3-Class Kidney Fibrosis Classification in Mice

Ross Enriquez, Alan Ly, Jonathan Ma, Aaron Lee

Toronto Metropolitan University
350 Victoria St, Toronto, ON M5B 2K3, Canada
{ross.enriquez, alan.ly, jonathan.ma, aaron.k.lee}@torontomu.ca

***Abstract* — Computer vision is a powerful technology that has many medical applications. Successfully classifying kidney fibrosis would provide a safe, non-invasive procedure to replace current methods. In this paper we train three models using transfer learning to classify photoacoustic and ultrasound images into three categories: severe, mild, and sham (control), achieving an accuracy of 87%. From this we learned that photoacoustic imaging is better suited for kidney fibrosis analysis, compared to ultrasonic imaging.**

***Keywords*— Transfer learning, Medical Image Classification, Kidney Fibrosis, MobileNetV2, VGG-16, ResNet50V2**

I. Introduction

The aim of this paper is to apply computer vision to classify kidney fibrosis within mice. In humans, kidney fibrosis is a condition characterized by scar tissues forming within the kidney, usually occurring in the late stages of various kidney diseases [1]. Fibrosis causes reduced blood flow and impaired kidney function, which are early signs of kidney failure. Therefore, identifying fibrosis is crucial for understanding the progression and severity of kidney disease. Furthermore, the ability to detect and analyze fibrosis through medical imaging alone has potential to provide a noninvasive method for diagnosis and monitoring.

A. Dataset / Data collection

The dataset used for this paper consists of medical image scans of mice kidneys. A surgical procedure called Ischemia Reperfusion Injury (IRI) was used to induce kidney fibrosis within mice, which were then scanned. The fibrosis was induced at 3 different severity levels, depending on the amount of time the kidney was exposed to IRI. This allows for 3 dataset classes: Sham, Mild, and Severe. The Sham group was exposed to the same surgical steps, but without the IRI procedure, which was used as a control group. Then, the kidneys were scanned using both ultrasound (US) and photoacoustic (PA) imaging at 750 nm and 850 nm respectively. The images were saved as .JPG files.

The goal of this paper is to classify mice into these three classifications of kidney fibrosis: sham, mild, and severe, with as high of an accuracy as possible. Later, the results will show whether imaging at 750 nm or 850 nm is more effective, and whether using ultrasound or photoacoustic imaging is more effective at producing accurate classification results.

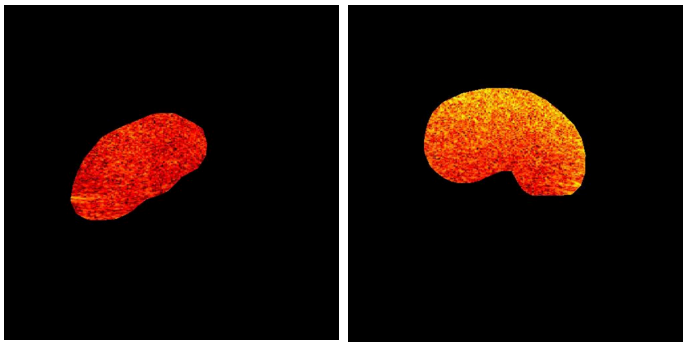

Fig. 1 Sample images from dataset (photoacoustic 850nm severe and mild)

II. Material and Methods

A. Model Selection

Rather than training a model from scratch, transfer learning was used to adapt an existing convolutional neural network to process the dataset. The three models used within this paper are MobileNetV2, VGG-16, and ResNet50V2.

It was important to find a model that either had a history of performing well for medical image classification, or is a well known multipurpose model. MobileNetV2 was chosen because it has a lightweight architecture which makes it faster to train. Furthermore, it has been used to achieve an accuracy of 82% when classifying breast cancer tumors [2]. VGG-16 is a model well known for being generalizable and having good performance on a variety of applications. Finally, ResNet50V2 is a deep model with many layers, which allows it to potentially perform well with sufficient training.

ResNet50V2 has also been documented to successfully classify COVID-19 in chest X-rays, with 88% accuracy [3]. This paper will conduct a comparative study of these three models.

B. Transfer Learning

Two different approaches to transfer learning were used: fine-tuning, and feature extraction. All three models were run using both types, in order to compare the different approaches.

1) *Fine-tuning the model*: In fine tuning, most of the layers were frozen except for the final layers. The final layers, representing the higher order features, were re-trained in order to adapt to processing medical images. The number of layers frozen and the number of layers re-trained is a variable hyperparameter that was experimented with to produce optimal results.

2) *Feature extraction*: Another method of transfer learning that can be used for this study is feature extraction. Using a pre-trained base convolutional model, training data can be passed to generate numerical features which can be extracted for further processing. Since these features are extracted using the representations learned from a pre-trained model, the entire model does not need to be re-trained. Instead, a classifier must be added and trained on the specific classification task.

B. Setup

A Python script was created using Tensorflow to set up and train the models. This was done both online through Google Colab, and locally using conda and cudnn.

C. Strategy

Due to the nature of the given dataset, data leakage was a potential issue that would undermine the results. When training the models, it was vital to ensure that images from one mouse are not used in both training and testing. If this were the case, the testing cases would have contained images that the models had already seen during training. This would have allowed the model to essentially "cheat" and achieve an artificially high validation accuracy. Therefore, the images associated with each mouse had to be clearly separated in order to prevent data leakage. This was done by using the images of one mouse for testing, and the images of the remaining three mice for training.

Below are some strategies which were implemented in order to increase the performance of the model and achieve a higher validation accuracy.

1) *Data Augmentation*: Since the dataset was small, data augmentation was used. This involved introducing random flipping and rotation to the existing images in order to generate new images and expand the dataset.

2) *Hyperparameter tuning*: Libraries such as GridSearchCV and RandomSearchCV were used to identify the optimal values for hyperparameters such as C, kernel, and gamma.

