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

Julia Christina Camacho
Texas Academy of Mathematics and Science

**Presented at the following competitions and symposiums in 2019:**

Intel International Science and Engineering Fair (4th Place in Computational Biology)
Texas Junior Academy of Science (1st Place in Computer Science,
2nd Place Overall in Physical Sciences - AJAS Qualifier)
Texas State Science and Engineering Fair (2nd Place in Computational Biology)
Fort Worth Regional Science and Engineering Fair (1st Place in Computer Science,
3rd Place Overall in Best of Fair, and multiple special awards)
TAMS Interdisciplinary Research Fair (1st Place in Computer Science)

**Abstract:**

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. However, few methods currently take genetics into account in assessing secondary-cancer risk. 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 then performed, and eight machine learning models were trained using all features and then 10 selected features. Then, to further improve prediction accuracy, four types of ensemble models (bagging, boosting, voting, and stacking) were constructed using combinations chosen to maximize model diversity. Models trained on 10 selected features proved to be more accurate than models trained on all features, and the highest ensemble accuracy was achieved through voting. Grid searching was used to optimize hyperparameters, and the model evaluation metrics used were classification accuracy and ROC AUC scores.

The 10 most important features identified through feature selection were radiation, age, and eight SNPs in genes such as BRCA2 and XRCC5; knowledge of these genetic variants is critical for primary cancer treatment and secondary cancer prevention. Additional analysis on these SNPs can be performed in the future to discover the particular mechanisms of secondary-CNS cancer development.

These computational models, among the first of their kind to utilize both clinical and genetic data, stand to dramatically improve current procedures for cancer treatment selection by predicting the risks of radiotherapy on a patient-by-patient basis. The ability to input different clinical and genomic features in generating predictions of treatment risk thus also has direct applications in precision medicine, where personalized regimens serve individual patients.

**Acknowledgement of Major Assistance:**

I conducted my research at the University of North Texas with assistance from Dr. Xuexia Wang, associate professor of mathematics, from the fall of 2018 to the summer of 2019. A complete statement on outside assistance is included at the end of this paper.

---

**Table of Contents**

## Introduction

In this project, I developed a computational method for the prediction of secondary central nervous system (CNS) cancer development in survivors of childhood cancer.

Secondary cancers, which develop as a result of initial radiation or chemotherapy treatments and are histologically unrelated to the initial (primary) cancer, are a major cause of morbidity and mortality in cancer survivors. For instance, in a 2016 study, 8.1% of cancer patients developed secondary malignancies. Out of these patients, 55% died as a result of their secondary cancer (Donin *et al*, 2016). CNS (brain and spinal cord) cancers are especially devastating, with pediatric 15-year mortality rates of 14% (Morris *et al*, 2007). In addition, CNS tumors can directly affect patients' thought processes and movements, which drastically increases the difficulty of life while patients undergo treatment.

Early prediction of the development of secondary cancer is crucial for determining optimal primary cancer treatment and secondary cancer prevention strategies, as development of secondary cancer can possibly be avoided altogether if correct treatments are chosen for patients. If secondary cancer risk can be determined for patients, then current treatment recommendation procedures for primary cancers can be improved with more accurate assessments of the total risks associated with various types of cancer treatments such as radiation, chemotherapy, and immunotherapy.

Recent research has revealed that significant inter-individual variability exists in the risk of developing secondary CNS cancers (Zhao *et al*, 2013). This disclosure suggests that genetics may play a role in patient susceptibility, which only serves to increase the importance of patient-by-patient risk evaluation.

However, few methods currently take genetics into account when assessing secondary-cancer risk, which would increase predictive accuracy and allow for more effective precision-medicine approaches to patient treatment. To address this need, I developed a computational method for the prediction of secondary CNS cancer through ensemble machine learning approaches utilizing both clinical and genetic patient data.

