A Novel Transformation of Genetic Information into Images for Improved Primary Cancer Classification

Xu, Sidra (School: The Harker School)

Cancer is a leading cause of human death worldwide, yet quick and accurate identification of the primary cancer site is still lacking. Owing to its ability to differentiate similar tumors and deliver quick results, somatic point mutation-based cancer classification has generated growing interest. However, with a status quo of 65% accuracy, significant improvements are still needed for practice use. To address this issue, a novel approach is proposed in this study: gene expression embedding-augmented, mutation-based cancer classification with convolutional neural networks (CNNs). This methodology allows the algorithm to harvest information in both somatic mutations and gene expressions without the need for the latter from patients, often not available in a clinical setting. It also leverages the incredible success of CNNs through the conversion of genetic information into images. More specifically, genes are clustered by functional annotations and transformed into images by encoding expression embedding vectors in the blue channel and mutation frequencies in the red channel. Images are inputted into InceptionV3 for classification with transfer learning using pretrained ImageNet weights. The model currently provides a prediction accuracy of up to 77%, a significant improvement from previous studies, confirming the huge potential of the proposed image conversion methodology. Feature ranking of the model using class activation mapping and gene ontology revealed multiple genetic markers and biological processes specific to each cancer type, which could be targeted for therapeutic drug development. This is the first time a new dimension of information, ie. the spatial dependencies hidden behind genetic data, has been unlocked for improved cancer classification.

Awards Won: