 70.6442%. The results indicate that the Adam optimizer outperformed the SGD optimizer in terms of accuracy, while the SGD optimizer achieved a slightly higher AUC (Figure 4).

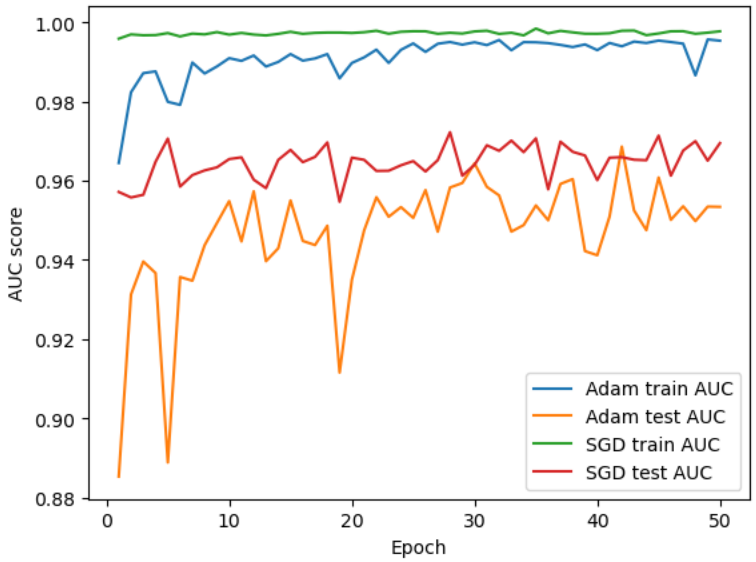

Figure 4: PneumoniaMNIST Train and Test AUC Scores vs Epoch using DAM

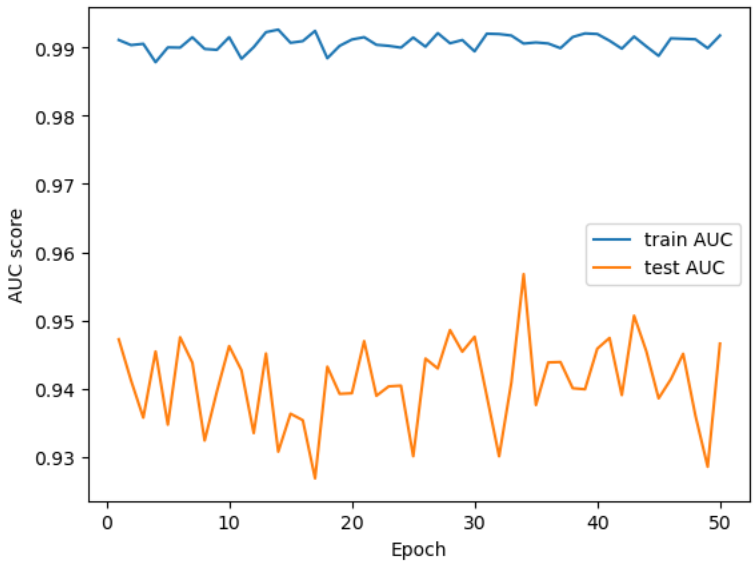

Figure 5: PneumoniaMNIST Train and Test AUC Scores vs Epoch using LibAUC with Dropout

In addition, I leveraged the pretrained ResNet-18 (28) model from DAM and fine-tuned it using LibAUC, I experimented with ResNet-18 models with and without dropout. With a dropout rate of 0.7, and batch size 64, the ResNet-18 model achieved a test AUC of 0.9545 and an accuracy of 63.2756% at epoch 8. When the batch size was increased to 128, the model achieved a test AUC of 0.9568 and an accuracy of 62.3526% at epoch 34. The test AUC for batch size 128 is slightly higher than that for batch size 64 (Figure 5). For the ResNet-18 model without dropout, the best test AUC was 0.9613, with an accuracy of 61.8045% at epoch 19 (Figure 6).

Although the models' AUC performance surpassed the benchmark, the accuracy scores were consistently lower. This discrepancy highlights the need for further refinement and optimization of the models to improve their overall performance and achieve more balanced results across both AUC and accuracy metrics.

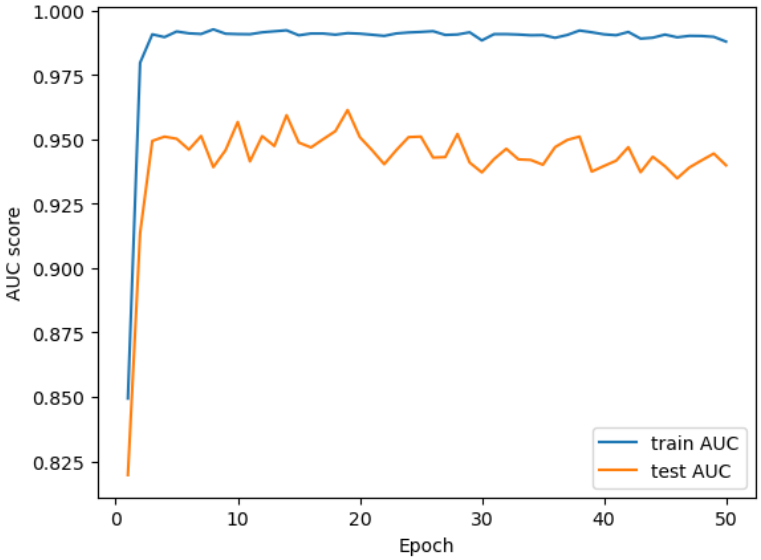

Figure 6: PneumoniaMNIST Train and Test AUC Scores vs Epoch using LibAUC without Dropout