**Materials, Methods, and Procedures**

I had three key aims in my project methodology: first, to utilize both clinical and genetic data in developing accurate machine learning predictors for cancer development; second, to determine the genetic features most correlated with secondary development; and third, to identify the biological significance of such features.

I took several key steps (shown below in Figure 1) in developing my machine learning models: data preparation, feature selection, training and testing of individual models, ensemble model construction and subsequent training and testing, and finally, genetic analysis of the most important features.

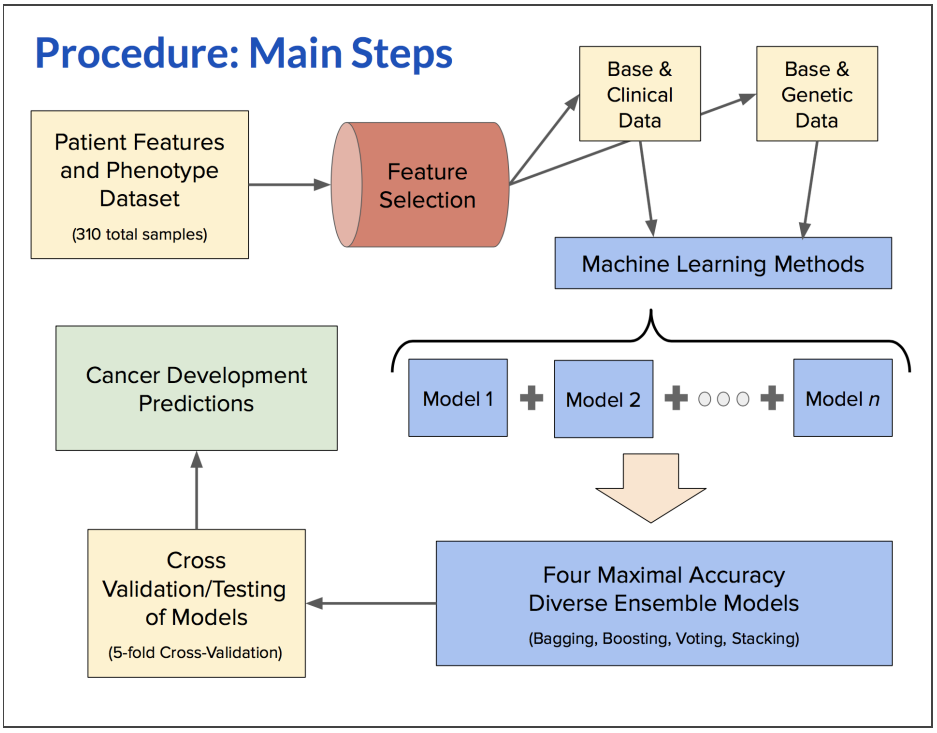



**Figure 1.** Key methodological steps.

To begin, I obtained genetic and clinical data from a 2017 Children's Oncology Group (COG) study. In this study, 121 medical institutions contributed cancer patient data, resulting in a total dataset of 310 individuals, or data samples, in total. Each patient had been treated with cranial radiation therapy (CRT) during their primary cancer treatment. Case/control matching was present: 82 of those patients were treated as cases, as they had developed secondary CNS cancer, and 228 were treated as controls, since they had not developed secondary CNS cancer (shown in Table 1). Thus, I set up the problem as a binary classification challenge.

| Number of **Cases** (sec. cancer developed) | Number of **Controls** (no sec. cancer developed) |
|---|---|
| 82 | 228 |

**Table 1.** The number of cases and controls in the dataset.

The data consisted of 96 features in total: 4 base features related to patient background, 3 clinical features, and 89 genetic features in the form of SNPs, or Single-Nucleotide Polymorphisms, a type of genetic point mutation that can play a major role in carcinogenesis.

