DEEP LEARNING CLASSIFICATION OF ABLATED ATRIAL TISSUE IN MULTI-EXCITATION HYPERSPECTRAL IMAGES

Arpi Hunanyan[1,2], Nazeli Ter-Petrosyan[1*], Fernando Villarruel[3], Tigran Soghomonyan[3], Narine Sarvazyan[1,3,4], Aram Butavyan[1], and Varduhi Yeghiazaryan[1]*

[1]Akian College of Science and Engineering, American University of Armenia, Yerevan, Armenia
[2]Institut Polytechnique de Paris, Palaiseau, France
[3]L. A. Orbeli Institute of Physiology NAS RA, Yerevan, Armenia
[4]George Washington University, Washington, DC, United States

ABSTRACT

Multi-excitation hyperspectral imaging (ME-HSI) is an emerging modality that captures rich spectral information for each spatial pixel over a range of excitation wavelengths. To compare ME-HSI to traditional hyperspectral imaging (HSI), we employ five deep learning approaches for pixel-level classification of radiofrequency-ablated bovine left atrial tissue samples. The dataset comprises eight ME-HSI samples, with four excitation wavelengths. We consider three data setups: (1) individual HSI cubes for each excitation wavelength, (2) concatenation of all HSI cubes, and (3) majority voting across predictions for individual HSI cubes. For training, we use randomly selected circular regions, gradually increasing the number of regions to observe model performance dynamics. The two ME-HSI setups that use data from all four excitations show a small overall improvement in performance over the individual HSI cubes. The five-click setup (2% training data) generates the most stable results in all models. GiGCN outperforms other models in overall accuracy.

Index Terms— Multi-excitation hyperspectral imaging, tissue classification, ablated tissue, deep learning, GNN

1. INTRODUCTION

Atrial fibrillation (AF) is the leading cardiac arrhythmia among adults, closely associated with major cardiovascular complications [1]. An effective intervention for treating AF is radiofrequency (RF) ablation (RFA), a minimally invasive procedure in which the heat generated by high-frequency electrical currents is used to eliminate the source of electrical activity in the atrial tissue. However, a major challenge lies in accurately identifying and confirming the extent of tissue damage during the procedure. Postoperative AF recurrence is often attributed to viability gaps—intact myocardial regions between ablated zones that continue to conduct electrical impulses [2]. Given these limitations, there is an increasing demand for real-time, non-invasive imaging modalities capable of visualizing ablation effects intraoperatively with high specificity and spatial resolution.

Hyperspectral imaging (HSI) captures rich spectral information for each spatial pixel across a wide range of wavelengths. Unlike RGB imaging, which captures only three channels per pixel, HSI produces a 3D data structure called a hyperspectral cube, characterized by two spatial and one spectral (λ_{em}) dimensions. The multiple applications of HSI in the biomedical field are increasingly combined with machine learning approaches [3, 4, 5]. The potential of HSI to reveal the extent of tissue damage during surgical RF ablation procedures in atrial heart tissue was explored in [6]. Following this work, [7] presented a comparative analysis of 45 deep learning (DL) classification methods for identifying ablated regions in atrial tissue HSI. The study revealed that many architectures known to perform well on common HSI datasets yielded subpar results on this type of data, likely due to the substantially smaller size of the training set and the higher complexity of the target. Building on traditional hyperspectral methods, [8] introduced a multi-excitation HSI (ME-HSI) approach for enhanced tissue differentiation, in which multiple 3D HSI cubes are captured at different excitation wavelengths (λ_{exc}) and then combined into a single 4D dataset, resulting in improved target identification. [9, 10] employed ME-HSI for nerve and ligament differentiation using convolutional neural networks (CNNs) with up to 99% accuracy.

In this study, we use the ME-HSI technique to create a dataset of RF-ablated bovine atrial tissue. We then apply a selection of well-performing and architecturally diverse models from the study in [7] for the task of ablated atrial tissue classification. We cover six different experimental setups: four individual 3D HSI cubes captured at different excitation wavelengths, a combined 3D cube formed by concatenating all four cubes along the spectral dimension, and a majority voting approach across the four excitations. Additionally, we examine

Thanks to the Afeyan Family Foundation, RA HESC (24PostDoc/2-2G002), and EU (ERA-CHAIR NAS-SAR project) for funding.

* Equal contribution.

how increasing the training data and the user effort in providing click-based annotations affects model accuracy. Complementary experiments and discussion are reported in [11].

2. DATASET

Sample collection: Bovine hearts were obtained from a local butcher in Yerevan surroundings. Left atria (LA) tissue pieces were dissected, and immediately transported on ice to the laboratory. Eight LA pieces from three different cows were used.

RFA execution and HSI acquisition: RFA lesions were created on each LA piece using an *EPT-1000 XP* machine with a 4 mm *Blazer* catheter (*Boston Scientific*) set at 8 W power for 15 s. The catheter tip was placed perpendicular to the tissue during ablation. To account for tissue heterogeneity, four to eight lesions were made on each LA tissue piece, marking the center of each lesion with a needle. Immediately after RFA—and up to six hours after heart extraction—the samples were imaged using an HSI setup consisting of: (i) 300 W xenon lamp outfitted with four filters centered at 360, 370, 380 and 390 nm (all with 10 nm FWHM; *Thorlabs*) and (ii) a Nuance FX (*Perkin Elmer*) imaging system. Four cubes were obtained, with the above-mentioned λ_{exc} and 31 emission wavelengths (λ_{em} = 420–720 nm with 10 nm step), while spatial dimensions varied depending on tissue size.

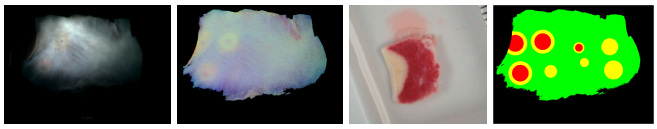

Fig. 1: Imaging pipeline for LA tissue, left-to-right: 370 nm excitation cube before preprocessing, after preprocessing, TTC-stained lesion slide, and ground truth mask derived from TTC staining (red, yellow, green for ablated, unsure, unablated pixels, respectively). Eight ablated regions were expected, but only four were reliably revealed by TTC staining.

Histological analysis: After completion of HSI experiments, LA tissue slides were extracted from each RFA lesion center and stained with 1.0% TTC solution, allowing measurement of the lateral size of the lesions—used to generate ground truth masks; see Fig. 1 for a sample.

Data preprocessing and ground truth mask generation: The acquired data were imported into MATLAB and corrected as follows: (i) exposure time and lamp intensity corrections for each λ_{exc} cube. (ii) combination of corrected cubes into a multi-excitation dataset and pixel-wise normalization to the maximum value.

The center coordinates of each lesion, together with the lateral size (from TTC) were used to generate a mask of lesions—assuming lesions to be circular. As tissue slicing breaks muscle fibers, leading to overstimated lesion sizes, the outer 30% of the lesion area was labeled as 'unsure' tissue and excluded from DL training/test sets and the inner 70% was labeled as 'ablated'. When the ablation center was unclear, an approximated center was identified, labeling the entire circle as 'unsure'. The background was identified by filtering pixels with a significant signal at the maximum emission wavelength, and was excluded from the analysis.

3. METHODS

To investigate the usability of the ME-HSI dataset for ablated region classification, we selected a diverse set of DL architectures representing various learning paradigms [7]. These models include graph neural networks (GNNs), transformer, CNNs, hybrid architectures, and multilayer perceptron (MLP), ensuring comprehensive analysis across different context-learning strategies.

GiGCN [12] is a GNN designed specifically for HSI tasks. It uses internal and external graphs to hierarchically extract features from within and around superpixel regions. The model captures spatial context through graph-based learning. Sellars et al. [13] proposed another GNN tailored for HSI analysis. It extracts both spectral and spatial features from superpixels and constructs weighted graph representations for classification. Like GiGCN, it emphasizes spatial context learning. HybridSN [14] is a CNN-based model developed for HSI classification. It combines 3D convolutional layers for spectral–spatial feature extraction with 2D CNN layers for further spatial processing. LeViT [15] is a vision transformer that merges CNN and transformer designs. Originally introduced for RGB images, it replaces standard linear projections with convolutional embeddings. In this study, it is trained in the spatial context. MLP [16] refers to a basic four-layer fully connected neural network. It serves as a baseline model with no explicit spatial or spectral feature extraction mechanisms.

SS-ConvNeXt [17], a hybrid architecture combining convolutional and transformer-based components, was considered for classification but excluded due to high computational cost; future work will include it for comparative evaluation.

We used a set of three statistical evaluation metrics. The primary metric employed was overall accuracy (OA), which measures the proportion of correctly classified instances out of the total predictions. Additionally, we applied average accuracy (AA), which calculates the mean accuracy for each class, providing a balanced evaluation across classes. To further analyze the agreement between predicted and actual classifications, we used the Kappa statistic, or Cohen's Kappa (κ), a measure of inter-rater agreement for categorical data. This metric ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement at the level of chance. It is given by the formula: $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the observed agreement, and p_e is the expected agreement between the ground truth mask and the classifier output.

We evaluated HSI classification using a training pipeline based on simulated user clicks. The corresponding training

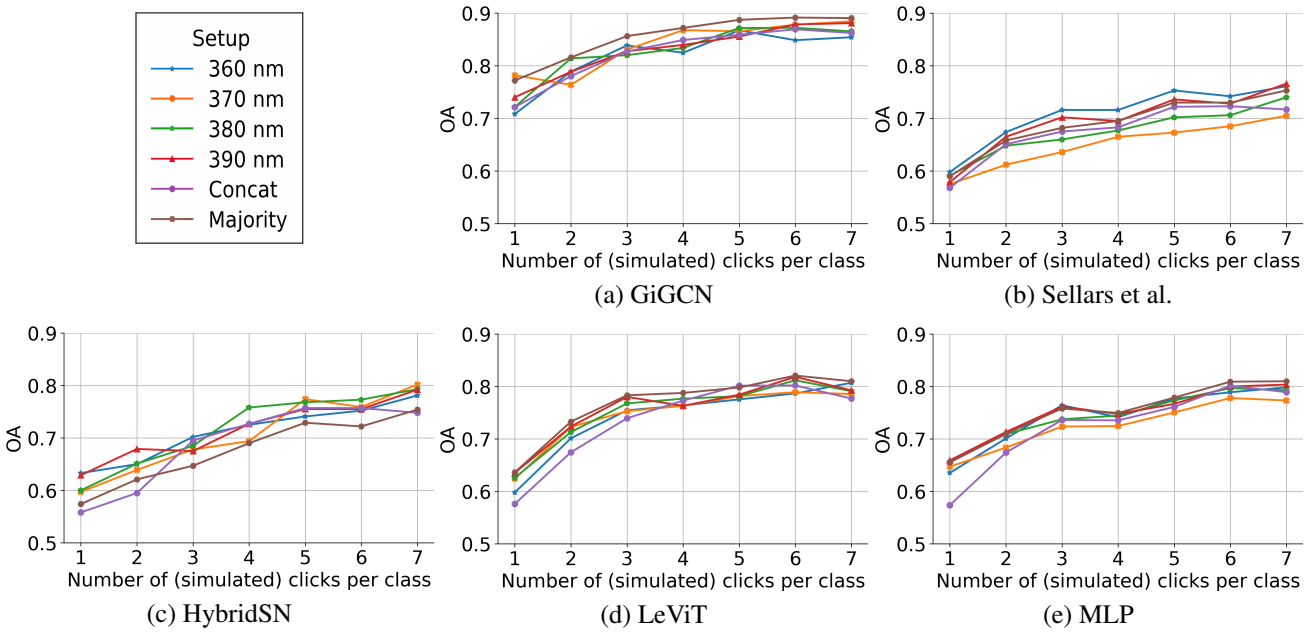

Fig. 2: Average OA scores of the five DL methods ((a)–(e)) for the six data setups (excitation of 360 nm, 370 nm, 380 nm, 390 nm, concatenation, and majority voting) and the numbers of simulated clicks per class ranging from 1 to 7.

masks were generated by placing circular regions (each covering 0.2% of the total image area) randomly within the annotated regions for each class (ablated/unablated). Per class, 1–7 clicks were generated, simulating different interaction levels. The models were trained using five randomly selected training splits per click level to ensure stability and repeatability.

Due to limited ablated regions in few cases, it was not possible to generate masks for higher click levels (6–7 clicks), so those cases were excluded (one sample for both 6 and 7 clicks, and another for 7 clicks).

Overall, we defined six experimental data setups. We compared results across individual excitation channels (four in total), their concatenation, and majority voting. Majority voting was applied across channels; in the case of tied votes, class assignment was resolved by random selection. The experiments were conducted on an Ubuntu 24.04 system equipped with an Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU.

4. RESULTS

Fig. 2 presents the average OA scores across the five DL methods for the six different data setups, with the number of simulated clicks per class ranging from 1 to 7. Across all models and setups, a clear upward trend is observed with OA improving as more clicks, i.e. training data, are provided. Notably, increasing from 1 to 5 clicks can boost accuracy by up to 30% in certain setups. This upward trend is generally consistent up to the 5-click setup, after which the results become more varied: in some cases, the accuracy score reaches a plateau, while in others it continues to rise or even decrease. This variability can be partly explained by the exclusion of a couple of samples from the 6- and 7-click setups; this could potentially affect the representativeness of the results. Given the strong and stable performance at 5 clicks, we fix this as the standard setup for all following experiments.

