Understand the Human Body, Improve Human Condition

Early Detection of Non-Small Cell Lung Cancer using HDSCA 3.0

Harinarayana Mellacheruvu[1], Jiyoun Seo[2], Jaden Mullin[3], James Hicks[3], Peter Kuhn[3]

[1]*Department of Quantitative Biology, Convergent Science Institute in Cancer, University of Southern California, Los Angeles, CA*

[2]*Department of Computational Biology and Bioinformatics, Convergent Science Institute in Cancer, University of Southern California, Los Angeles CA*

[3]*Department of Biological Sciences, Convergent Science Institute in Cancer, University of Southern California, Los Angeles CA*

USC Michelson Center
Convergent Science Institute in Cancer

Introduction

Non-Small Cell Lung Cancer (NSCLC) accounts for 85% of all lung cancers, and a large proportion of patients with NSCLC are diagnosed late stage. Furthermore, the five-year survival rate drops massively when NSCLC progresses from early to late stage. Thus, it is imperative that we can diagnose NSCLC early in its course where preventative measures can be employed. Early detection of NSCLC is one of the main goals of the Stanford Study (SU) where we aim to implement circulating rare events identified using the High-Definition Single Cell Assay (HDSCA) to identify the different clinical groups in the SU cohorts, distinguish between malignant NSCLC and benign controls, and utilize this biomarker in clinical settings given the current lack of biomarker usage in NSCLC detection and diagnosis due to sensitivity and specificity deficiency. Lung nodules identified by PET-CT imaging have an extremely high false positive ate and current prediction models aren't optimal when stratifying benign and malignant nodules. Using HDSCA 1.0 which focused only on CK+/CD45- cells, it was found that these tumor cell aggregates enhanced diagnostic and prediction accuracy of NSCLC when used in conjunction with clinical, demographic, and PET-CT variables. However, the prediction model built had poor diagnostic accuracy when data from the second cohort from the study was implemented. With the improved HDSCA 3.0 assay, which utilizes an unsupervised hierarchical clustering algorithm, OCULAR, to find all types of rare events, rather than giving bias to only CK+/CD45- cells, we want to see whether the classification accuracy of NSCLC vs benign controls improves and whether that accuracy is preserved when data from the second cohort is included.

Objective: With the transition from detecting CTCs to detecting rare events, we want to see whether HDSCA 3.0 data will add to the diagnostic capability of prediction models. We used a multitude of computational workflows in order to better characterize circulating rare events from NSCLC patients and benign controls, with the goal of detecting patterns which can be utilized to help achieve NSCLC early detection.

HDSCA Platform

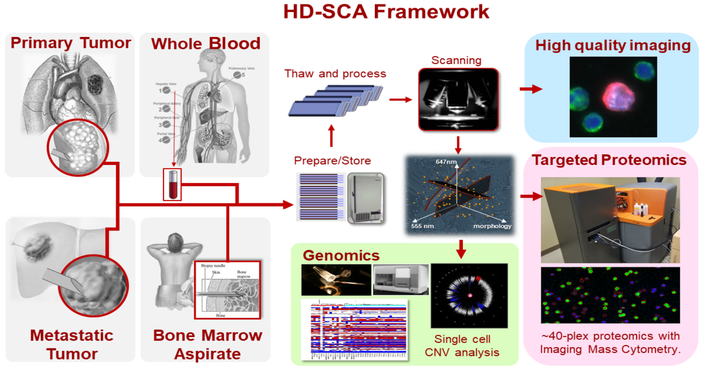

Figure 1: The High-Definition Single Cell Assay is used to understand cancer's basic unit of disease, the cell. Circulating tumor cells in the blood are a rare one in a million occurrence, leading to a needle in the haystack problem. HDSCA 3.0 uses a "no cell left behind" approach to interrogate all cells from a patient sample slide, using an unsupervised hierarchical clustering algorithm, OCULAR, to group "common" and "rare" cell events.

Input Data and Methods

The input data for our learning and computational analysis included 1102 cells from 44 patients, collected in a diagnostic workup and one-time blood draw from high-risk patients prior to FDG PET Imaging. Of the 44 patients, 22 patients had malignant NSCLC and the other 22 patients were benign controls. Furthermore, our 22 NSCLC patients consisted of 15 stage I NSCLC patients, 2 stage II NSCLC patients, 1 stage III NSCLC patient, and 4 stage IV NSCLC patients. Each circulating rare event is defined by morphological and channel specific parameters derived from the R EBImage package. Each cell is also labeled with a class, either benign versus malignant, or benign vs stage I NSCLC for our sub-classification, in order to classify cells as either originating from a benign control or a NSCLC patient. In previous studies, using cell feature data and HDSCA 1.0 from 131 patients, including 6119 cells in the training set and 1452 cells in the test set, a support vector machine algorithm was used to predict benign from malignant NSCLC, with an Area Under The Curve of 0.6142 (AUC = 0.6142). In addition to a support vector machine algorithm, we employed a random forest classifier which is based on decision trees, in conjunction with HDSCA 3.0 data, in order to assess and compare performance