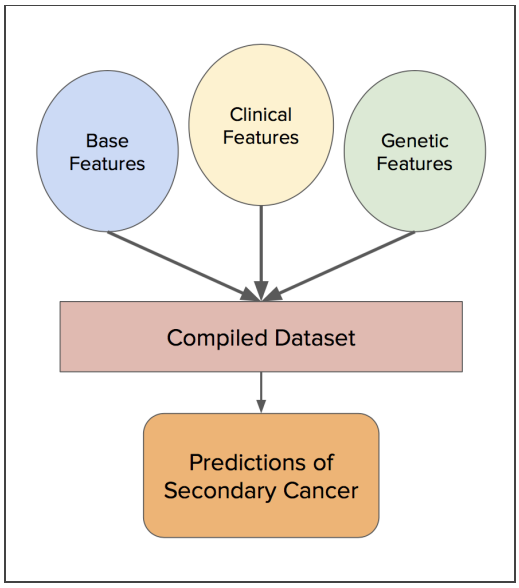



**Figure 2.** Visualization of the dataset composition.

The base features consisted of the current age, age at diagnosis, sex, and ethnicity of each patient. The clinical features were the quantitative dosage amount of CRT that the patient received during treatment for their primary cancer, as well as the histological types of their primary and secondary cancers. The types of primary cancers consisted of acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), bone, brain, carcinoma, germ cell, Hodgkin's and non-Hodgkin's lymphoma, kidney, neuroblastoma, papilloma, retinoblastoma, and sarcoma. The types of secondary cancers were glioma and meningioma, the two most common variants of secondary CNS cancer. To prepare these data for usage with algorithms, I used One Hot Encoding to convert the categorical values to numerical values.

The genetic data consisted of SNPs chosen based on their presence in certain genes that are currently suspected to be related to the development of secondary CNS tumors. For each SNP in a sample, I used additive coding to represent the genomic data as the number of times (1, 2, 3, etc.) the SNP was present in the patient's genome.

| Number of **Base Features** | Number of **Clinical Features** | Number of **Genetic Features** |
|---|---|---|
| 4 | 3 | 89 (SNPs) |

**Table 2.** Breakdown of the data features utilized.

Next, I began training and testing individual machine learning models. I chose to use 8 individual models, with 4 being tree-based (Random Forests, Extra Trees, AdaBoost, and XGBoost), and 4 not tree-based (K-Nearest Neighbors (KNN), Radial Basis Function Support Vector Machines (RBF SVM), Multilayer Perceptron Neural Networks (MLP Neural Networks), and Quadratic Discriminant Analysis (QDA)). Out of the tree-based models, Random Forests and Extra Trees were bagging-based, while AdaBoost and XGBoost were boosting-based.

I then trained and tested each model using 5-fold cross-validation (CV), with accuracy defined as the proportion of the number of correctly predicted values (*true positives* + *true negatives*) to the total number of values (shown in Equation 1). I first trained using the base and clinical features

alone, then added the genetic data to the subset to train using all features together. However, I found that for some of the models, adding the genetic data did not immediately increase accuracy (described further in the Results section). I hypothesized that this was most likely due to high noise in those data, and employed feature selection to reduce the noise.

I tried two different methods for feature selection: Gini impurity F-scores and permutation shuffling, and used the same four tree-based models from earlier: Random Forests, Extra Trees, AdaBoost, and XGBoost. I averaged the predictions of the models and extracted the top 10 most important features for use in retraining the individual machine learning models. I also conducted exploratory dataset analysis and calculated Pearson correlation coefficients for each of the features.

Upon retraining and retesting using these feature subsets, I found that utilizing the features selected using F-scores, which were calculated through the decrease in Gini impurity resulting from splitting on a feature, resulted in higher accuracy scores than those that I had achieved after training with all 96 data features. Through this feature selection process, I thus determined what subset to use in my next steps.