3) *Group Kfold*: Due to the small size of our dataset, cross validation was used in place of a normal training / testing split. The dataset was split into k groups, each of which were used as testing against the rest of the groups. This allows for testing and training on the whole dataset, rather than on only a certain percentage. Cross validation with both 4 and 5 folds was performed. It was especially important to separate mouse data to ensure no data leakage occurred.

4) *Early Stopping*: In order to prevent overfitting, early stopping was implemented. When the validation accuracy stopped increasing and began to deviate from the training accuracy, the training was stopped. This was used to determine the optimal number of training epochs.

III. Results

A. MobileNetV2

For MobileNetV2, multiple training methods were used. Originally, the methods that were used include group 4-fold cross-validation, data augmentation, regularization, dropout, transfer learning with feature extraction, oversampling, and using an SVM classifier with hyperparameter tuning was used. The hyperparameters that were used when tuning the SVM classifier were the C regularization parameter (0.1, 1, 10, 100) and kernel (linear, rbf, poly). Table 1 shows the best hyperparameters for each dataset using GridSearchCV, Table 2 presents the performance metrics for each dataset, identified using GridSearchCV, and Fig. 2 through Fig. 5 will show

the resultant confusion matrices for each fold of each dataset.

TABLE 1
OPTIMAL HYPERPARAMETERS FOR EACH DATASET

Hyperparameter	Datasets			
	PA 750	PA 850	US 750	US 850
C	0.1	0.1	0.1	0.1
kernel	linear	linear	linear	linear

TABLE 2
PERFORMANCE METRICS FOR EACH DATASET TYPE USING AN SVM CLASSIFIER

Type	Fold	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)
750 PA	1	37.00%	0.49	0.37
	2	48.00%	0.50	0.48
	3	34.00%	0.33	0.34
	4	64.00%	0.69	0.64
850 PA	1	47.00%	0.61	0.47
	2	45.00%	0.49	0.45
	3	41.00%	0.53	0.41
	4	26.00%	0.24	0.26
750 US	1	27.00%	0.42	0.27
	2	50.00%	0.49	0.50
	3	30.00%	0.28	0.30
	4	31.00%	0.36	0.31
850 US	1	46.00%	0.61	0.46
	2	35.00%	0.37	0.35
	3	31.00%	0.35	0.31
	4	48.00%	0.31	0.48

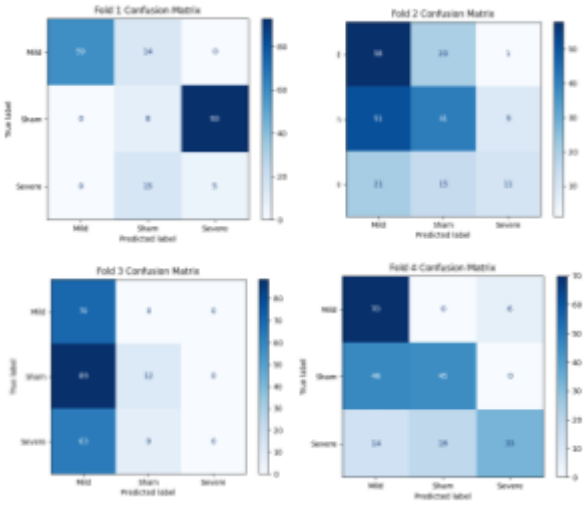

Fig. 2 Original confusion matrices generated from MobileNetV2 with the 750nm Photoacoustic Dataset.

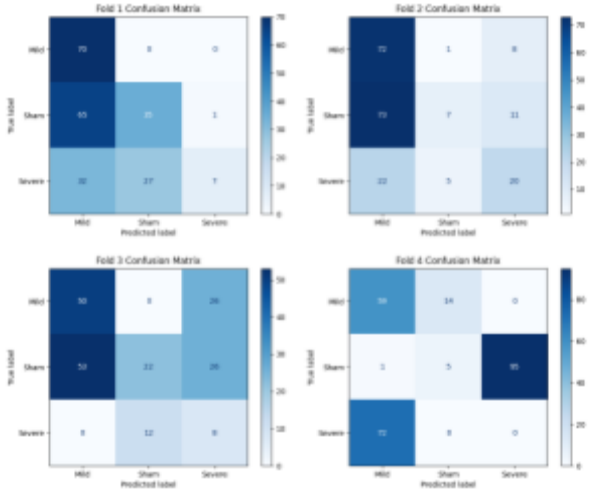

Fig. 3 Original confusion matrices generated from MobileNetV2 with the 850nm Photoacoustic Dataset.

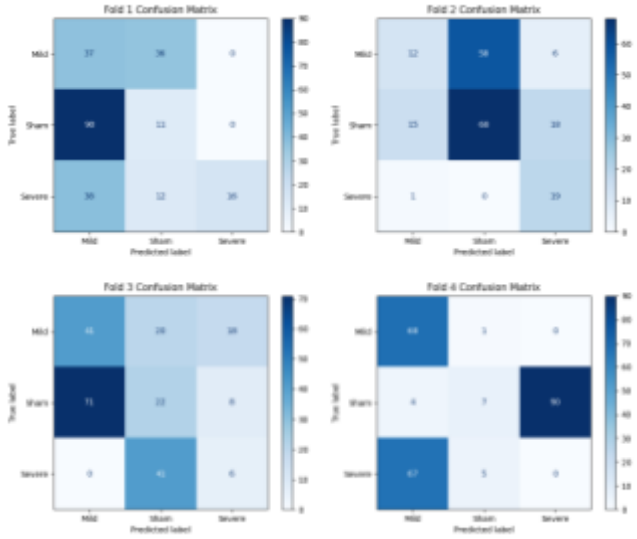

Fig. 4 Original confusion matrices generated from MobileNetV2 with the 750nm Ultrasound Dataset.