2.3 ChestMNIST

2.3.1 Dataset and Preprocessing

The *ChestMNIST* dataset is based on the NIH-ChestXray14 dataset. This dataset contains 112,120 frontal-view X-ray images of 30,805 unique patients and is accompanied by 14 text-mined disease labels. These labels can be formulated as a multi-label binary-class classification task. The dataset is split into three parts: training, validation, and testing sets, with 78,468, 11,219, and 22,433 data points, respectively.

Each image consists of a single channel, and the labels are defined as follows: {'0': 'atelectasis', '1': 'cardiomegaly', '2': 'effusion', '3': 'infiltration', '4': 'mass', '5': 'nodule', '6': 'pneumonia', '7': 'pneumothorax', '8': 'consolidation', '9': 'edema', '10': 'emphysema', '11': 'fibrosis', '12': 'pleural', '13': 'hernia'}.

The original source images with dimensions of $1 \times 1024 \times 1024$ were resized to $1 \times 28 \times 28$ to reduce computational complexity while maintaining the essential information required for classification. The official data split was used to ensure consistency and comparability with other studies utilizing the same dataset.

2.3.2 Hyperparameters

The specific hyperparameters chosen for DAM can be found in Table 5, while those for LibAUC are presented in Table 6.

Table 5: ChestMNIST Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	200
Batch Size	64
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	Adam
Weight Decay	0.0001

Table 6: ChestMNIST Hyperparameters for LibAUC ResNet-18

Hyperparameter	Value
Number of Epochs	50
Batch Size	128
Learning Rate	0.1
Epoch Decay	0.003
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG

2.3.3 Experiments and Results

Using the DAM approach, the highest test AUC of 0.5487 and accuracy of 63.88% were attained at epoch 19. Upon observing the imbalanced nature of the data (Table 7), I decided to incorporate the class weighting technique with the ResNet-18 model to address this issue. This method helps adjust the model's learning process by assigning different weights to each class, thus reducing the impact of the dominant classes and enabling better generalization.

However, the best test AUC of 0.4916, achieved at epoch 16 after applying class weighting, did not result in an improvement compared to the highest test AUC obtained before implementing class weighting. This suggests that, in this case, the class weighting technique may not have effectively addressed the imbalanced data issue or improved the model's performance.

Therefore, I opted to utilize the pre-trained ResNet 18 model from ImageNet for the classification of the ChestMINST dataset. This approach led to the best test AUC of 0.5552 and an accuracy of 63.92% at epoch 188, demonstrating an improvement in the model's overall performance (Figure 7).

Table 7: Label counts for diseases in the ChestMNIST dataset

Label Index	Disease	Count
0	Atelectasis	50401
1	Cardiomegaly	1686
2	Effusion	6445
3	Infiltration	9483
4	Mass	2313
5	Nodule	2292
6	Pneumonia	307
7	Pneumothorax	1935
8	Consolidation	1060
9	Edema	503
10	Emphysema	676
11	Fibrosis	535
12	Pleural	761
13	Hernia	71

Furthermore, I leveraged the pre-trained ResNet-18 (28) model from DAM and fine-tuned it using LibAUC. The best test AUC 0.4181 was achieved at epoch 16 (Figure 8). This result was lower than the best AUC obtained with DAM.

In comparison with the benchmark performance, which has an AUC of 0.768 and an accuracy of 94.7%, my models did not surpass the benchmark in either metric. The highest AUC achieved with my models was 0.5552 using the pre-trained ResNet 18 model from ImageNet, while the benchmark AUC is considerably higher at 0.768. Additionally, the highest accuracy attained was 63.92%, which is significantly lower than the benchmark accuracy of 94.7%.

These findings suggest that further refinements and optimizations are necessary to improve the models' performance and reach or surpass the benchmark metrics. It is important to note that the task at hand is a multi-class classification problem, which inherently presents additional challenges and