In order to further increase prediction accuracy, I constructed ensemble models using bagging, boosting, voting, and stacking. Ensemble learning aims to combine several machine learning models to decrease overfitting and increase overall prediction accuracy. To construct diverse model combinations, I computed Pearson correlations between the output predictions of each of the individual models and chose 7 low-correlation model combinations (using a threshold of $correlation < 0.2$), along with a combination consisting of all the individual models.

I utilized these 8 combinations in constructing soft voting ensembles and stacking ensembles. For each of the stacking ensembles, I tested 3 different meta-classifiers: logistic classification and two types of Random Forests, which resulted in 24 different combinations in total. In addition to this, I developed an adaptive stacking method which started with one model and only added models to the ensemble if they increased prediction accuracy. Along with voting and stacking, I also optimized the hyperparameters of the bagging methods (Random Forests and Extra Trees) and boosting methods (AdaBoost and XGBoost) through a grid search method.

After training and testing each of the ensemble models, I selected the highest prediction accuracy scores for each ensemble category and used t-tests to compare these scores with the highest of the individual models. Additionally, to further verify the improved quality of the ensemble models, I computed the AUC (Area Under the Curve) scores of ROC (Receiver Operating Characteristic) curves, defined as the area under the curve of the true positive rate, or *sensitivity*, plotted against the false positive rate, or *1-specificity* (shown in Equations 2 and 3), for the highest-performing model.

Then, to investigate the generalizability of the models to new data, I compiled another dataset by adding whole-exome sequencing data to the original genetic data. This new dataset thus consisted of 259 patients (data samples) and 1491 features. There were 64 cases (patients who had developed secondary cancer), and 195 controls (patients who had not developed secondary

cancer). I then repeated the entire procedure, from feature selection to ensemble model construction, and obtained a new selected subset of data features and predictive models with even more promising accuracy scores.

To analyze the biological significance of the genetic features that I had discovered through feature selection to be highly correlated with secondary cancer development, I utilized the NCBI dbSNP database to identify the genomic locations of the mutations and the oncological significance of the genes that they were located in.

$$accuracy = \frac{TP+TN}{total\ predictions} \qquad true\ pos.\ rate = \frac{TP}{TP+FN}$$

$$false\ pos.\ rate = \frac{FP}{FP+TN}$$

**Equations 1, 2, and 3.** Accuracy metrics employed to evaluate the predictive models.

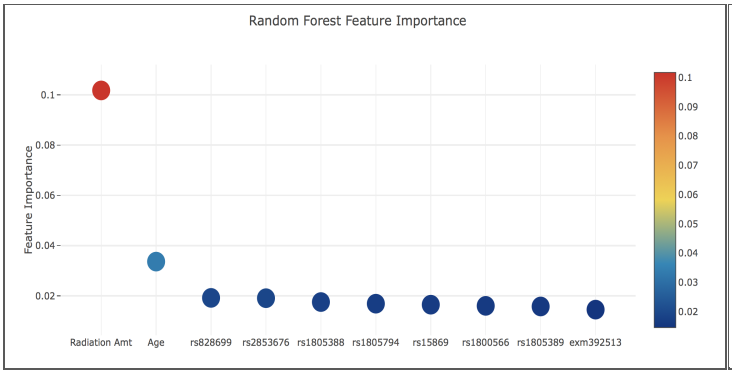
**Results (Data and Findings)**

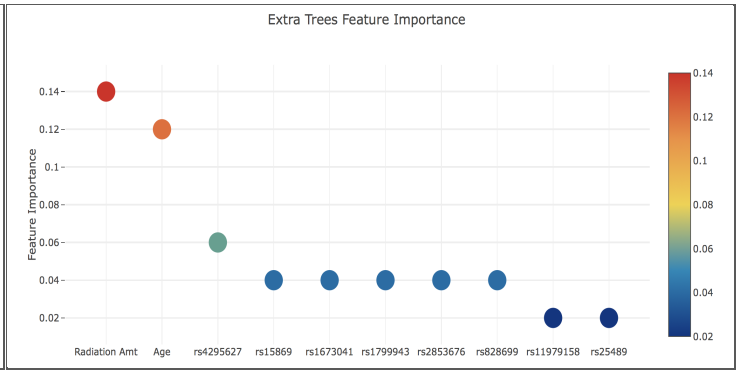
These are the accuracy scores for the individual models after training and testing with base and clinical features only and then all features together. For almost 40% of the models, adding the genetic data did not immediately increase model accuracy (most likely due to high noise). To remedy this, I performed feature selection to find a feature subset for training that would result in higher accuracies.

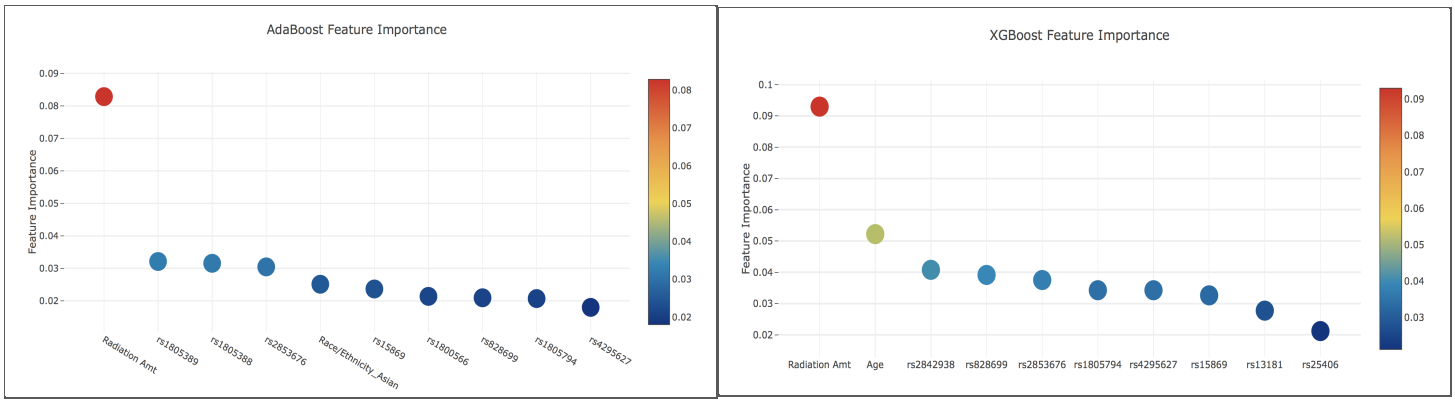
| Model | Score with Base and Clinical Features | Score with Base, Clinical, and Genetic Features | Improvement in score from Base/Clinical? |
|---|---|---|---|
| **Random Forests** | 0.72 | 0.73 | Yes |
| **Extra Trees** | 0.68 | 0.74 | Yes |
| **AdaBoost** | 0.71 | 0.71 | **No (same)** |
| **XGBoost** | 0.74 | 0.73 | **No (decrease)** |
| **KNN** | 0.71 | 0.74 | Yes |
| **RBF SVM** | 0.72 | 0.71 | **No (decrease)** |
| **Neural Networks** | 0.71 | 0.74 | Yes |
| **QDA** | 0.57 | 0.69 | Yes |

**Table 3.** The accuracy scores of the individual machine learning models before feature selection.

Below are the top 10 most important features (measured through Gini F-scores, as described in Materials, Methods, and Procedures) that were selected by the 4 tree-based models (Random Forests, Extra Trees, AdaBoost, and XGBoost).