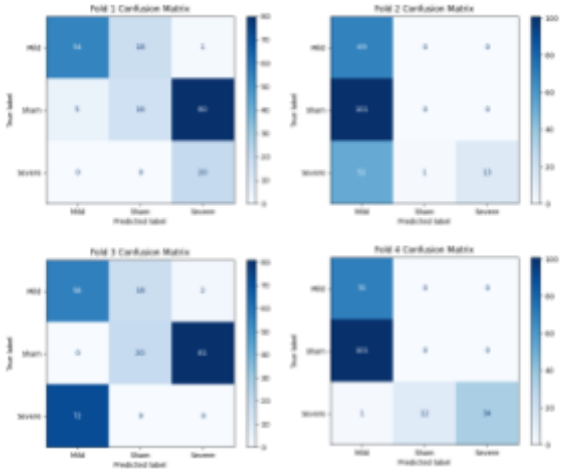

Fig. 5 Original confusion matrices generated from MobileNetV2 with the 850nm Ultrasound Dataset.

However, with the sub par results from the combination of methods mentioned earlier, a slightly better approach was adopted. While there wasn't a significant increase regarding the average validation accuracy, the difference between the original approach and the newer approach was the addition of SMOTE in oversampling, and using group 5-fold cross-validation instead of 4. The best hyperparameters for each dataset was the same as Table 1. Table 3 presents the performance metrics of each dataset, identified using GridSearchCV and Fig. 6 through Fig. 9 will show the resultant confusion matrices for each fold of each dataset.

TABLE 3
PERFORMANCE METRICS FOR EACH DATASET TYPE USING AN SVM CLASSIFIER

Type	Fold	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)
750 PA	1	72.00%	0.74	0.72
	2	40.00%	0.42	0.40
	3	33.00%	0.30	0.33
	4	71.00%	0.73	0.71
	5	56.00%	0.68	0.56
850 PA	1	44.00%	0.57	0.44
	2	66.00%	0.76	0.66
	3	32.00%	0.27	0.32
	4	36.00%	0.31	0.36
	5	33.00%	0.35	0.33
750 US	1	62.00%	0.66	0.62
	2	48.00%	0.62	0.48
	3	31.00%	0.29	0.31
	4	32.00%	0.28	0.32
	5	82.00%	0.90	0.82
850 US	1	28.00%	0.27	0.28
	2	29.00%	0.29	0.29
	3	62.00%	0.72	0.62
	4	47.00%	0.77	0.47
	5	37.00%	0.58	0.37

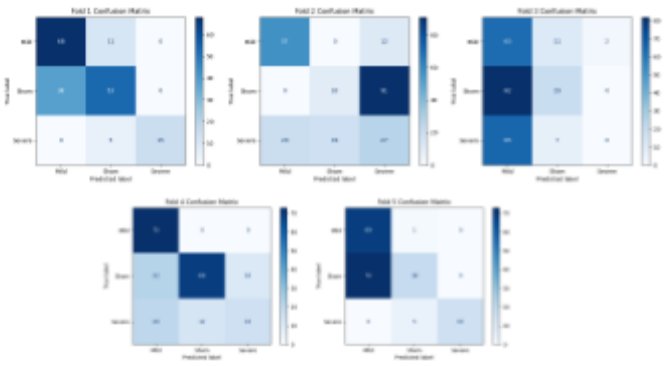

Fig. 6 Updated confusion matrices generated from MobileNetV2 with the 750nm Photoacoustic Dataset.

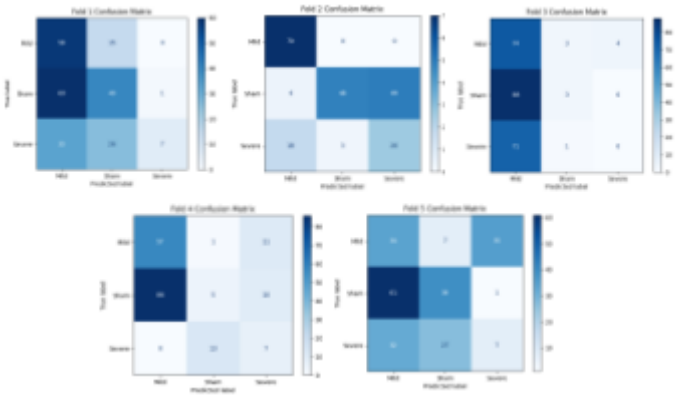

Fig. 7 Updated confusion matrices generated from MobileNetV2 with the 850nm Photoacoustic Dataset.

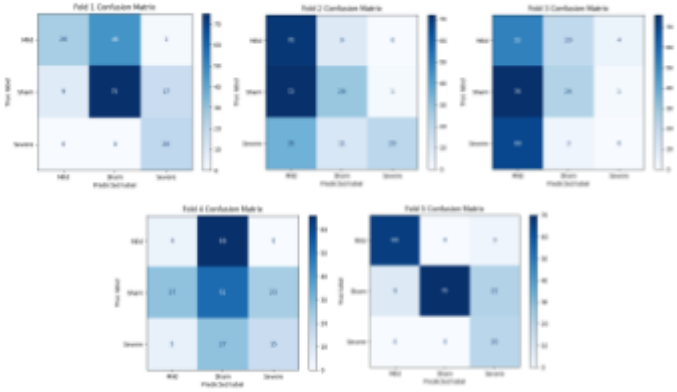

Fig. 8 Updated confusion matrices generated from MobileNetV2 with the 750nm Ultrasound Dataset.

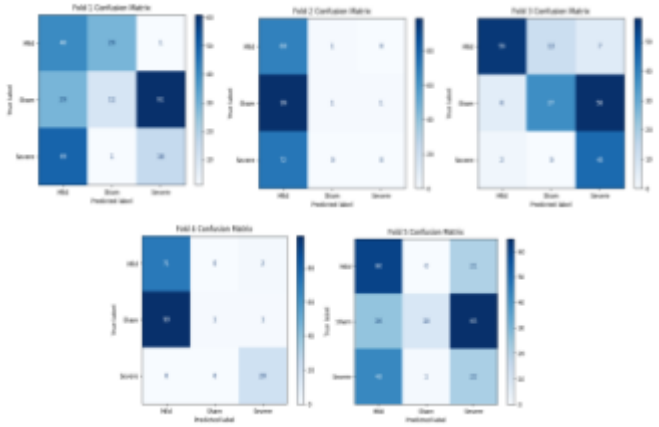

Fig. 9 Updated confusion matrices generated from MobileNetV2 with the 850nm Ultrasound Dataset.

