BioRx: An Integrative NLP Approach to Early Survival and Recurrence Prediction and Novel Biomarker Discovery in Unstructured Text-Based Clinical Narratives for Diabetes Patients

Banda, Snikitha (School: Notre Dame High School San Jose)

Type I and Type II Diabetes are the number one cause of kidney failure, affecting 422 million people and being responsible for 1.5 million deaths worldwide. The leading upstream cause of diabetic death is a lack of clinically approved biomarkers, causing inaccurate survival and recurrence predictions for patient risk stratification and treatment management. Thus, my project presents the first low-cost, noninvasive computational platform to predict the survival and recurrence rates of diabetes patients one year in advance using newly discovered genetic biomarkers in patients' unstructured clinical narratives. The first step is to train the word2vec model to text process ~89 million clinical narratives from 678 patients by standardizing notes, eliminating noise, and preserving semantic meaning. By developing these robust, encoded representations, the model generates a dataset of document-level embeddings, which is leveraged to train an unsupervised Multilayer Perceptron (MLP) network. The model simultaneously evaluates the effect of prognostic factors on survival and predicts a rate for each patient. The identification and diagnosis models achieved AUROCs of 0.91 and 0.95 respectively, successfully exceeding all incumbent clinical techniques. This technique can be generalized to any disease types documented in EHRs and will lead to better clinical outcomes and spare patients from unnecessary aggressive therapies. Ultimately, this platform helps solve one of precision medicine's primary limitations and is a pragmatic tool with industrial viability that offers clinicians the ability to leverage its personalized predictions to amplify diagnostic accuracy, determine robust treatments earlier, and save millions of lives.

Awards Won:

Second Award of $2,000