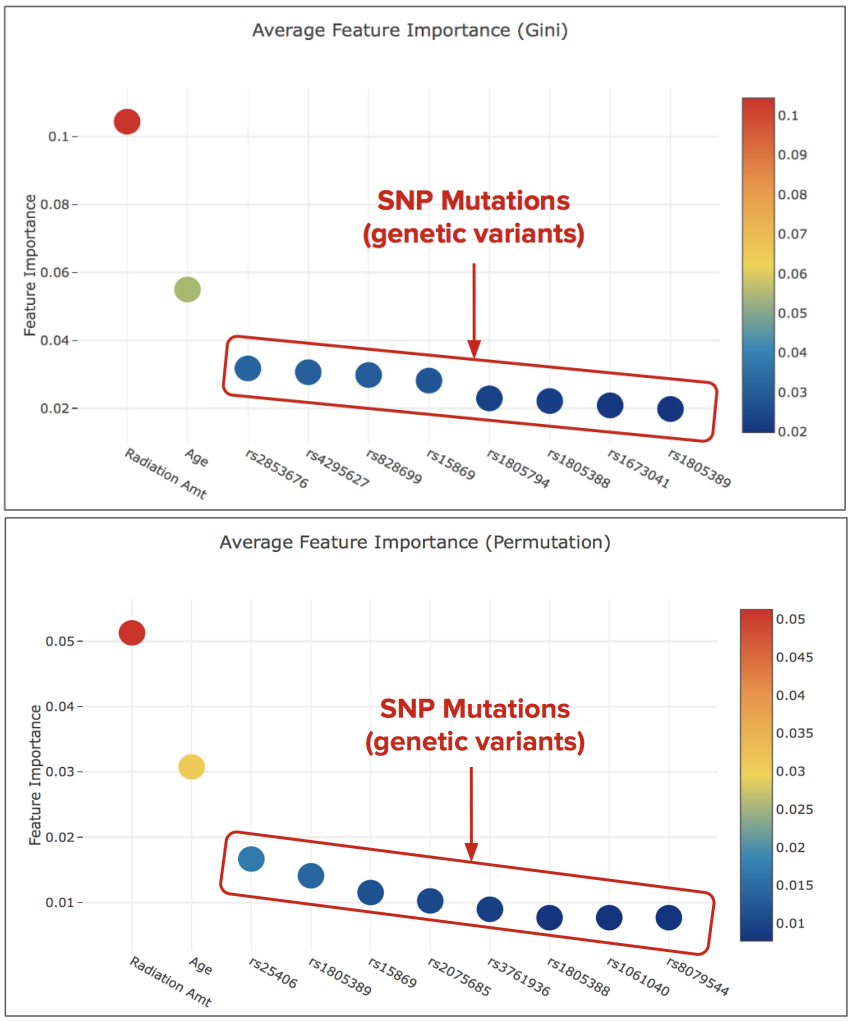





**Figures 3, 4, 5, and 6.** Top 10 most important data features, as selected by 4 tree-based models through Gini methodology.

Below are the top 10 features as calculated through both Gini impurity F-scores and permutation shuffling, obtained by averaging the predictions of the 4 tree-based models used for both methods. Radiation dosage amount and patient age, as well as 8 genetic variants, were found to be crucial to model prediction. (The significance of the genetic variants are elaborated upon further later in Results as well as Discussion and Conclusions.)




**Figures 7 and 8.** Top 10 most important features from the dataset (predictions from 4 models averaged), selected by Gini and permutation shuffling methods.

To the right are the Pearson correlation coefficients between each of the top 10 features selected through Gini F-scores. Pearson correlation indicates the relationship between two variables, or the extent to which data points follow a linear correlation, and are calculated by dividing the covariance by the product of the standard deviations. The largest correlations (most overlap between the features, and thus most relation) were between the SNPs rs1805388 and rs1805389, rs15869 and rs1805389, and rs285376 and rs1805388.

However, the overall relatively low correlations between the 10 features indicated that feature selection resulted in distinct and helpful features being chosen.