B. VGG-16

For VGG-16, multiple training methods were used including feature extraction, 4-fold cross validation, CatBoost and transfer learning with fine-tuning layers. Originally, the usage of feature extraction, oversampling with SMOTE, and an SVM classifier with hyperparameter tuning was used. The hyperparameters that were tuned for the SVM classifier included the regularization parameter C (with values 0.01, 0.1, 1, 10 and 100), the kernel type (linear, rbf, poly, sigmoid) and the kernel coefficient gamma (scale, auto, 0.01, 0.1, 0.5). Group 4-fold cross-validation was applied to ensure robustness. Table 4 presents the performance metrics for each dataset, identified through GridSearchCV. Fig. 10 through Fig. 13 display the confusion matrices for each fold of each dataset.

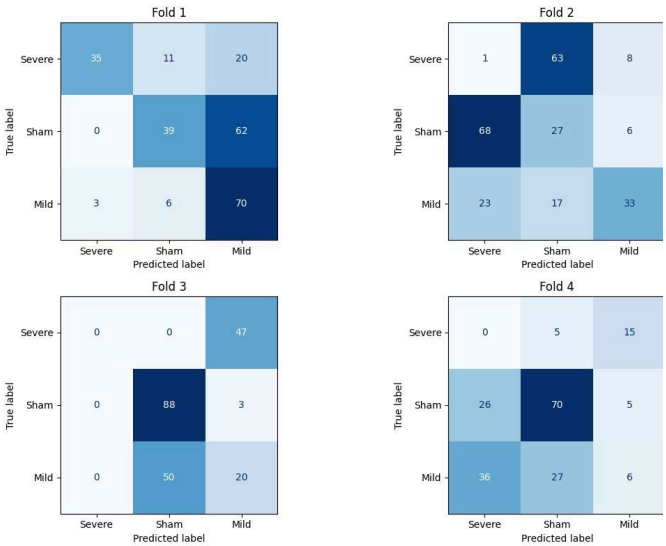

Fig. 10 Confusion matrices generated from VGG-16 using an SVM classifier with the 750nm Photoacoustic Dataset.

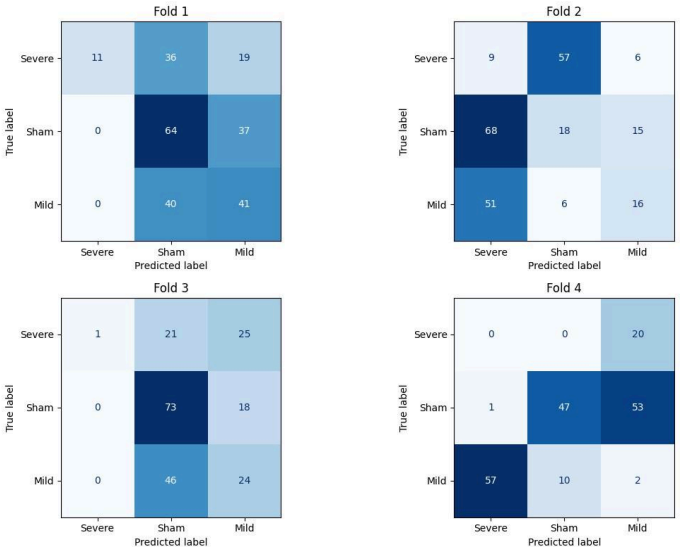

Fig. 11 Confusion matrices generated from VGG-16 using an SVM classifier with the 850nm Photoacoustic Dataset.

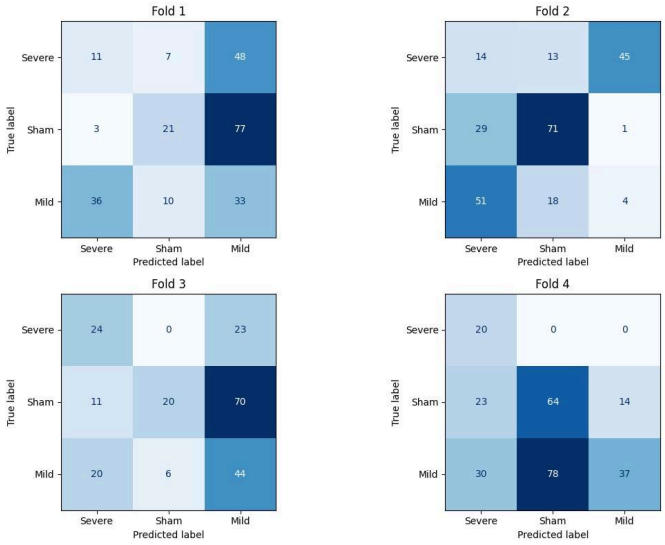

Fig. 12 Confusion matrices generated from VGG-16 using an SVM classifier with the 750nm Ultrasound Dataset.

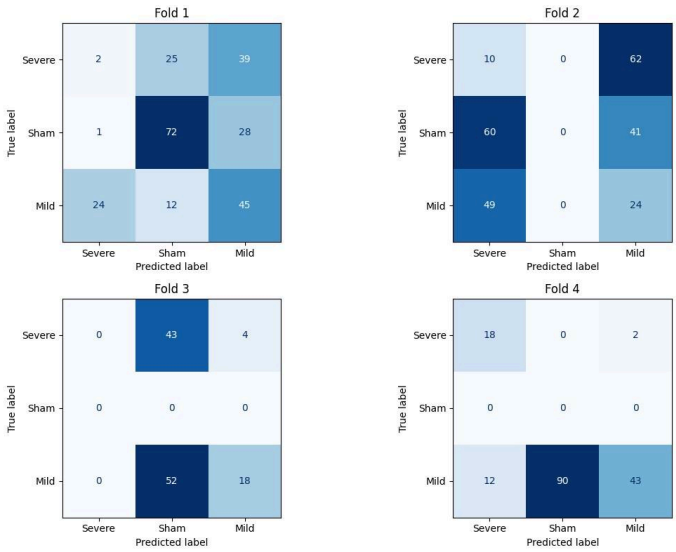

Fig. 13 Confusion matrices generated from VGG-16 using an SVM classifier with the 850nm Ultrasound Dataset.

