

Pathology image-based lung cancer subtyping using deep-learning features and cell-density maps

Mustafa I. Jaber¹, Christopher W. Szeto², Bing Song³, Liudmila Beziaeva⁴, Stephen C. Benz², Patrick Soon-Shiong³, and Shahrooz Rabizadeh^{1,3}

¹NantOmics LLC., Culver City, CA; ²ImmunityBio, Santa Cruz, CA; ³ImmunityBio, Culver City, CA; ⁴NantHealth, Culver City, CA

Abstract

In this paper, we propose a patch-based system to classify non-small cell lung cancer (NSCLC) diagnostic whole slide images (WSIs) into two major histopathological subtypes: adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). Classifying patients accurately is important for prognosis and therapy decisions. The proposed system was trained and tested on 876 subtyped NSCLC gigapixel-resolution diagnostic WSIs from 805 patients – 664 in the training set and 141 in the test set. The algorithm has modules for: 1) auto-generated tumor/non-tumor masking using a trained residual neural network (ResNet34), 2) cell-density map generation (based on color deconvolution, local drain segmentation, and watershed transformation), 3) patch-level feature extraction using a pre-trained ResNet34, 4) a tower of linear SVMs for different cell ranges, and 5) a majority voting module for aggregating subtype predictions in unseen testing WSIs. The proposed system was trained and tested on several WSI magnifications ranging from x4 to x40 with a best ROC AUC of 0.95 and an accuracy of 0.86 in test samples. This fully-automated histopathology subtyping method outperforms similar published state-of-the-art methods for diagnostic WSIs.

Introduction

The most common form of lung cancer, non-small cell lung cancer (NSCLC), is further categorized into three major histopathological subtypes: ~40% Adenocarcinoma (LUAD), ~30% squamous cell carcinoma (LUSC), and the rest is large cell (undifferentiated) carcinoma. Classifying patients accurately is important for prognosis and therapy decisions. Classification currently requires manual review of whole slide images (WSIs) by medically trained pathologists, and incurs significant delays and cost. Here we present an automated algorithm to differentiate LUAD and LUSC diagnostic WSIs.

The problem of identifying lung cancer type from pathology images has been discussed extensively in recent years, and has been addressed most successfully using deep learning convolutional neural networks [1]. Most algorithms in literature such as those generated by Yu *et al* [2] and Coudray *et al* [3] were developed on images of fresh-frozen tissue sections [4]. These sections are primarily obtained to preserve molecular structures such as RNA that rapidly degrade, and are not representative of typical diagnostic workflows. In Yu, a feature selection process was performed on a set of engineered features (generated using CellProfiler software) to describe ten dense patches of size 1000 x 1000 px at x40 magnification (250 μ m²). Their algorithm achieved a receiver operating characteristic area under the curve (ROC AUC) of 0.75 using an SVM classifier with a Gaussian kernel. Coudray used a fully-trained Inception v3 deep network and achieved a state-of-the-art record performance with a ROC AUC of 0.95 when tested on frozen tissue WSIs from 137 LUAD and LUSC patients from TCGA [3]. The networks in Coudray were

trained and tested on x5 and x20 magnifications with 1024 μ m² and 256 μ m² patches, respectively.

In contrast to frozen sections, formalin-fixed paraffin-embedded (FFPE) biopsies are larger, and cellular morphology better preserved thus apparent in diagnostic WSIs. This allows the pathologist to examine multiple slides and different areas of sampled tissue to give a more accurate diagnosis [4]. The fidelity of morphology in diagnostic WSIs from FFPE sections has enabled use of cell identification/detection modules for cancer type classification as described in Vu *et al* [5]. They tested their algorithm on 32 diagnostic WSIs from TCGA (16 LUAD and 16 LUSC cases) with two post-processing classification techniques. The authors trained a ResNet32 for patch-level classification then applied majority voting to achieve accuracy of 0.78. They also achieved accuracy of 0.81 by using a random forest regression model as a post processing technique for classification.

In this paper, an image-based lung cancer subtyping method based on deep-learning feature vectors and cell-density maps is proposed for FFPE diagnostic WSIs (Figure 1). The results from our fully-automated histopathology subtyping method show that it outperforms similar published state-of-the-art methods for diagnostic WSIs.