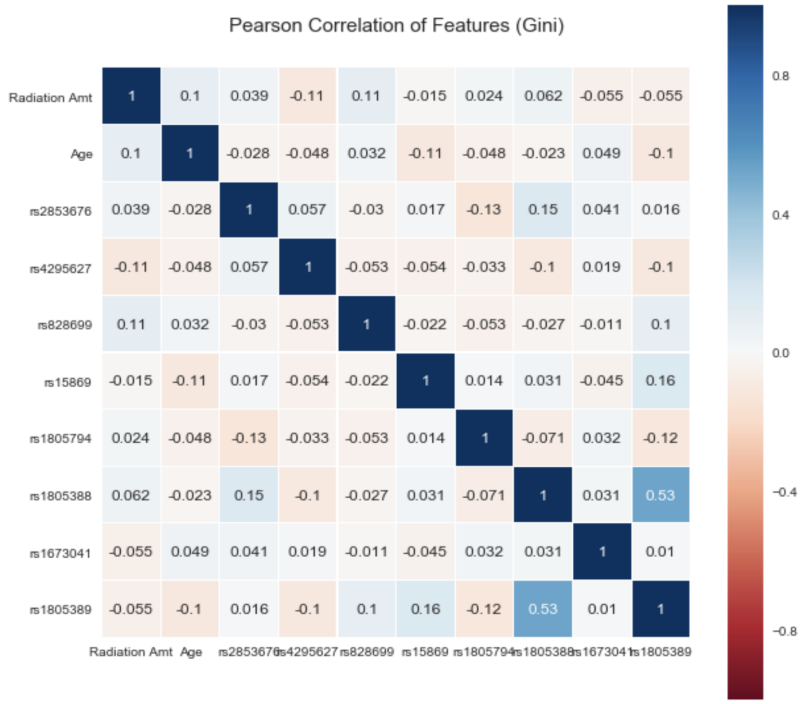


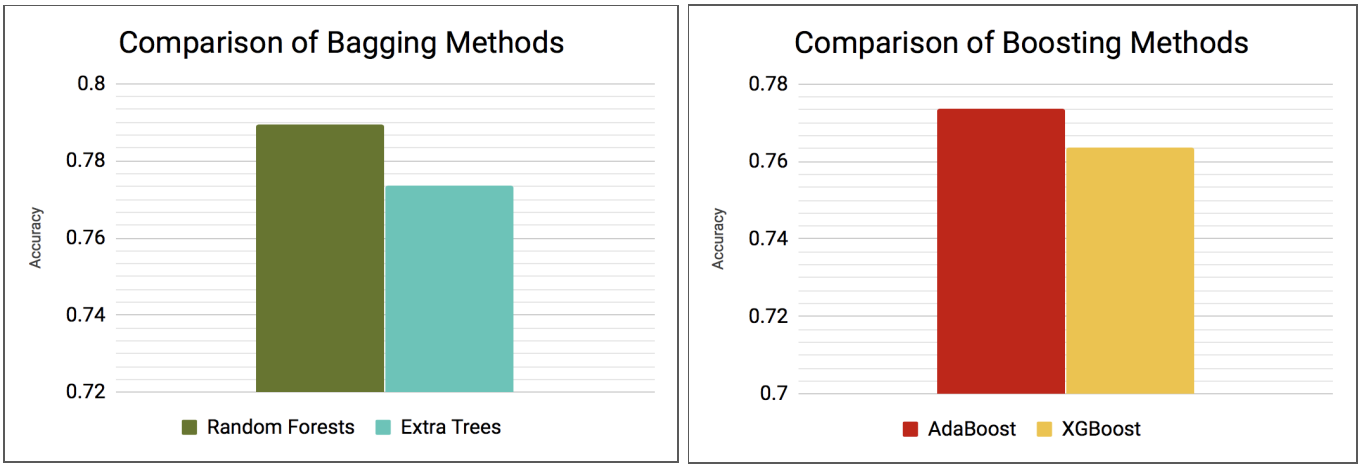

**Figure 9.** Pearson correlation for the top features.

Below are the accuracy scores for the individual models, this time after training and testing with the top 10 features chosen by both of the feature selection methods. Utilizing the features chosen by Gini selection resulted in significant improvements in accuracy (compared to the scores calculated before feature selection) for nearly 90% of the models, as shown by the bolded numbers.