TABLE 4
PERFORMANCE METRICS FOR EACH DATASET TYPE USING AN SVM CLASSIFIER

Type	Fold	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)
750 PA	1	59.00%	0.68	0.59
	2	25.00%	0.32	0.25
	3	52.00%	0.38	0.52
	4	40.00%	0.45	0.40
850 PA	1	47.00%	0.59	0.47
	2	17.00%	0.24	0.17
	3	47.00%	0.57	0.47
	4	26.00%	0.45	0.26
750 US	1	26.00%	0.35	0.26
	2	36.00%	0.35	0.36
	3	40.00%	0.55	0.40
	4	45.00%	0.59	0.45
850 US	1	48.00%	0.42	0.48
	2	14.00%	0.08	0.14
	3	15.00%	0.49	0.15
	4	37.00%	0.91	0.37

Since the combination of methods mentioned earlier, which included feature extraction and hyperparameter tuning of the SVM classifier produced suboptimal results, a more streamlined approach was adopted. This approach focused on transfer learning with fine-tuning of the last 4 layers, early stopping and the addition of class weights to handle class imbalance. This method yielded the most accurate results. Table 5 shows the performance metrics for different dataset types using transfer learning across class categories. Fig. 14 through Fig. 17 show the resultant confusion matrices for each dataset type.

TABLE 5
PERFORMANCE METRICS FOR DIFFERENT DATASET TYPES USING TRANSFER LEARNING

Type	Class	Accuracy	Precision	Recall	F1
750 PA	Severe	87.00%	0.86	0.95	0.91
	Sham	87.00%	0.96	0.79	0.87
	Mild	87.00%	0.80	0.91	0.85
850 PA	Severe	73.00%	0.90	0.14	0.24
	Sham	73.00%	0.66	0.90	0.76
	Mild	73.00%	0.79	0.99	0.88
750 US	Severe	53.00%	0.46	0.53	0.49
	Sham	53.00%	0.55	0.88	0.68
	Mild	53.00%	0.88	0.09	0.16
850 US	Severe	46.00%	0.17	0.08	0.11
	Sham	46.00%	0.60	0.52	0.56
	Mild	46.00%	0.44	0.70	0.54

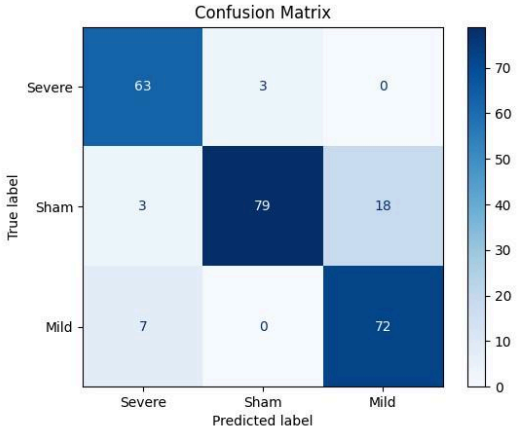

Fig. 14 Confusion matrix generated from VGG-16 using Transfer Learning and Fine-Tuning with the 750nm Photoacoustic Dataset.

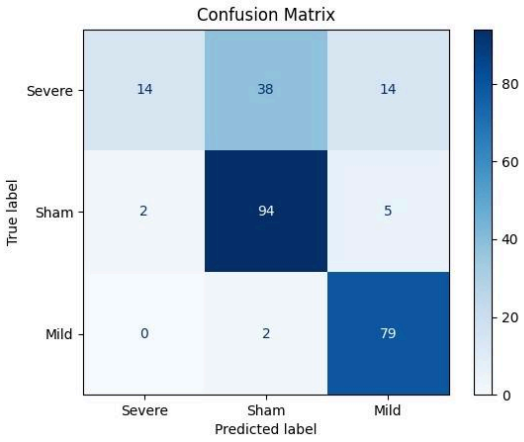

Fig. 15 Confusion matrix generated from VGG-16 using Transfer Learning and Fine-Tuning with the 850nm Photoacoustic Dataset.

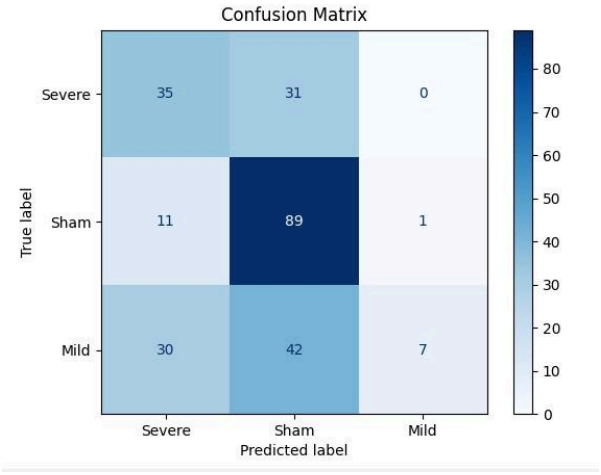

Fig. 16 Confusion matrix generated from VGG-16 using Transfer Learning and FIne-Tuning with the 750nm Ultrasound Dataset.

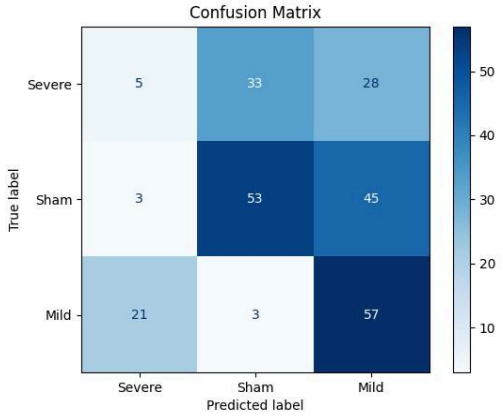

Fig. 17 Confusion matrix generated from VGG-16 using Transfer Learning and Fine-Tuning with the 850nm Ultrasound Dataset.