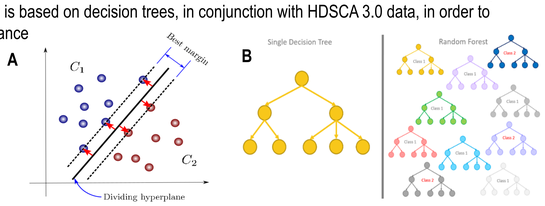

Figure 2: Conceptual visual representation of A) Support Vector Machine and B) Random Forest

Classification Models

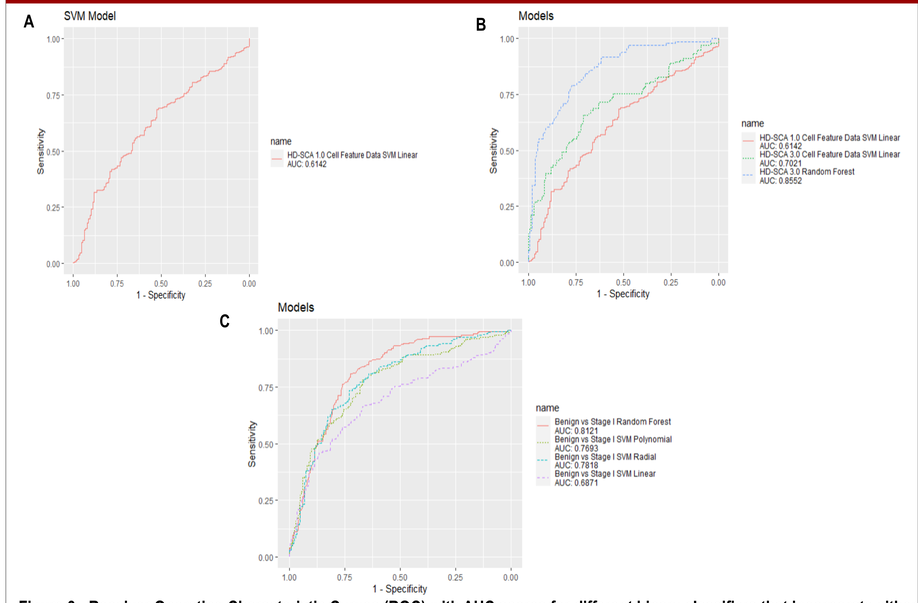

Figure 3: Receiver Operating Characteristic Curves (ROC) with AUC scores for different binary classifiers that incorporate either HDSCA 1.0 or HDSCA 3.0 single cell data. A) Support vector machine that utilizes HDSCA 1.0 data from 131 patients, including 6119 cells in the training set and 1452 cells in the test set. B) Support Vector Machine and Random Forest classifier utilizing HDSCA 3.0 single cell data. C) Performance of Random forest classifier compared to different support vector machines when classifying benign vs stage I NSCLC

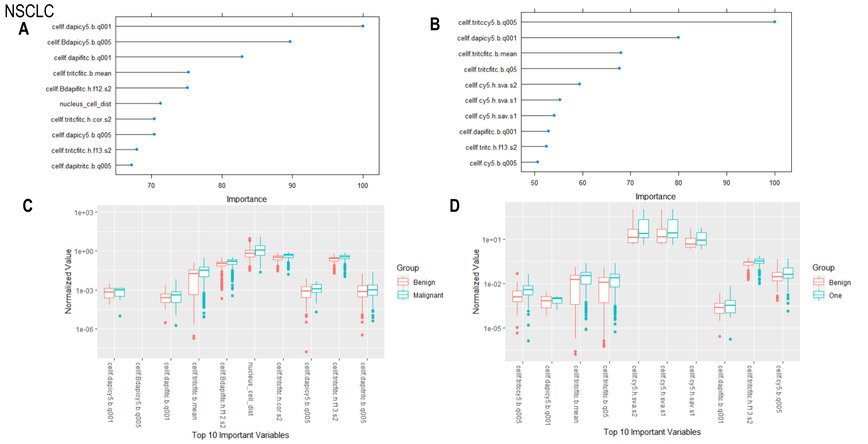

Figure 4: Top 10 most important variables when training the random forest classifier with HDSCA 3.0. There is a mix of basic and hairlick features that contribute to the model's capability, with basic features representing statistics on pixel intensity and hairlick features quantifying pixel texture. A) Most important variable for RF classifier for benign vs NSCLC is the 1st percentile intensity of the blended DAPI and CY5 channel. B) Most important variable for RF classifier for benign vs stage I NSCLC is the fifth percentile intensity of the blended TRITC and CY5 channel. C) Difference in top 10 variables for cells from benign controls and NSCLC patients, with values appearing to be higher in malignant patients. D) Difference in top 10 variables for cells from benign controls and stage I NSCLC patients.

