Machine Learning based Feature Selection of RNA Seq data

RNA-Sequencing or RNA-Seq is a sequencing used to study the presence and quantification of RNA in a biological sample. Cancer RNA-Seq helps scientists in various ways like determining which variants are expressed in cancer sample,

The Cancer Genome Atlas Project (TCGA) is one of the largest cancer genomics programs, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. It hosts over 2.5 Petabytes of genomic, epigenomic, transcriptomic, and proteomic data.

Our Aim was to use machine learning to uncover the genes and corresponding proteases responsible for the onset of Head and Neck Squamous Carcinoma (HNSC), commonly known as Oral Cancer.
We chose to use RNA Seq gene expression data as it is widely used as the database of choice for training models and is aligned with our problem statement.

HTSeq - FPKM(Fragments per Kilobase of transcript per Million mapped reads) dataset of 3 cancers - Oral(HNSC), Thyroid (THCA), and Esophageal(ESCA) was downloaded from UCSC Xena, and combined. The idea was that since THCA and ESCA are physiologically similar to HNDC, a model that performs well on these will definitely perform well on the other cancer types. The final dataset was 60483(genes/features) × 1287 (Sample patients).

Extreme Gradient Boost(XGBoost) algorithm, which is an ensemble-based algorithm, was chosen as it generally performs really well and is used widely in academia for similar purposes.

Due to hardware limitations, GridSearch based hyperparameter tuning was not possible. Despite this, the model gave a mean accuracy of 99.22% over 5 k-fold validations.

Xgboost reported the most important features(54) by Gain which was plotted. The studies of the roles of these genes hold the answer to our study.

Secondly, the dataset was also passed to Boruta which is an algorithm specifically for feature extraction. At the time of writing this, the output is being studied.

There is immense value in further extending our efforts to arrive at a better answer. Differential gene expression analysis, feature extraction techniques, curation of different datasets are some of the possible avenues of research.