C. ResNet50V2

For ResNet50V2, transfer learning methods using feature extraction and fine-tuning were used. For feature extraction, an SVM classifier was used with tuned hyperparameters and 4-fold cross-validation. The most common hyperparameters for tuning an SVM classifier are the regularization parameter *C*, *kernel*, and the kernel coefficient *gamma* [6]. Table 6 shows the best hyperparameters for each dataset found through RandomSearchCV. Fig. 18 through Fig. 21 show the resultant confusion matrices for each fold of each dataset.

TABLE 6
OPTIMAL HYPERPARAMETERS FOR EACH DATASET

Hyperparameter	Datasets			
	PA 750	**PA 850**	**US 750**	**US 850**
C	0.066904211664988801	0.074176520348718831	12.746711578215052	12.746711578215052
kernel	linear	linear	sigmoid	sigmoid
gamma	0.012561043700013555	0.012030178871154668	0.005762487216478602	0.005762487216478602

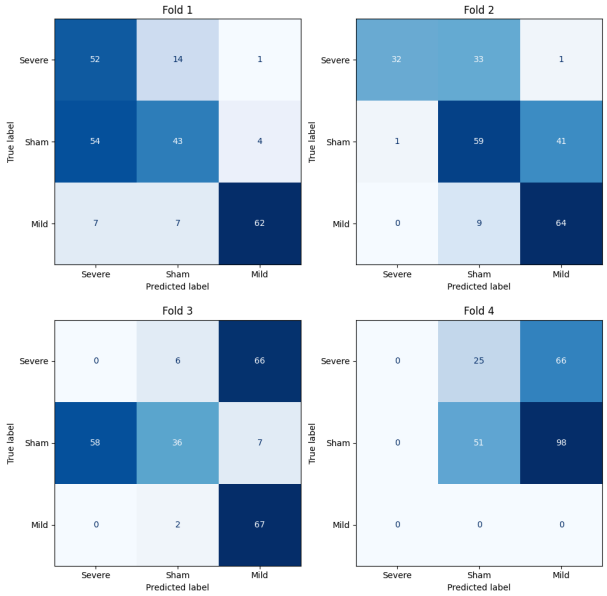

Fig. 18 Confusion matrices generated from ResNet50V2 using feature extraction with the 750nm Photoacoustic Dataset.

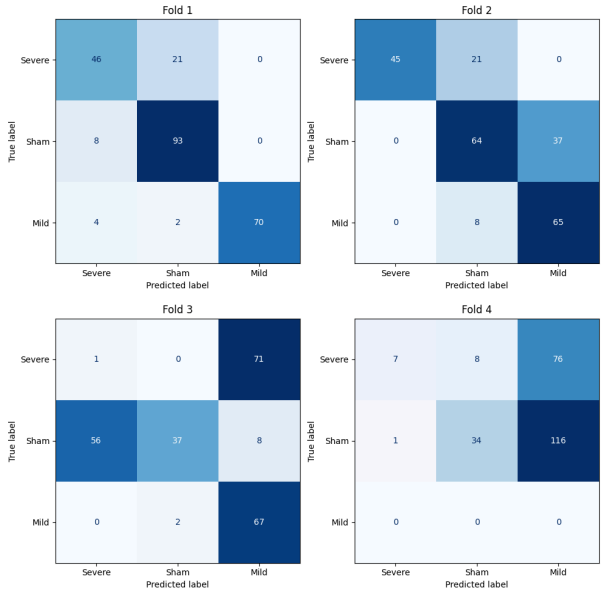

Fig. 19 Confusion matrices generated from ResNet50V2 using feature extraction with the 850 nm Photoacoustic Dataset.

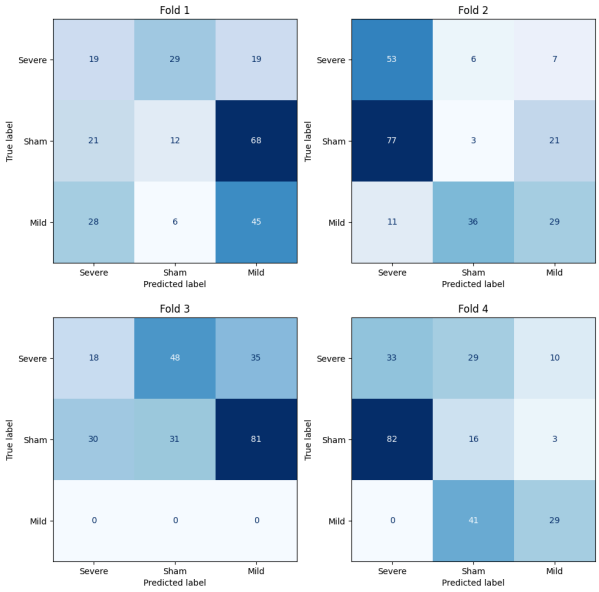

Fig. 20 Confusion matrices generated from ResNet50V2 using feature extraction with the 750 nm Ultrasound Dataset.

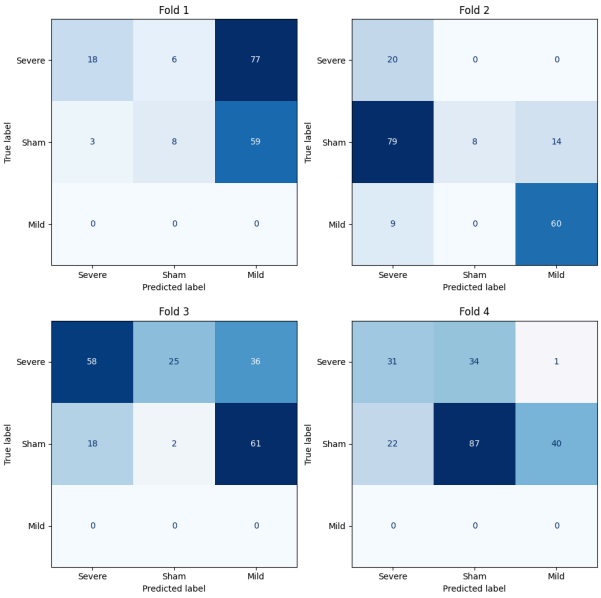

Fig. 21 Confusion matrices generated from ResNet50V2 using feature extraction with the 850 nm Ultrasound Dataset.