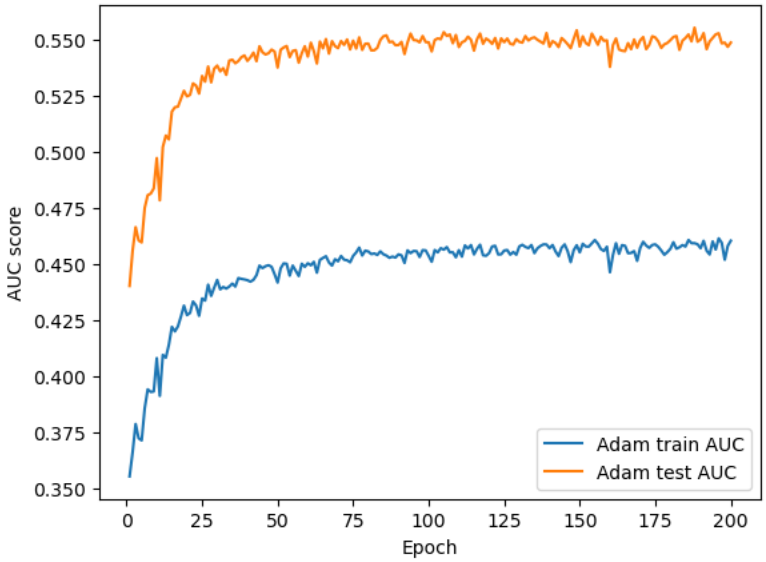

Figure 7: ChestMNIST Train and Test AUC Scores vs Epoch using DAM

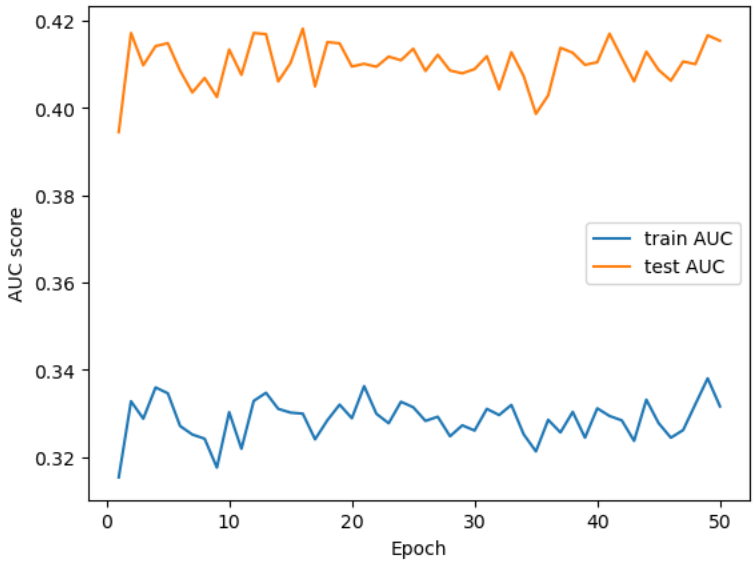

Figure 8: ChestMNIST Train and Test AUC Scores vs Epoch using LibAUC

complexity due to the multiple disease classes involved. Considering this aspect, exploring alternative approaches and techniques tailored to multi-class classification problems might yield better insights and potentially enhance the performance of my models.

3 3-D Data Analysis

In this section, I analyze four 3-D medical image datasets, namely NoduleMNIST3D, AdrenalMNIST3D, VesselMNIST3D, and SynapseMNIST3D. Each of these datasets comprises 3-D images representing a range of medical conditions. The objective is to accurately classify these images based on their respective labels.

To ensure consistency and enhance the performance of my models, I apply the same data augmentation techniques to all four datasets. Initially, I employed the following basic data augmentation techniques::

- Resampling to a 1x1x1 voxel size,
- Cropping or padding to a size of 28x28x28 voxels,
- Converting images to the canonical orientation.

However, the initial results were not satisfactory, leading me to incorporate additional augmentation techniques for better performance. These techniques include:

- Random affine transformations with scales between 0.9 and 1.1, rotations between -10 and 10 degrees, and translations between -0.1 and 0.1, all with a probability of 0.5,
- Random elastic deformation with a maximum displacement of 5 voxels in each direction, a grid of 5x5x5 control points, locked borders, and a probability of 0.5,
- Random flipping along all three axes with a probability of 0.5, and
- Adding random noise with a standard deviation between 0.01 and 0.05 and a probability of 0.5.

Applying these data augmentation methods helps improve the robustness and generalization capabilities of my models, allowing them to better recognize and classify the medical images in each dataset.

For this study, I use the ResNet-18 3D architecture on all four datasets. The ResNet-18 3D model is first pre-trained using Deep AUC Maximization (DAM), an optimization technique for the Area Under the Receiver Operating Characteristic Curve (AUROC), which is especially relevant for imbalanced datasets and medical applications.

After pretraining with DAM, I fine-tune the model using LibAUC, a library specializing in AUC-based learning objectives optimization.

By utilizing the pre-trained ResNet-18 3D model and fine-tuning it with LibAUC, I aim to enhance the classification performance of NoduleMNIST3D, AdrenalMNIST3D, VesselMNIST3D, and SynapseMNIST3D datasets.

3.1 NoduleMNIST3D

3.1.1 Dataset and Preprocessing

The *NoduleMNIST3D* dataset is derived from the LIDC-IDRI, a large public lung nodule dataset containing thoracic CT scan images. The dataset is designed for both lung nodule segmentation and 5-level malignancy classification tasks. In this study, the focus is on binary classification, where cases with malignancy levels 1 and 2 are categorized as the negative class (benign), and cases with malignancy levels 4 and 5 are categorized as the positive class (malignant). Cases with a malignancy level of 3 are ignored for this task.