We first present a brief description of the proposed system below, then provide details of all core components of the proposed system, and finally we show experimental results and comparison of performance to other algorithms reported in the literature.

The proposed system

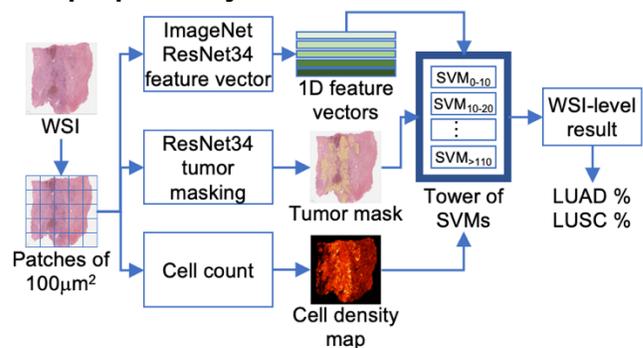


Figure 1. Block diagram of the automated adeno/squamous-cell NSCLC classifier. The classifier is based on tumor patches (represented by 1D vector in logits layer of pre-trained – ImageNet – RedNet34) of pre-specified ranges of cell count used in tower of SVMs.

A block diagram of the proposed adeno/squamous-cell NSCLC classification system is shown in Figure 1. In brief, the proposed system consists of the steps as follows. Input WSIs are first cropped into non-overlapping 100 μ m² patches and tumor patches are identified using a trained tumor/non-tumor ResNet34 module that detects and excludes stroma and adjacent non-tumor regions. Cell counts in all patches are generated using an image

segmentation module (generated by color deconvolution, local drain segmentation, and watershed transformation modules). Cell counts in tumor patches are then sorted into different ranges. Corresponding patches (represented by 1D vector in the logits layer of ResNet34) are assigned to SVM classifiers based on cell range. Finally, a majority-voting post processing technique is used to call input WSIs as LUAD or LUSC.

Auto-generated tumor masking

This module has two main steps. The first was development of ‘gold standard’ (ground truth) masks of tumor regions based on expert (pathologist) opinion and outlines. To assist pathologists manually annotating WSIs, we developed a few-shot (automated learning based on a limited number of examples) classification system where the human expert first initiates the masking process by selecting several tumor and adjacent non-tumor points in a given WSI. The masking system then groups patches similar to initial selections into the two groups: tumor and non-tumor regions. This process is iterated and refined as needed till the pathologist approves the final gold standard tumor mask. This process was applied to hundreds of diagnostic WSIs from TCGA.

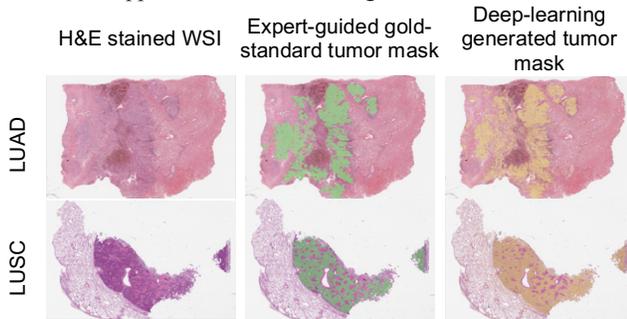


Figure 2. Assessment of trained tumor masks. Examples of deep-learning generated tumor masks based on the expert-guided system.

The second step comprises the use of masks from the aforementioned expert-guided system (in addition to patched WSIs) to train a ResNet34 module that can auto-generate tumor masks for WSIs. This module operates similarly to the cancer region detection method described in Li *et al* [6].

Applying this tumor/non-tumor masking module to unseen TCGA adeno- and squamous-cell NSCLC diagnostic WSIs, an average of 23.6K patches per image were found to contain tissue and of those 10.4K patches (44.1%) were classified as tumor by the masking module. Figure 2 shows two lung WSIs with gold standard and deep-learning generated tumor masks.

Cell density maps

In this module, a pre-processing step of hematoxylin and eosin (H&E) stain separation was applied to the input WSI. A local drain cell segmentation technique was applied to the H-channel, followed by watershed transformation for patch-based cell counting. Various methods for nuclei detection and cell segmentation are described in Irshad *et al* [7]. Figure 3 shows example 100 μm^2 patches from adeno and squamous-cell NSCLC where individual cells are shown with pseudo-colors. Cell counts are also given in the figure.