Fig. 2 also supports a comparative analysis of the six data setups. For the GiGCN, LeViT, and MLP models, the majority voting setup consistently outperforms the other setups across most click numbers. In contrast, the same setup results in subpar or even the worst performance for Sellars et al. and HybridSN. This inconsistency can be explained by the poor individual performance of these models in single-excitation setups, which, when aggregated via majority voting, further reduces OA. With only 1 click, the concatenated setup often performs the worst across most models. However, as the number of clicks increases, its performance improves significantly. For instance, in the case of LeViT, the concatenated setup transitions from the lowest performer at 1 click to the best performer at 5 clicks. This can be attributed to the fact that the concatenated setup includes four times more spectral information than individual excitation setups, requiring more training data (i.e. clicks) to achieve optimal performance.

Table 1 presents the OA, AA and κ scores across all five models and six data setups, with the training setup fixed at 5 clicks. Among the models, GiGCN achieves the highest OA scores in most setups, while LeViT obtains the best AA scores. This difference can be explained by the models' distinct architectures. GiGCN, a GNN, uses superpixels for clas-

Table 1: Mean OA, AA, and κ scores, $\pm$standard deviation, for five DL methods, six data setups, and training setup of 5 simulated clicks per class. Per column, top three scores are in boldface, underlined, and italisized, respectively. In the last row, per measure, the top score is in boldface.

	360 nm			370 nm			380 nm			390 nm			Concat			Majority		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
GiGCN	**.867** $\pm$.083	.743 $\pm$.138	**.336** $\pm$.243	**.866** $\pm$.101	.707 $\pm$.164	.310 $\pm$.308	**.871** $\pm$.082	.704 $\pm$.148	*.312* $\pm$.271	**.856** $\pm$.095	.717 $\pm$.135	*.307* $\pm$.254	**.859** $\pm$.087	.794 $\pm$.109	**.393** $\pm$.241	**.887** $\pm$.077	.726 $\pm$.123	**.375** $\pm$.252
Sellars et al.	.753 $\pm$.096	.809 $\pm$.074	.277 $\pm$.188	.673 $\pm$.128	.750 $\pm$.076	.196 $\pm$.154	.702 $\pm$.113	.767 $\pm$.081	.220 $\pm$.159	.736 $\pm$.107	.788 $\pm$.077	.262 $\pm$.182	.722 $\pm$.103	.786 $\pm$.090	.244 $\pm$.190	.730 $\pm$.109	.792 $\pm$.077	.255 $\pm$.172
HybridSN	.741 $\pm$.129	*.812* $\pm$.068	.280 $\pm$.164	*.774* $\pm$.096	.824 $\pm$.057	*.304* $\pm$.163	.768 $\pm$.120	.824 $\pm$.074	.311 $\pm$.176	.755 $\pm$.113	.827 $\pm$.073	.298 $\pm$.184	.757 $\pm$.110	*.822* $\pm$.062	.286 $\pm$.148	.729 $\pm$.104	*.817* $\pm$.066	.258 $\pm$.134
LeViT	*.775* $\pm$.105	**.830** $\pm$.066	*.310* $\pm$.172	.782 $\pm$.135	**.840** $\pm$.077	**.354** $\pm$.211	.782 $\pm$.127	**.830** $\pm$.086	**.335** $\pm$.211	.784 $\pm$.115	**.840** $\pm$.070	**.341** $\pm$0.205	.802 $\pm$.119	**.848** $\pm$.070	.362 $\pm$.192	.798 $\pm$.115	**.850** $\pm$.064	.360 $\pm$.202
MLP	.777 $\pm$.117	.821 $\pm$.065	.331 $\pm$.208	.750 $\pm$.117	*.805* $\pm$.078	.285 $\pm$.187	*.773* $\pm$.115	*.820* $\pm$.076	.318 $\pm$.194	*.767* $\pm$.124	*.823* $\pm$.074	.320 $\pm$.199	*.761* $\pm$.129	.832 $\pm$.071	*.309* $\pm$.176	*.779* $\pm$.117	.831 $\pm$ .072	*.335* $\pm$.205
Average	.783	.805	.307	.769	.785	.290	.771	.789	.299	.780	.799	.306	.780	**.816**	**.319**	**.785**	.803	.317