The *NoduleMNIST3D* dataset consists of 1,633 samples, split into training, validation, and test sets with a ratio of 7:1:2. The training set has 1,158 samples, the validation set has 165 samples, and the test set has 310 samples. Each sample has one channel and is spatially normalized with a spacing of $1mm \times 1mm \times 1mm$. The images are center-cropped to a size of $28 \times 28 \times 28$ voxels.

3.1.2 Hyperparameters

The specific hyperparameters chosen for DAM can be found in Table 8, while those for LibAUC are presented in Table 9.

Table 8: NoduleMNIST3D Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	Adam
Weight Decay	0.0001

Table 9: NoduleMNIST3D Hyperparameters for LibAUC

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Epoch Decay	0.003
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG

3.1.3 Experiments and Results

I first applied the DAM approach, where the best AUC of 0.8962 and accuracy of 21.14% were achieved at epoch 3. However, the accuracy was not satisfactory, and I decided to include additional data augmentation techniques to address this issue. With these additional augmentations, the best AUC improved to 0.9201, and the accuracy slightly increased to 21.18% at epoch 25. Despite the improvement in AUC, the training appeared to be overfitting, and the accuracy remained relatively low.

To further improve the model's performance, I incorporated the class weighting technique into the cross-entropy loss, taking into account the imbalanced dataset with 295 positive samples and 863 negative samples. This approach led to a reduction in overfitting during training, but the test accuracy did not show significant improvement. The best model achieved an AUC of 0.9207 and an accuracy of 19.69% at epoch 20, with only a slight improvement in AUC (Figure 9).

Next, I used the pre-trained ResNet-18 3D model from DAM and fine-tuned it with LibAUC. The best test AUC reached 0.9240, and the accuracy was 17.15% at epoch 27 (Figure 10). This approach resulted in a more considerable improvement in AUC compared to previous experiments.

When comparing my results to the benchmark, which has an AUC of 0.863 and an accuracy of 84.4%, I observe that my model surpasses the benchmark in terms of AUC, demonstrating the model's enhanced ability to distinguish between classes. However, the accuracy remains significantly lower than the benchmark, indicating the need for further optimization and refinements to improve the overall performance.

3.2 AdrenalMNIST3D

3.2.1 Dataset and Preprocessing

The *AdrenalMNIST3D* dataset is a newly developed 3D shape classification dataset containing shape masks from 1,584 left and right adrenal glands, corresponding to 792 patients. These shape masks were collected from *Zhongshan Hospital Affiliated to Fudan University* and annotated by expert

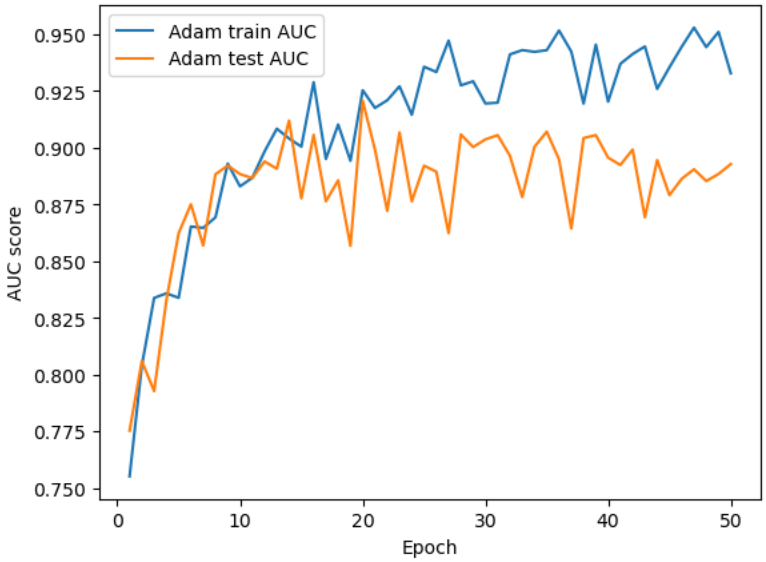

Figure 9: NoduleMNIST3D Train and Test AUC Scores vs Epoch using DAM

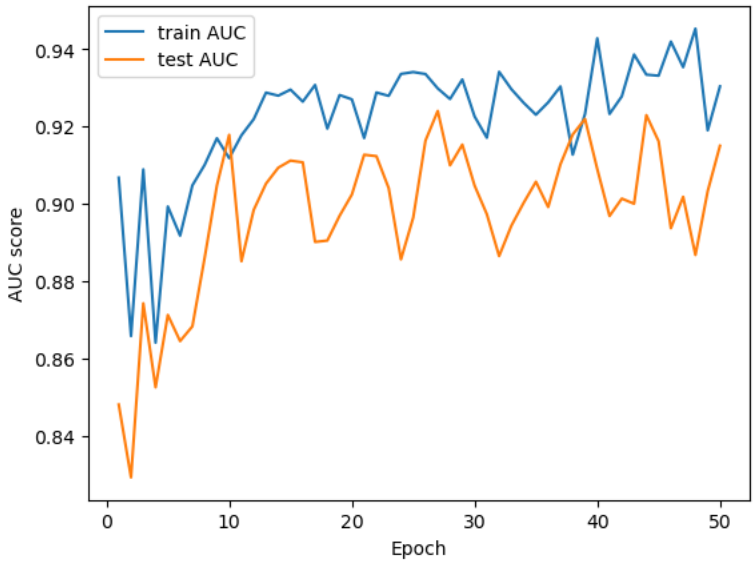

Figure 10: NoduleMNIST3D Train and Test AUC Scores vs Epoch using LibAUC

endocrinologists using abdominal computed tomography (CT) scans. Each 3D adrenal gland shape is labeled as either a normal adrenal gland or an adrenal mass (hyperplasia).