Tower of SVMs

A set of binary classifiers - linear SVMs here - were trained on 1D tumor patch representations from the logits layer of the residual neural network. For our system, we use ResNet34 that has been pre-trained on the ImageNet database. Cell counts of tumor

patches in use guided 1D representations to different SVMs. The proposed tower of SVMs has twelve binary classifiers that target patches of 100 μm^2 with cell counts from 0 to 10 cells per patch, 10 to 20 cells per patch, etc. up to >110 cells per patch.

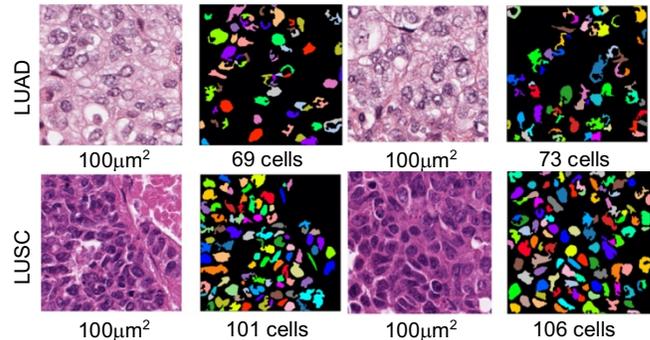


Figure 3. Cell density maps generated by stain separation, local drain, and watershed transformation modules.

Experimental results and discussion

The system was trained and tested on 876 subtyped NSCLC gigapixel-resolution diagnostic WSIs from 805 patients obtained from TCGA sources. Samples were randomly split into training (711 WSIs from 664 patients) and testing (165 WSIs from 141 patients) sets.

Our analysis of training data and the SVMs built is shown in Table 1, which lists training patch counts and patch-level SVM classification accuracy of all cell ranges. The total number of training patches used was 5,320K sampled from 711 training WSIs. This is about 7.48K tumor patches per training WSI. The table also shows that all classifiers are estimated with a sufficient number of samples (min = 169.26K and max = 737.35K train patches). Although the WSI level training data was 50.35% LUAD and 49.65% LUSC, this distribution was not found for some higher cell ranges. As seen in Table 1, ~60% of training patches with > 90 cells per patch were squamous-cell NSCLC. To identify the cell count range that was most predictive of subtype, we compared accuracies using the lift metric, which takes into account the prevalence of the majority class to adjust for changes in label imbalances across ranges. We found SVMs at intermediate-to-high ranges (40 to 90 cells per patch) performed similarly with the highest in the 60 to 70 cells per patch range when compared to their own majority class size. Overall the weighted average accuracy of the proposed tower of SVMs on training patches was found to be 72.14%.

Table 1. Tower of SVMs training patches and accuracy.

Ranges of cell counts	Training data		% accuracy (training)	Lift	
	100 μm^2 patches	%			
0 to 10	188.67K	48.08	51.92	73.08	1.41
10 to 20	169.26K	52.95	47.05	71.93	1.36
20 to 30	272.89K	54.70	45.30	70.86	1.30
30 to 40	418.68K	54.29	45.71	70.72	1.30
40 to 50	565.16K	52.72	47.28	70.75	1.34
50 to 60	671.68K	51.27	48.73	71.39	1.39
60 to 70	737.35K	49.80	50.20	71.75	1.43
70 to 80	708.27K	48.09	51.91	72.08	1.39
80 to 90	594.32K	45.40	54.60	72.67	1.33
90 to 100	427.83K	41.99	58.01	73.68	1.27
100 to 110	272.69K	38.19	61.81	75.61	1.22
110 to MAX	293.26K	38.62	61.38	73.82	1.20

The proposed automated adeno/squamous-cell NSCLC classifier was tested on 165 test WSIs from 141 patients. The test WSIs included 94 LUAD and 71 LUSD cases. Notice that random separation of patents into training and testing sets generated data sets with slightly different data distributions. That is, training set had 50.35% LUAD and 49.65% LUSC WSIs while the test set had 56.97% LUAD and 43.03% LUSC WSIs.