sification. These superpixels may include both ablated and unablated regions but are labeled as a single class, which can lead to incorrect labels within a superpixel. Therefore, while the model accurately identifies the ablated areas, it struggles with delineating precise boundaries. The pixelated appearance resulting from this approach is evident in Fig. 3, especially for GiGCN and Sellars et al., the two GNN-based methods. In contrast to the GNN models, LeViT produces smoother region classifications, as it utilizes spatial information in addition to spectral features. This leads to visually smoother results, which can also be observed in HybridSN, another model that integrates both spectral and spatial data. LeViT's high AA scores can be attributed to its consistent performance across both classes, but this score may also be partially driven by its tendency to overlabel, producing large uniform regions that cover entire ablated areas. This behavior is visible in Fig. 3 for the 360 nm and concatenated setups, where LeViT completely captures the ablated class while performing worse on the unablated class, resulting in a high AA score even with imbalanced performance. Despite its simplicity, the third-best performer overall is MLP. It relies solely on spectral information and classifies each pixel independently, leading to a speckled appearance in the classification maps, visible in Fig. 3, especially along the boundaries of ablated and unablated regions. These mixed regions suggest classification uncertainty, where the model struggles to confidently assign a class. A similar pattern of ambiguity can also be observed in the majority voting setup, indicating that boundary regions remain challenging for consistent classification.

Fig. 2 illustrates that the majority voting setup for GiGCN consistently outperforms other configurations, achieving performance levels closely paralleling those of the concatenated setup. The average scores presented in the last row of Table 1 clearly indicate that, overall, the concatenation and majority voting setups outperform the other configurations. Across the four single-excitation setups, there is no clear dominancy. The only exception is the method by Sellars et al., where the 360 nm setup demonstrates better performance, while for other setups, no consistent leader emerges. Further experiments in [11] reveal the impact of other ME-HSI setups.

5. CONCLUSION

We investigated the use of ME-HSI for the DL classification of RF-ablated tissue in samples of bovine left atria. We compared six data setups: individual HSI cubes from four excitation wavelengths, a concatenation of all four cubes along the emission axis, and majority voting among the four cubes. The DL classification algorithms spanned multiple architecture types. Among these, the GNN-based GiGCN and the transformer LeViT showed most promising results in the different setups. The two setups using data from all four excitations showed a small overall improvement of performance over the individual HSI cubes, highlighting the potential of ME-HSI for the task. Future experiments should help further identify the best utilization strategies for ME-HSI.

6. REFERENCES

[1] Ayodele Odutayo et al., "Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis," *BMJ*, vol. 354, 2016.

[2] Hakan Oral et al., "Clinical significance of early recurrences of atrial fibrillation after pulmonary vein isolation," *JACC*, vol. 40, no. 1, pp. 100–104, 2002.

[3] Baowei Fei, "Hyperspectral imaging in medical applications," in *DHS&T*, vol. 32, pp. 523–565. Elsevier, 2019.