To maintain patient privacy, the source CT scans are not provided. Instead, the dataset comprises the real 3D shapes of adrenal glands and their binary classification labels. The center of each adrenal gland is calculated, and the center-cropped volume of $64mm \times 64mm \times 64mm$ is resized to $28 \times 28 \times 28$.

The dataset is randomly split at the patient level into a training set of 1,188 samples, a validation set of 98 samples, and a test set of 298 samples. The dataset includes one channel, and the labels are defined as follows: '0' represents normal adrenal glands, while '1' represents adrenal hyperplasia.

3.2.2 Hyperparameters

The specific hyperparameters chosen for DAM can be found in Table 10, while those for LibAUC are presented in Table 11.

Table 10: AdrenalMNIST3D Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	Adam
Weight Decay	0.0001

Table 11: AdrenalMNIST3D Hyperparameters for LibAUC

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Epoch Decay	0.003
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG

3.2.3 Experiments and Results

In the experiments for the AdrenalMNIST3D dataset, several approaches were employed to optimize the model's performance. Notably, the dataset exhibited class imbalance with 259 positive samples and 929 negative samples. To address this issue, the class weighting technique was incorporated into the model training process.

Initially, the Deep AUC Maximization (DAM) approach was applied with a batch size of 64, yielding a test AUC of 0.8531 and an accuracy of 41.2483% at epoch 7. Next, the batch size was increased to 128, resulting in the best test AUC of 0.8324 and an accuracy of 64.1544% at epoch 16.

To further enhance the model's performance, more data augmentation techniques and class weighting were introduced, and the batch size was adjusted to 32. This led to a significant improvement in the test AUC and accuracy, reaching 0.8980 and 68.5570%, respectively, at epoch 45 (Figure 11).

After pretraining the model with DAM, the pre-trained ResNet-18 3D model was fine-tuned using LibAUC. This approach achieved the highest test AUC of 0.8984 but a reduced accuracy of 15.8188% at epoch 49 (Figure 12).

When compared to the benchmark metrics of AUC 0.827 and ACC 0.721, my experiments demonstrated a notable improvement in AUC performance but fell short of achieving higher accuracy. Overall, the optimizations and techniques applied to the model resulted in considerable progress

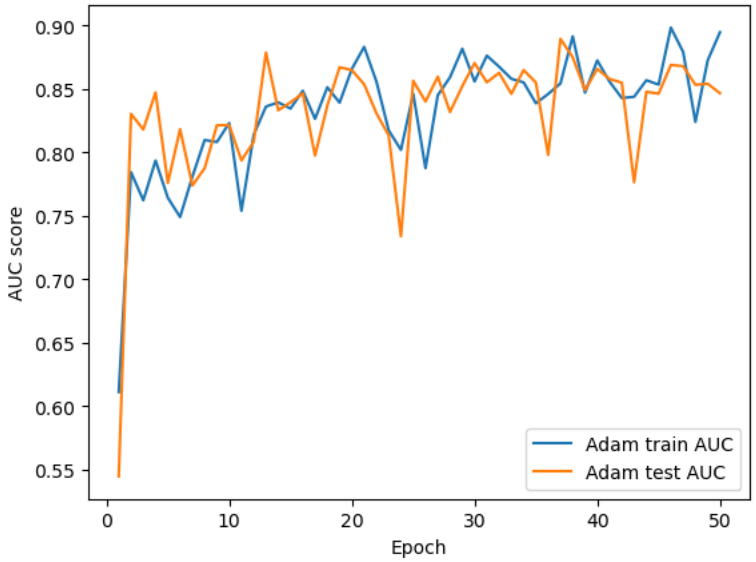

Figure 11: AdrenalMNIST3D Train and Test AUC Scores vs Epoch using DAM

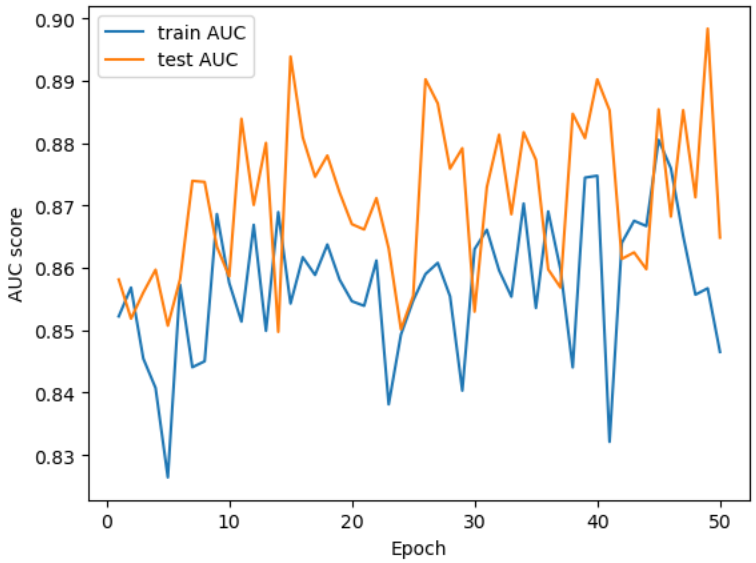

Figure 12: AdrenalMNIST3D Train and Test AUC Scores vs Epoch using LibAUC

in terms of AUC, suggesting their effectiveness in enhancing the classification performance for the AdrenalMNIST3D dataset.

3.3 VesselMNIST3D

3.3.1 Dataset and Preprocessing

The *VesselMNIST3D* dataset is a binary-class dataset derived from the open-access 3D intracranial aneurysm dataset, *IntrA*. *IntrA* comprises 103 3D models of entire brain vessels, reconstructed from MRA images. The dataset contains 1,694 healthy vessel segments labeled as '0' (vessel) and 215 aneurysm segments labeled as '1' (aneurysm). These segments were generated automatically from the complete models.

To preprocess the dataset, non-watertight meshes were fixed using *PyMeshFix*, and watertight meshes were voxelized into $28 \times 28 \times 28$ voxels using *trimesh*. The source dataset was then divided into training, validation, and test sets with a $7 : 1 : 2$ ratio. This resulted in the following distribution of samples: Train: 1,335 datapoints, Validation: 192 datapoints, and Test: 382 datapoints. The task for this dataset is binary classification, with each data point containing a single channel.

3.3.2 Hyperparameters

The specific hyperparameters chosen for DAM can be found in Table 12, while those for LibAUC are presented in Table 13.

Table 12: VesselMNIST3D Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	Adam
Weight Decay	0.0001

Table 13: VesselMNIST3D Hyperparameters for LibAUC

Hyperparameter	Value
Number of Epochs	50
Batch Size	64
Learning Rate	0.001
Epoch Decay	0.003
Weight Decay	0.00001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG

3.3.3 Experiments and Results

I started with the DAM approach to address the VesselMNIST3D dataset's binary classification task. The initial model achieved a best AUC of 0.9601 and an accuracy of 56.8377% at epoch 17. This demonstrated that the DAM approach provided a solid foundation for differentiating between vessel and aneurysm classes.

To further improve the model, I introduced additional data augmentations and applied a class weighting technique to the cross-entropy loss function. I implemented this technique to address the class imbalance in the dataset, where there were 150 positive samples (aneurysm) and 1185 negative samples (vessel). By incorporating these adjustments, the model's performance significantly improved, resulting in a best AUC of 0.9845 and an accuracy of 51.6178% at epoch 29 (Figure 13).