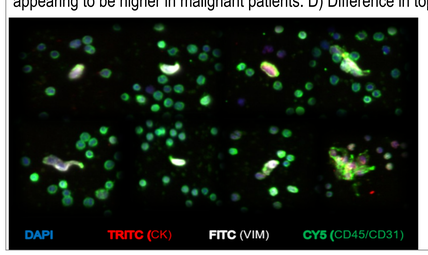

Figure 5: Cells with highest predicted probability of belonging to NSCLC patient when using random forest classifier. We see many CD45 and vimentin positive cells, as well as the presence of CD31. The cells observed are of different shape compared to traditional white blood cells and display different staining expression profiles.

Deep Learning Classifier

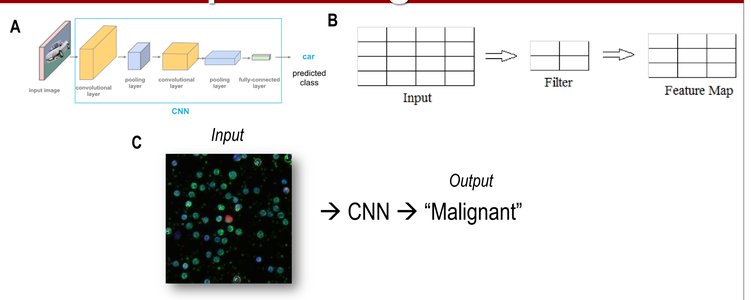

Figure 6: Relevance and applications for CNN and deep learning architecture with HDSCA 3.0 single cell data. A) Conceptual overview of convolutional neural network, which can be fitted to image data as is in order to make a binary classification. CNNs consist of multiple different types of layers, each with a specific purpose aimed to analyze the image. B) The convolutional layer, which contains filters (matrices) which are applied to the input to detect specific features and create a feature map. Work is done on these feature maps, including pooling which summarizes the features in the feature map and reduces the computation needed, and flattening, which reduces dimensionality, which are then passed to more layers in the neural network in order to output a class. Convolutional neural networks can find meaning and patterns from images and input that humans may struggle to find. C) Application of convolutional neural network to HDSCA 3.0 single cell images.

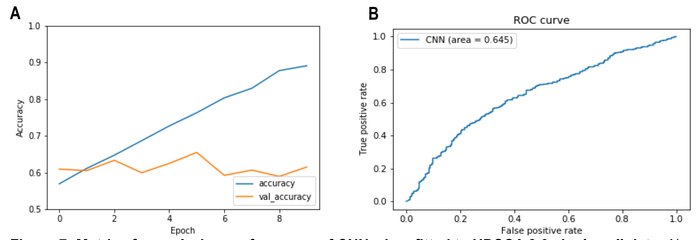

Figure 7: Metrics for analyzing performance of CNN when fitted to HDSCA 3.0 single cell data. A) Training and Validation Accuracy when fitting convolutional neural network to HDSCA 3.0 single cell data**.** Over 10 epochs, with 1 epoch representing the entire dataset being passed forward and backward through the neural network, training accuracy reached 0.9 and validation accuracy reached 0.61. B) ROC curve using AUC as a metric to assess performance of our benign vs malignant classification for convolutional neural network classifier. We observe an AUC score of 0.645 when classifying cells from the 44 patients.

Conclusions and Future Directions

Using supervised machine learning methods, as well as deep learning architectures, we can preliminarily classify cells that originate from benign controls versus NSCLC patients with cell morphological and channel specific data in either numerical parameter or image-based form. However, further analysis is needed for both algorithms. For our machine learning algorithms, we need to continue further expanding our OCULAR cell dataset, calibrate out model and tune its hyperparameters, and continue testing with different models for a conclusion to be made on whether we can identify the different clinical groups in the SU cohorts using HDSCA 3.0. Our deep learning classifier can also be optimized by changing the number of filters used during the convolution process, as well as the number of layers through which inputs are passed, and more obvious changes such as altering the image size. Furthermore, we can explore different forms of data augmentation such as using generative adversarial networks (GANs) and autoencoders to increase the quality and amount of available data to be analyzed respectively and increase model performance. We can also look towards other deep learning architectures such as residual neural networks, capsule networks which expand upon the CNN.

Acknowledgments

First and foremost, we want to thank all patients in the SU cohorts, who contributed their blood and tissue specimens to research and without whom this study would not have been possible. Appreciations also to Xiomara Villasenor and the rest of our technical team of USC for processing and cryobanking the blood samples. Rafael Nevarez, Nicholas Matsumoto, Shoujie Chai, provided support in the technical data analysis. Great appreciations also to my mentor Jiyoun Seo for guiding me throughout this project, and principal investigators Dr. Kuhn and Dr. Hicks for giving me the opportunity to work in such an amazing environment.

CONTACT

Harinarayana Mellacheruvu: mellache@usc.edu
Kuhn-Hicks Laboratory, CSI-Cancer: http://kuhn.usc.edu/

SCAN ME