Introduction

Breast cancer is the second most prevalent cancer worldwide, with nearly 2 million new cases each year. Histopathology evaluation, which involves pathologists' visual assessment of microscopic tissue slices, is indispensable for the diagnosis of breast cancer. However, studies have shown that the billions of pixels from the slide scans contain a wealth of untapped biological signal that has yet to be interpreted. In this project, we combine convolutional neural network (CNN) based image analysis of these scans with transcriptomic analyses to uncover novel molecular and morphological profiles associated with hormone receptor status and genomic subtypes.

Methods

- **Data from 1099 BIRCa patients was collected from the TCGA cohort.**
- **Whole slide images were tiled at their densest location.**
- **200 256x256 tiles per slide were used.**
- **Caffe2 was used with the Inception V3 CNN to predict clinical/genomic parameters.**
- **Tumor vs. Normal**
- **Hormone Status**
- **PAM50 Status**
- **Rotations were used for balanced tile-level training sets.**
- **Models were initialized without pretrained model weights, while hyperparameters were determined through transfer learning from lung cancer models.**
- **Hold-out cross validation was conducted.**

Hormone Receptor Status:

- **ER/PR status is normally determined with the aid of antibodies or immunohistochemistry stains, as opposed to directly from an H&E stained slide.**

Results

- **The CNN-based classifiers showed strong performance over every examined task.**
- **Our CNNs replicated the strong tumor vs. normal performance described by Liu, et al. in the TCGA cohort.**
- **Our analysis extends to histological status, hormone receptor status, and PAM50 status for the first time.**
- **To correlate strong CNN predictive performance with genomic and morphological features, we investigated (i) ER and PR status and (ii) PAM50 status in greater detail.**

**Hormone Receptor Status:**

- **Hormone receptor classifiers had 92.6% accuracy over patients with discordant receptor status (ER/PR- or ER-/PR+).**
- **Ignoles that PR-specific knowledge was learned by classifiers.**
- **Ridge regression over RNAseq expression had AUCs of 0.910.838 for determining ER/PR status (+/-) respectively.**
- **Genes associated with ER/PR status were enriched in immune-related terms, suggesting that the morphological signal might come from immune infiltration.**
- **56% and 64% of significant GO terms were immune-related for ER and PR status respectively.**
- **Terms included innate immune response, defense response to bacterium, and regulation of STAP protein.**
- **The Random Forest trained lymphocyte detector had 95% true positive/negative accuracy rate over pathologist labeled slides.**
- **A threshold convolutional node in the ER classifier was localized preferentially with lymphocytes (R = 0.55263, p < 0.05) compared to cell nuclei (R = 0.03364, p > 0.05).**

**PAM50 Status:**

- **2 models were constructed to classify PAM50 status: (i) a single 4 class classifier and (ii) three binary classifiers arranged in a tree.**
- **The penultimate feature vector of the Luminal AB binary classifier was extracted, and its correlation with microscopic expression was examined.**
- **4 Class Model:**
  - Distributions of reported confidence associated with correctly/incorrectly classified patients were compared with the resulting clusters
- **SNE was conducted over microarray data for PAM50 gene, predictions were compared with the resulting clusters.**

Conclusions

- **A majority of misclassifications involved classifying Luminal B patients as HER-2 Enriched/Luminal B.**
- **Consistent with existing understanding of PAM50-hierarchical clustering over microarray data grouped Luminal B and HER-2 enriched, while RNA-seq t-SNE clustering was unable to resolve Luminal A from B.**
- **The Luminal A/B binary classifier was able to separate the two classes with high AUC (0.860-0.939).**
- **t-SNE dimensionality reduction of 6000+ features from the Luminal A/B classifier revealed a composite feature with significant correlations with a set of PAM50 gene expression values.**

Acknowledgements

K. J. is a Howard Hughes Medical Institute Student Fellow, and BB is supported by NIH Ruth L. Kirschstein National Research Service Award #F31CA170311-01 and #T32CA179857-01. We thank the NVIDIA Credit for Research: Microsoft Azure for Research Award, and the NVIDIA-GPU Grant Program for their computational support.