Predicting the Development of Secondary Central Nervous System Cancer through Ensemble Learning Methods

Camacho, Julia Christina (School: Texas Academy of Mathematics and Science)

Secondary cancers, which develop as a result of initial radiation or chemotherapy treatments, are a major cause of morbidity and mortality in cancer survivors. Early prediction of the development of secondary cancer is crucial for determining optimal treatment and prevention strategies, and significant inter-individual variability in the risk of developing secondary CNS (central nervous system) cancers suggests that genetics may play a role in patient susceptibility. This project developed a computational method for the prediction of secondary CNS cancer through ensemble learning approaches utilizing both clinical and genetic data. Data, including radiation doses and 89 SNPs (single nucleotide polymorphisms), were obtained from a 2017 COG (Children's Oncology Group) study. Feature selection was performed, and 8 machine learning models were trained using all features and then 10 selected features. Then, 4 types of ensemble models (bagging, boosting, voting, and stacking) were constructed using combinations chosen to maximize model diversity. Grid searching was utilized to optimize hyperparameters, and the model evaluation metrics used were accuracy and ROC AUC scores. The 10 most important features were radiation, age, and 8 SNPs in genes such as BRCA2 and XRCC5; knowledge of these genetic variants is critical for treatment and prevention. Naively adding genetic data directly to clinical models did not immediately increase prediction accuracy, which indicated that feature selection was needed to filter out noise in the genetic data. Models trained on 10 selected features were more accurate than models trained on all features, and the highest ensemble accuracy was achieved through voting.

Awards Won:

Fourth Award of $500