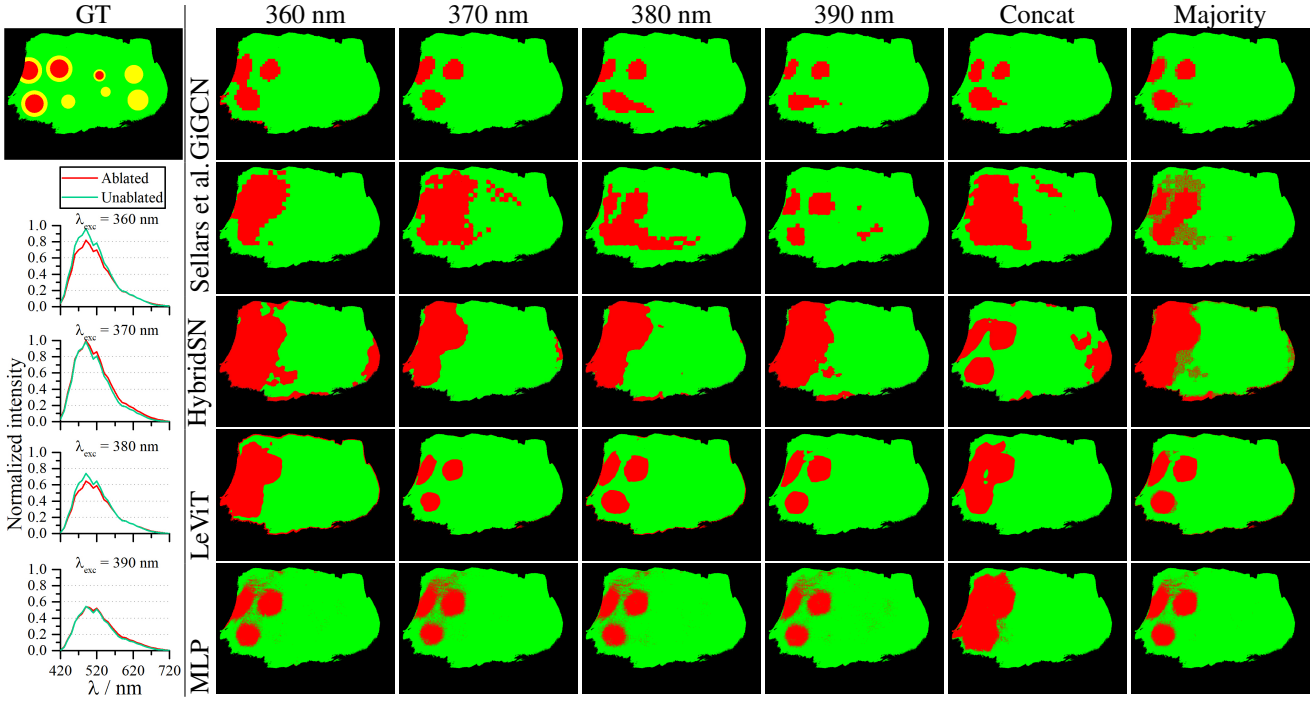

Fig. 3: Predictions (green for unablated, red for ablated) of the five DL methods for the six data setups and the fixed 5-clicks-per-class training. The leftmost column shows the ground truth and average spectra for the two classes. Notably, none of the approaches detected the 'unsure' ablated regions on the right of the image, illustrating sensitivity to training data selection.

[4] Rong Cui et al., "Deep learning in medical hyperspectral images: A review," *Sensors*, vol. 22, no. 24, pp. 9790, 2022.

[5] Uzair Khan et al., "Trends in deep learning for medical hyperspectral image analysis," *IEEE Access*, vol. 9, pp. 79534–79548, 2021.

[6] Luther M Swift et al., "Hyperspectral imaging for label-free in vivo identification of myocardial scars and sites of radiofrequency ablation lesions," *Heart Rhythm*, vol. 15, no. 4, pp. 564–575, 2018.

[7] Yeva Gabrielyan et al., "Comparative analysis of deep learning methods for classification of ablated regions in hyperspectral images of atrial tissue," *IEEE Access*, vol. 13, pp. 35029–35047, 2025.

[8] Narek Chilingaryan et al., "4D hyperspectral imaging for intraoperative tissue classification," in *MI 2025: CBI*. SPIE, 2025, vol. 13410, pp. 255–260.

[9] Naira Matosyan et al., "Combining 4D hyperspectral imaging with CNN for nerve and ligament differentiation," in *ISBI*. IEEE, 2025, pp. 1–5.

[10] Naira Matosyan et al., "Spectral pixels as images: CNN-based pixel classification of 4D hyperspectral data for nerve and ligament differentiation," in *MI 2025: IP*. SPIE, 2025, vol. 13406, pp. 565–573.

[11] Nazeli Ter-Petrosyan et al., "Comparative analysis of multi-excitation hyperspectral image configurations for ablated atrial tissue classification," in *WHISPERS*. IEEE, 2025.

[12] Sen Jia et al., "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE TNNLS*, vol. 35, no. 1, pp. 1157–1171, 2022.

[13] Philip Sellars et al., "Superpixel contracted graph-based learning for hyperspectral image classification," *IEEE TGRS*, vol. 58, no. 6, pp. 4180–4193, 2020.

[14] Swalpa Kumar Roy et al., "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE GRSL*, vol. 17, no. 2, pp. 277–281, 2019.

[15] Benjamin Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *ICCV*, 2021, pp. 12259–12269.

[16] Xiaofei Yang et al., "Extension of deep learning toolbox based on pytorch for hyperspectral data classification," `https://github.com/xiachangxue/DeepHyperX`, Accessed: 18 October 2024.

[17] Yimin Zhu et al., "Spatial–spectral ConvNeXt for hyperspectral image classification," *IEEE JSTARS*, vol. 16, pp. 5453–5463, 2023.