Table 2 shows the performance of the proposed automated adeno/squamous-cell NSCLC classifier on training and test WSIs for different patch sizes. The ROC AUC statistic and accuracy (with >50% of tested patches belonging to assigned label) are used as performance metrics herein. The table also shows that the proposed classifier achieves better performance with larger patch size.

accuracies are given in the figure. That is, starting from the top, the accuracy of correctly identified LUAD (Fig. 4a) and LUSC (Fig. 4b) WSIs decreases.

Table 2. Proposed automated adeno/squamous-cell NSCLC classifier performance on training and test WSIs for different patch sizes.

Patch size in μm^2	711 train WSIs (50.35 % LUAD & 49.65% LUSC)		165 test WSIs (56.97% LUAD & 43.03% LUSC)	
	ROC AUC	Accuracy	ROC AUC	Accuracy
100	0.93	0.85	0.93	0.84
200	0.94	0.85	0.93	0.85
300	0.94	0.86	0.94	0.86
1000	0.97	0.89	0.95	0.86

Fig. 5 shows that accuracy increases with patch size. WSIs accurately classified as either LUAD (left) or LUSC (right) using larger patch sizes show low accuracy when using patches of only $100\mu\text{m}^2$. If one uses a very large patch size of (for example) $4\text{K} \times 4\text{K}$ pixels at $\times 40$ magnification, a limitation is the reduced number of available patches to process and loss of resolution for further analyses such as intratumoral heterogeneity.

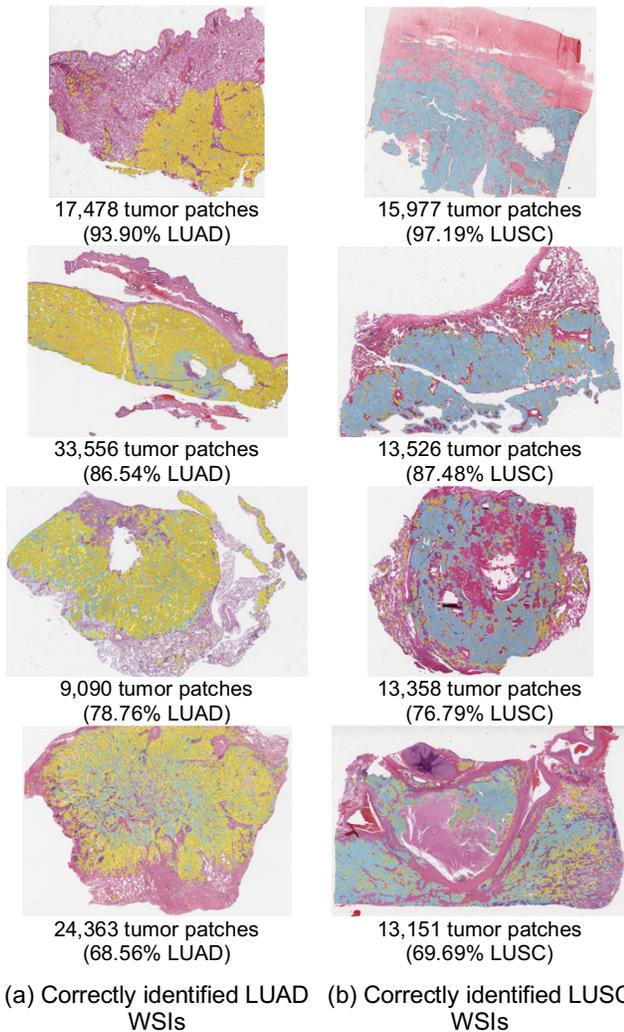


Figure 4. Test WSIs with overlaid label maps. Yellow indicates a tumor region classified as LUAD, and blue indicates a tumor region classified as LUSC.

Figure 4 shows classification results from the proposed system for two sets of WSIs from TCGA adeno (a) and squamous-cell (b) NSCLC wherein images are overlaid with colored label maps. Yellow indicates a region classified as LUAD, blue indicates a region classified as LUSC, and non-tumor regions are not colored. The classification maps are for $100\mu\text{m}^2$ patches from test WSIs; total number of tumor patches and accuracy of identification are also shown. Example test WSIs with variety of