With fine-tuning of ResNet50V2, all layers except the top 4 were frozen. Data augmentation was used to compensate for the small dataset, and a Dense layer with ReLU activation with a 0.5 Dropout to another Dense layer with a softmax activation for the classifier. The learning rate was set to 0.0001, and SMOTE class weight balancing was used, along with 4-fold cross-validation for 100 epochs and early stopping with patience of 5. Fig. 22 through Fig 25 show the resultant confusion matrices for each fold of each dataset.

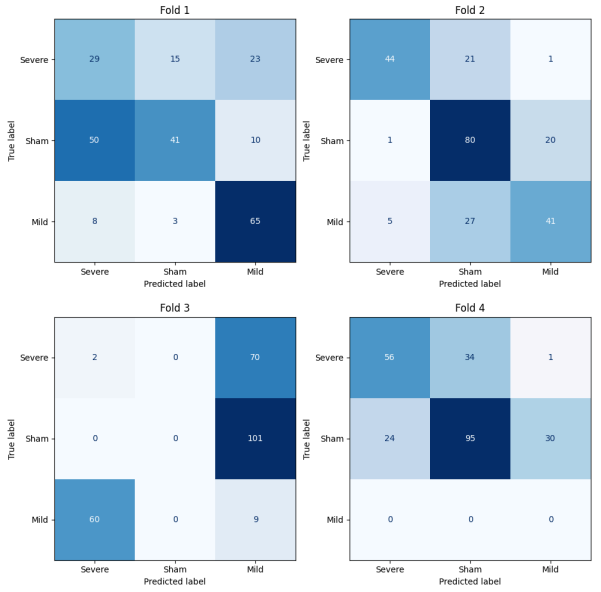

Fig. 22 Confusion matrices generated from ResNet50V2 using fine-tuning with the 750 nm Photoacoustic Dataset.

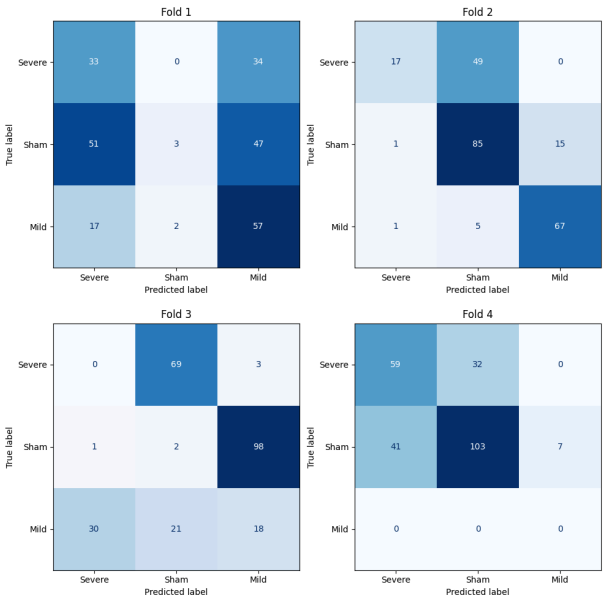

Fig. 23 Confusion matrices generated from ResNet50V2 using fine-tuning with the 850 nm Photoacoustic Dataset.

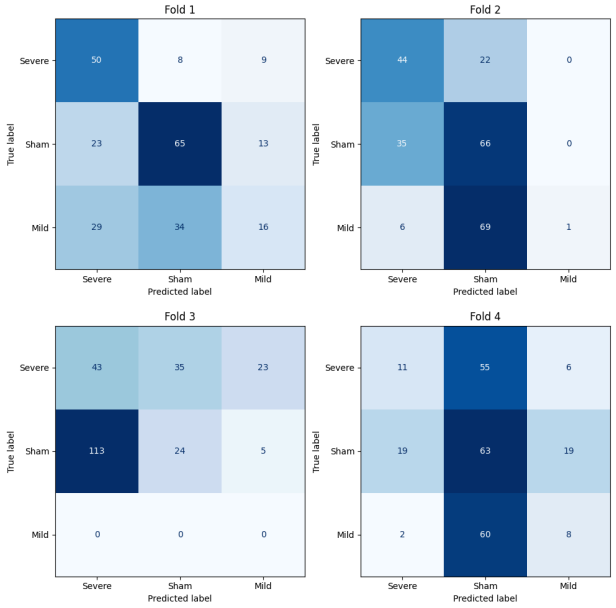

Fig. 24 Confusion matrices generated from ResNet50V2 using fine-tuning with the 750 nm Ultrasound Dataset.

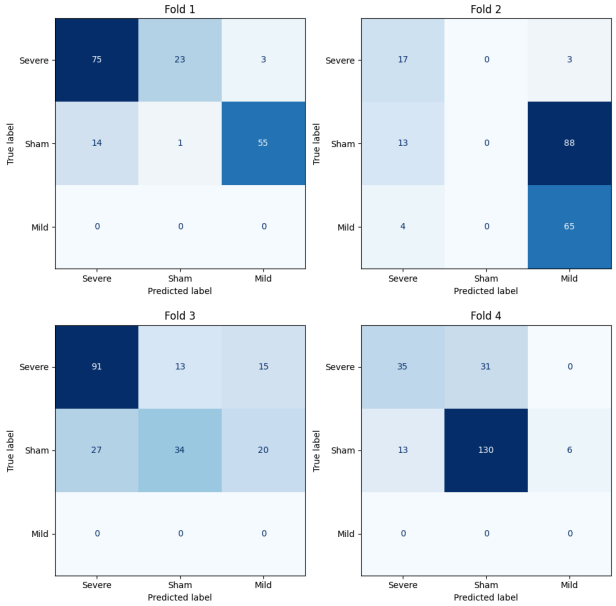

Fig. 25 Confusion matrices generated from ResNet50V2 using fine-tuning with the 850 nm Ultrasound Dataset.

In addition, several performance metrics were measured for each dataset as outlined in Table 7. Evidently, transfer learning by feature extraction performed much better than fine-tuning for ResNet50V2. This could potentially be attributed to the fact that the majority of model evaluation was spent on feature extraction, so more time fine-tuning the model could have improved its performance.

TABLE 7
PERFORMANCE METRICS FOR RESNET50V2 USING FINE-TUNING

Type	Fold	Accuracy (%)	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
750 PA	1	55.0	0.59	0.55	0.55
	2	69.0	0.71	0.69	0.69
	3	5.0	0.02	0.05	0.03
	4	63.0	0.72	0.63	0.67
850 PA	1	38.0	0.47	0.38	0.30
	2	70.0	0.75	0.70	0.67
	3	8.0	0.05	0.08	0.06
	4	67.0	0.70	0.67	0.68
750 US	1	53.0	0.52	0.53	0.50
	2	46.0	0.65	0.44	0.37
	3	28.0	0.35	0.28	0.28
	4	34.0	0.32	0.34	0.30
850 US	1	44.0	0.51	0.44	0.48
	2	43.0	0.20	0.43	0.28
	3	62.0	0.75	0.62	0.67
	4	77.0	0.78	0.77	0.77

IV. DISCUSSION

Table 8 summarizes the best average model accuracies for each dataset. While the accuracy of all models proved to be low on average, the highest accuracy achieved was 87%.

TABLE 8
RESULTING MODEL ACCURACIES FOR EACH DATASET

Transfer Learning Methods	Models	Accuracies per Dataset (%)			
		PA 750	PA 850	US 750	US 850
Feature Extraction	MobileNetV2	54.4	42.2	51.0	40.6
	VGG-16	44.5	34.25	36.75	28.5
	ResNet50V2	48.0	55.0	30.0	37.0
Fine-tuning	MobileNetV2	29.2	34.6	32.2	47.8
	VGG-16	87.0	73.0	53.0	46.0
	ResNet50V2	47.0	41.0	39.0	45.0

A. Model Selection and Transfer Learning Strategy

From this table, it is clear that VGG-16 was the superior model, performing better than the other two models in all datasets. However, the initial conditions of the model settings are vital to the performance, because the first attempts with VGG-16 resulted in poor accuracy. This shows that fine-tuning transfer learning was significantly more effective compared to feature-extraction transfer learning, for this use case.

B. Imaging Modality

All three models consistently performed better using photoacoustic images compared to the ultrasound images. The laser light used by photoacoustic imaging may provide information more correlated to fibrosis than the sound wave information from ultrasound imaging. This is because ultrasound imaging reveals structural information, whereas photoacoustic imaging can reveal information related to blood oxygen levels and hemoglobin concentration [4].

1) *Imaging Wavelength*: With MobileNetV2 and VGG-16, imaging at 750 nm produced higher accuracies compared to imaging at 850 nm. However, ResNet50V2 performed better on the 850 nm. Overall, 750 nm may be better, but it may depend upon the model's architecture and overall strengths and weaknesses.

V. Conclusions and Future Work

After training the three models using transfer learning, an accuracy of 87% was achieved by using VGG-16 to classify photoacoustic imaging at 750 nm.

A. Future Work

In this paper, classification is done for each dataset: PA 750 nm, PA 850 nm, US 750 nm, and US 850 nm. However, a future model could input all four datasets together and perform classification using information from all four dataset types. Ideally, both would be combined together. This method, called dual-modal imaging, is used in applications such as oncology and vascular imaging [5]. Using dual-modal imaging would provide the model with more information and could potentially provide a higher accuracy than using just one imaging modality alone.

Furthermore, the dataset used was limited. Future work can be done to expand the dataset and provide a more comprehensive set of data. Because of the difficulty of performing IRI on mice to collect the data, perhaps GANs or other image generation technology can be applied as a supplementary version of data augmentation.

Acknowledgements

We would like to acknowledge and thank our professor, Omar Falou, for his consistent support, guidance, and feedback, and for keeping the group on track throughout the course term.

References

[1] Panizo S, Martínez-Arias L, Alonso-Montes C, Cannata P, Martín-Carro B, Fernández-Martín JL, Naves-Díaz M, Carrillo-López N, Cannata-Andía JB. Fibrosis in Chronic Kidney Disease: Pathogenesis and Consequences. Int J Mol Sci. 2021 Jan 2;22(1):408. doi: 10.3390/ijms22010408. PMID: 33401711; PMCID: PMC7795409.

[2] Surya, A., Shah, A., Kabore, J., & Sasikumar, S. (2023). Enhanced Breast Cancer Tumor Classification using MobileNetV2: A Detailed Exploration on Image Intensity, Error Mitigation, and Streamlit-driven Real-time Deployment. arXiv preprint arXiv:2312.03020.

[3] Velu, S. (2023). An efficient, lightweight MobileNetV2-based fine-tuned model for COVID-19 detection using chest X-ray images. Mathematical Biosciences and Engineering, 20(5), 8400-8427.

[4] Beard P. (2011). Biomedical photoacoustic imaging. Interface focus, 1(4), 602–631. https://doi.org/10.1098/rsfs.2011.0028.

[5] N. Nyayapathi, E. Zheng, Q. Zhou, M. Doyley, and J. Xia, "Dual-modal photoacoustic and ultrasound imaging: from preclinical to clinical applications," Front. Photon., vol. 5, 2024, doi: 10.3389/fphot.2024.1359784.

[6] "RBF SVM Parameters," scikit-learn, https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html (accessed Nov. 30, 2024).