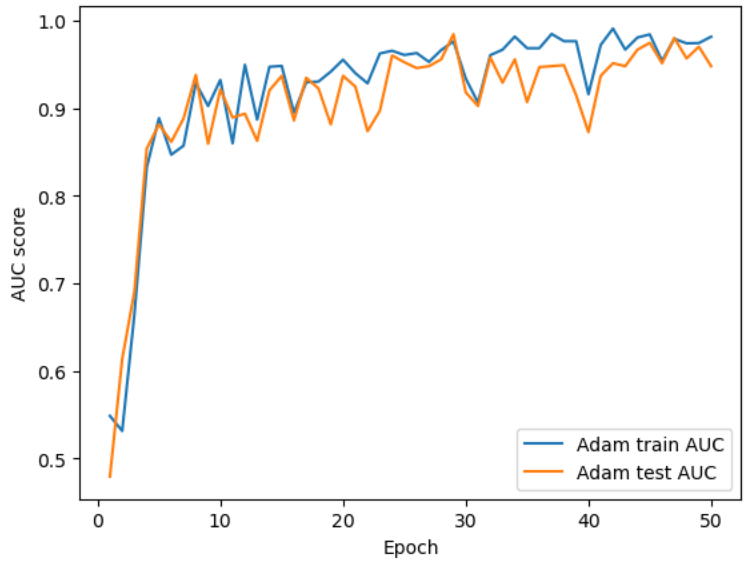

Figure 13: VesselMNIST3D Train and Test AUC Scores vs Epoch using DAM

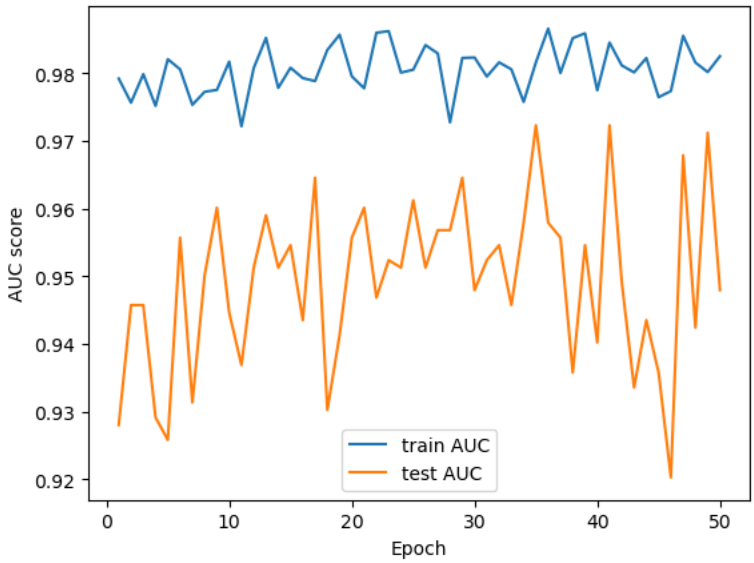

Figure 14: VesselMNIST3D Train and Test AUC Scores vs Epoch using LibAUC

These improvements indicate that the additional data augmentations and class weighting effectively enhanced the model's ability to distinguish between the two classes.

Following the success of the DAM approach, I decided to use the pre-trained ResNet-18 3D model from DAM in conjunction with the LibAUC approach. This combination further improved the model's performance, achieving the best test AUC of 0.9723 and an accuracy of 39.1885% at epoch 41 (Figure 14). This result suggests that the integration of the pre-trained ResNet-18 3D model with LibAUC contributed to the model's enhanced classification capabilities.

When comparing the results to the benchmark AUC of 0.874 and accuracy of 0.877, it is evident that my approach demonstrates a significant improvement in the AUC performance. The models I have developed provide a more effective solution for distinguishing between vessel and aneurysm classes in the VesselMNIST3D dataset, thereby showcasing the potential of my method for advancing research and applications in the field.

3.4 SynapseMNIST3D

3.4.1 Dataset and Preprocessing

The *SynapseMNIST3D* dataset is a novel 3D volume dataset designed for the binary classification task of determining whether a synapse is excitatory or inhibitory. This dataset is derived from a 3D image volume of an adult rat's brain, acquired using a multi-beam scanning electron microscope. The original data has a size of $100 \times 100 \times 100 \mu m^3$ and a resolution of $8 \times 8 \times 30 nm^3$. A sub-volume of $(30 \mu m)^3$ from the original data was previously utilized in the *MitoEM* dataset, which includes dense 3D mitochondria instance segmentation labels.

Three neuroscience experts segmented a pyramidal neuron within the entire volume and meticulously proofread all synapses on this neuron, assigning them either excitatory or inhibitory labels. For each labeled synaptic location, a 3D volume of $1024 \times 1024 \times 1024 nm^3$ was cropped and subsequently resized into $28 \times 28 \times 28$ voxels. The dataset is randomly split, with a $7:1:2$ ratio, into training, validation, and test sets, resulting in the following distribution: Training set: 1230 samples, Validation set: 177 samples, and Test set: 352 samples.

The dataset's binary classification task involves two classes, where 0' represents inhibitory synapses and 1' represents excitatory synapses.

3.4.2 Hyperparameters

The particular hyperparameters selected for DAM are detailed in Table 14, and the hyperparameters for LibAUC are provided in Table 15.

Table 14: SynapseMNIST3D Hyperparameters for DAM

Hyperparameter	Value
Number of Epochs	50
Batch Size	32
Learning Rate	0.001
Criterion	CrossEntropyLoss
Optimizer	Adam
Weight Decay	0.0001

Table 15: SynapseMNIST3D Hyperparameters for LibAUC

Hyperparameter	Value
Number of Epochs	50
Batch Size	64
Learning Rate	0.001
Epoch Decay	0.003
Weight Decay	0.0001
Margin	1.0
Loss Function	AUCMLoss
Optimizer	PESG

3.4.3 Experiments and Results

In the experiments conducted, two different approaches were used: DAM and LibAUC. The DAM approach initially showed promising results with a batch size of 64, achieving a test AUC of 0.9841 and ACC of 19.8239% at epoch 2. When the batch size was increased to 128, the performance improved significantly, reaching an AUC of 1.0000 and ACC of 94.0000% at epoch 25. Despite introducing additional data augmentations and applying a class weighting technique to the cross-entropy loss (accounting for 899 positive samples and 331 negative samples), the best model's ACC decreased to 21.7841% at epoch 26, while the AUC remained at 1.0000 (Figure 15).

