Weakly Supervised LSTM RNNs for Longitudinal Breast Cancer Recurrence Prediction via Unstructured Clinical Narratives

Sanyal, Josh (School: Homestead High School)

Breast cancer is the most common cancer in women globally, causing an estimated 626,700 deaths in 2018. As the late detection of breast cancer recurrence leads to a 1-1.5% curability rate, my project presents the first attempt to predict breast cancer recurrence, 1 year in advance, by leveraging unstructured clinical narratives in EHR over a patient's temporal visit sequence. I first develop robust, encoded representations of the unstructured clinical narratives. This is done through text preprocessing in which notes are standardized, noise is eliminated, and semantic meaning is preserved. Next, I train unsupervised word2vec embeddings across a corpus of ~100 million notes from 3 institutions for increased generalization. Finally, robust document-level embeddings are computed as a weighted average of word vectors in the document. Using the document-level embeddings, I train a weakly supervised, many-to-many, stacked stateful LSTM RNN for longitudinal recurrence prediction. The model is validated, trained, and tested on ~154K notes from 940 distinct patients, achieving 0.841 ROC AUC on the hold-out test set. The model's comprehensive, fully automated analysis of unstructured clinical narratives with limited ground truth labels can also be generalized to other types of cancers and diseases to offer key insights into future clinical events. Unlike previous "black-box" models, my model can be visualized, engendering trust in clinicians to leverage and utilize its personalized predictions as a pivotal part in determining adequate treatment. This allows clinicians to improve patient survival by treating recurrences earlier and even preventing them from occurring in the first place.