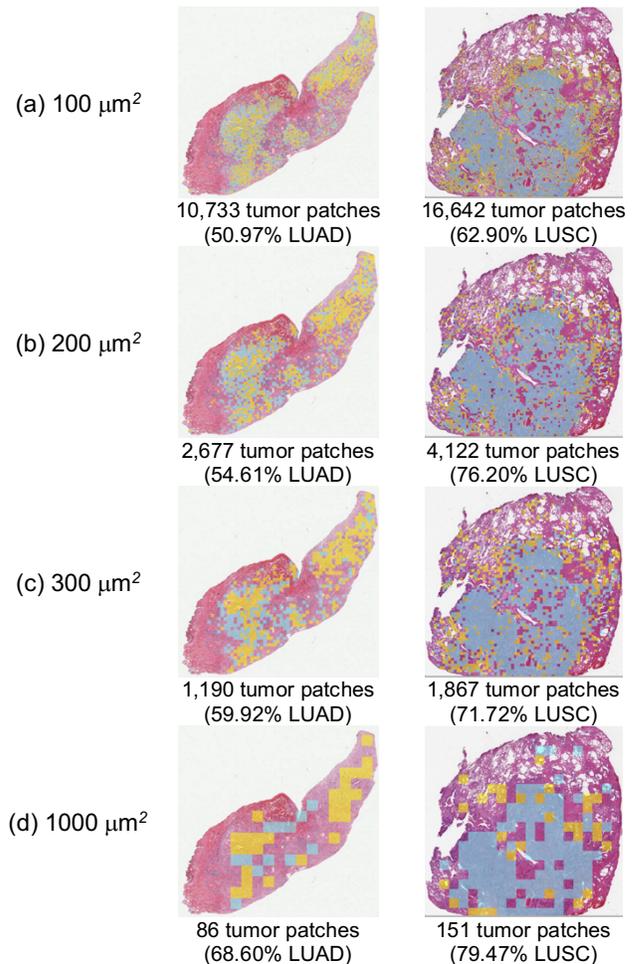


Figure 5. Larger patch size improved classification accuracy. Yellow indicates a region classified as LUAD and blue indicates a region classified as LUSC. Percentage accuracy is shown beneath each image.

Table 3. Performance of the proposed classifier relative to state-of-the-art methods. Our algorithm showed comparative or better performance.

	Train and test data	Test patients	Feature space	Classifier	Patch size		Performance	
					μm^2	Pixels	ROC AUC	Accuracy
Yu, K.-H. <i>et al.</i> (2016) [2]	LUAD and LUSC frozen tissue WSIs from TCGA	x	250 engineered features	SVM with gaussian kernel	250	1K @ x40	0.75	x
				SVM with linear kernel	250	1K @ x40	0.70	x
Coudray, N. <i>et al.</i> (2018) [3]	LUAD and LUSC frozen tissue WSIs from TCGA	N = 137	Partially-trained Inception v3		256	512 @ x20	0.85	x
			Fully-trained Inception v3		256	512 @ x20	0.95	x
			Fully-trained Inception v3		1024	512 @ x5	0.94	x
Vu, Q. <i>et al.</i> (2019) [5]	LUAD and LUSC diagnostic WSIs from TCGA	N = 32	Fully-trained ResNet32 with majority voting		128	256 @ x20	x	0.78
			Fully-trained ResNet32 with random forest regression		128	256 @ x20	x	0.81
Proposed algorithm	LUAD and LUSC diagnostic WSIs from TCGA	N = 141	512 from ImageNet-based ResNet34	Tower of 12 linear SVMs based on cell density	100	400 @ x40	0.93	0.84
					200	800 @ x40	0.93	0.85
					300	1200 @ x40	0.94	0.86
					1000	4K @ x40	0.95	0.86

Table 3 summarizes performance of the proposed classifier in comparison to other reported state-of-the-art methods. Yu *et al.* [2] developed a methodology for patch selection based on stain density and used 250 engineered features to represent such patches in SVM classifiers. Their algorithm was trained and tested on LUAD and LUSC frozen tissue WSIs from TCGA with 0.70 and 0.75 ROC AUC values for SVM with a Gaussian kernel and SVM with a linear kernel, respectively. This indicates that our deep-learning system outperforms similar efforts that were based on CellProfiler (engineered) features. A set of deep-learning systems presented in Coudray [3] are also included in Table 3. These systems were also trained and tested on LUAD and LUSC frozen tissue WSIs from TCGA. They achieved a ROC AUC of 0.95 with a fully-trained Inception v3 classifier using a patch size of 256 μm^2 . The number of patients in their test set (N = 137) was slightly lower than ours (N = 141). Finally, in a 2019 paper by Vu *et al.* [5], a LUAD vs. LUSC fully-trained ResNet32 classification scheme was described that was trained and tested on TCGA diagnostic WSIs. Their limited set of test WSIs (N = 32) shows a WSI-level classification accuracy of 0.81, its accuracy is thus inferior to that achieved by our automated adeno/squamous-cell NSCLC classifier (0.84 to 0.86 using classifiers with different patch sizes) on a much larger test set. In summary, Table 3 shows that the proposed algorithm was tested using a larger number of test patients (N = 141) and achieved comparative or better overall performance.

Conclusions

Herein, we report our development and validation of an adeno/squamous-cell NSCLC classification system for FFPE preserved tissue using patch-level cell count information and representation from pre-trained ResNet34 (ImageNet). This fully-automated histopathology subtyping system outperforms similar published state-of-the-art methods for diagnostic WSIs. The system also generated maps of (tumor) regions-of-interest within WSIs, providing novel spatial information on tumor organization. Our results showed tumor patches of 100 μm^2 with 60 to 70 cells distinguished LUAD from LUSC better than other cell-density ranges. Moreover, the proposed system showed LUSC WSIs have

many tumor regions with high cell density in comparison to LUAD. The system also illustrated that cropping input WSI into larger patches for processing increased the overall adeno/squamous-cell NSCLC classification accuracy.

Acknowledgments

The authors would like to thank John Yahya Elshimali, M.D., FCAP, FASCP for his digital pathology support and Patricia Spilman for editing this manuscript.

References

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp 436-444 (2015) doi.org/10.1038/nature14539
- [2] K.-H. Yu, C. Zhang, G. Berry, R. Altman, C. Ré, D. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Communications*, vol. 7, Article number: 12474 (2016) doi:10.1038/ncomms12474
- [3] N. Coudray, P. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559-1567 (2018) doi.org/10.1038/s41591-018-0177-5
- [4] A. BenTaieb and G. Hamameh, "Deep learning models for digital pathology," Available online: <https://arxiv.org/abs/1910.12329> (accessed on Nov 05, 2019).
- [5] Q. Vu, S. Graham, T. Kurc, M. To, M. Shaban, T. Qaiser, N. Koohbanani, S. Khurram, J. Kalpathy-Cramer, T. Zhao, R. Gupta, J. Kwak, N. Rajpoot, J. Saltz, and K. Farahani, "Methods for segmentation and classification of digital microscopy tissue images," *Front Bioeng Biotechnol.* vol. 7 (2019) doi: 10.3389/fbioe.2019.00053
- [6] Z. Li, Z. Hu, J. Xu, T. Tan, H. Chen, Z. Duan, P. Liu, J. Tang, G. Cai, Q. Ouyang, Y. Tang, G. Litjens, and Q. Li, "Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study," Available online: <https://arxiv.org/abs/1803.05471> (accessed on Nov 05, 2019).
- [7] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential," *IEEE Rev Biomed Eng.* Vol. 7, pp. 97-114 (2014) doi: 10.1109/RBME.2013.229580

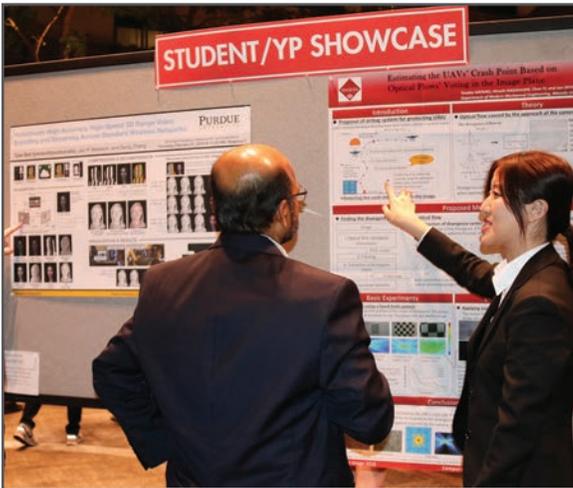
JOIN US AT THE NEXT EI!

IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!



- **SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •**
- **INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •**

www.electronicimaging.org