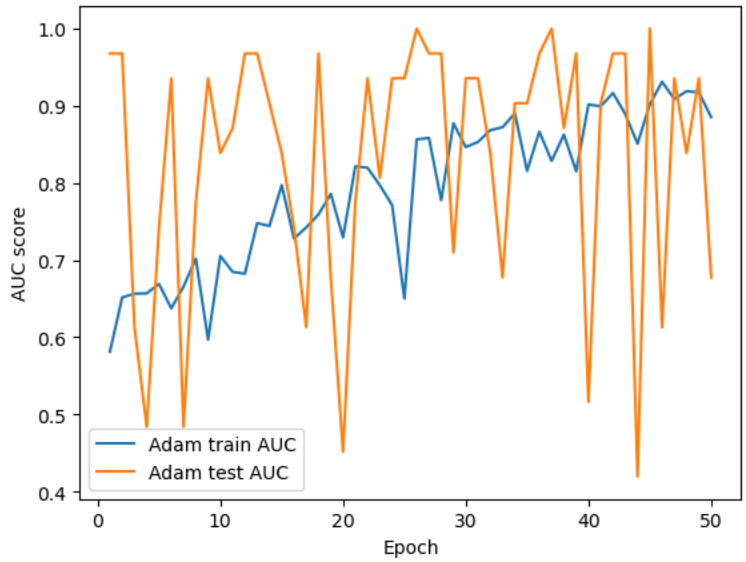

Figure 15: SynapseMNIST3D Train and Test AUC Scores vs Epoch using DAM

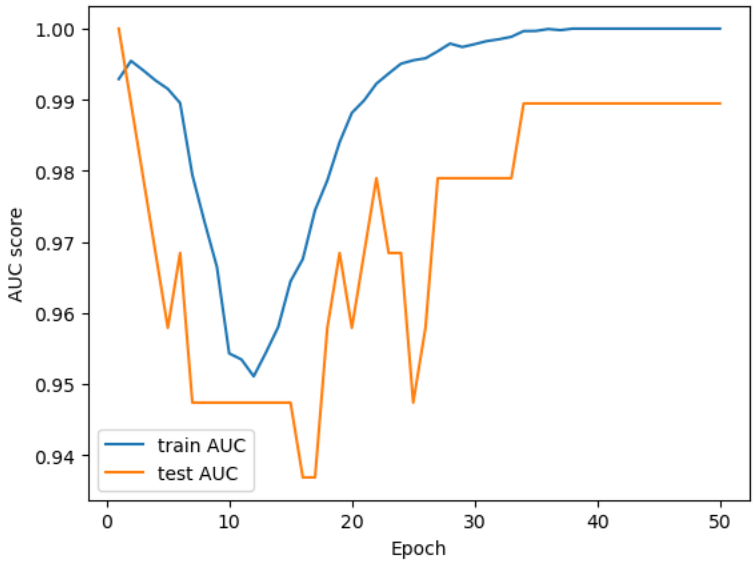

Figure 16: SynapseMNIST3D Train and Test AUC Scores vs Epoch using LibAUC

In the next stage, the pre-trained ResNet 18 3D model from DAM was employed in the LibAUC approach. This change resulted in a considerable performance boost, with a test AUC of 1.0000 and ACC of 94.1705%, both achieved at epoch 1 (Figure 16).

Comparing the results of the implemented approaches to the benchmark performance, with an AUC of 0.820 and an ACC of 0.745, demonstrates that both the DAM and LibAUC approaches outperformed the benchmark. The LibAUC approach, in particular, provided a substantial improvement in both AUC and ACC metrics. This suggests that leveraging a pre-trained model from the DAM approach and incorporating it into the LibAUC approach can lead to enhanced performance in the binary classification task.

4 Conclusion

In conclusion, this study explored the generalization potential of Deep AUC Maximization (DAM) in medical image classification tasks involving small datasets by utilizing the LibAUC library. The research encompassed seven medical image classification tasks from the MedMNIST dataset, with the goal of enhancing DAM performance through various network structures and surpassing the benchmark results outlined in the MedMNIST paper. All datasets underwent pre-training with DAM using ResNet-18 and fine-tuning with LibAUC.

Employing a comprehensive approach, the study pursued multiple avenues for improvement and showcased innovative concepts to strengthen LibAUC's generalization capabilities. The 2-D data analysis results revealed mixed success in exceeding benchmark AUC values, with certain models displaying higher AUC scores yet lower accuracy. Dropout regularization proved crucial in boosting the model's generalization abilities. Particularly, in the multi-class ChestMNIST classification task, the results did not reach the benchmark for AUC or accuracy, indicating a need for additional refinement and optimization.

In 3-D datasets, the ResNet-18 3D model was pre-trained using DAM and fine-tuned with LibAUC. This approach yielded varied success across different datasets, outperforming the benchmark AUC in NoduleMNIST3D, AdrenalMNIST3D, VesselMNIST3D, and SynapseMNIST3D. However, accuracy remained below the benchmark in certain cases, underscoring the necessity for ongoing optimization and refinements. The SynapseMNIST3D dataset displayed the most significant improvement in both AUC and accuracy, highlighting the effectiveness of combining the pre-trained DAM model with the LibAUC approach.

Overall, this study advances the development of more robust and adaptable DAM-based models for medical image classification tasks, especially those with small datasets. The results underscore the potential of the LibAUC library in elevating classification model performance for 2-D and 3-D medical image datasets. Future research should concentrate on refining models, investigating alternative methods and techniques specific to multi-class classification issues, and addressing the challenges arising from imbalanced datasets to enhance performance in both AUC and accuracy metrics.

References

[1] Jiancheng Yang, Rui Shi, DonglaiWei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), Jan 2023.

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[3] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546-6555. 2018.

[4] Yuan, Zhuoning and Qiu, Zi-Hao and Li, Gang and Zhu, Dixian and Guo, Zhishuai and Hu, Quanqi and Wang, Bokun and Qi, Qi and Zhong, Yongjian and Yang, Tianbao. 2022. LibAUC: A Deep Learning Library for X-risk Optimization. `https://libauc.org/`.