| Model | Score with No Feature Selection | Score with 10 Features Selected (Permutation) | Score with 10 Features Selected (Gini) | Standard Deviation of Scores with 10 Features Selected (Permutation & Gini averaged) |
|---|---|---|---|---|
| **Random Forests** | 0.73 | 0.74 | **0.76** | 0.040 |
| **Extra Trees** | 0.74 | 0.71 | **0.77** | 0.050 |
| **AdaBoost** | 0.71 | 0.75 | **0.77** | 0.040 |
| **XGBoost** | 0.73 | 0.72 | **0.77** | 0.055 |
| **KNN** | 0.74 | 0.75 | **0.77** | 0.055 |
| **RBF SVM** | 0.71 | 0.75 | **0.76** | 0.030 |
| **Neural Networks** | **0.74** | 0.73 | 0.60 | 0.085 |
| **QDA** | 0.69 | 0.70 | **0.72** | 0.060 |

**Table 4.** Accuracy scores of individual models before and after conducting feature selection.

After training and testing the individual models, I increased accuracy further through ensemble techniques. Below are the accuracy scores obtained from training and testing the hyperparameter-optimized bagging and boosting methods. The Random Forests model was more accurate than the Extra Trees model, with an accuracy score of 0.790 compared to 0.774. The best hyperparameters (found through grid search) were 50 trees with a maximum depth of 6 for Random Forests, and 5 trees with no maximum depth for Extra Trees. On the other hand, the AdaBoost and XGBoost models performed comparably, with accuracy scores of 0.774 and 0.764, respectively. The best hyperparameters were 50 tree estimators for AdaBoost, and a maximum depth of 2 for XGBoost.




**Figures 10 and 11.** Comparison of the accuracy scores of the 4 hyperparameter-optimized bagging and boosting ensemble models.

Below are the accuracy scores of the 8 diverse (combinations chosen based on a threshold of *correlation* < *0.2*) soft voting ensemble models. For voting, KNN combined with QDA, Random Forests, Extra Trees, and XGBoost produced the highest accuracy score of 0.796.

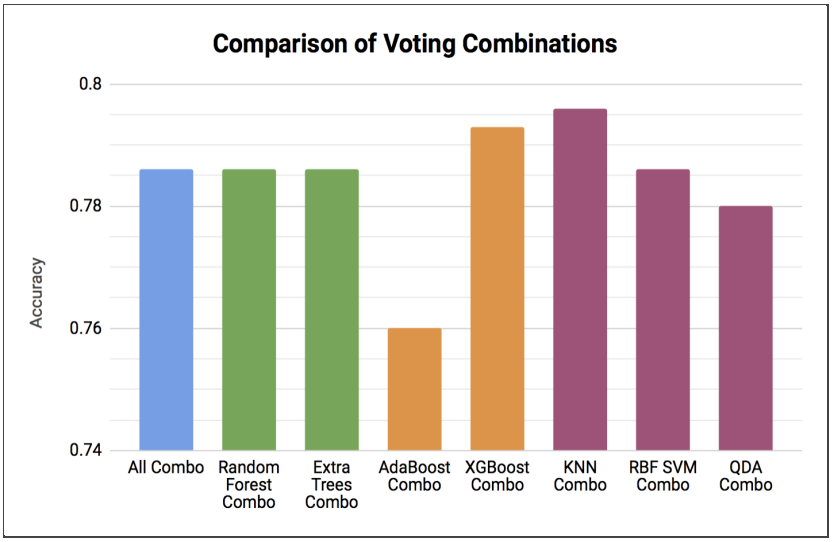



**Figure 12.** Comparison of the accuracy scores of the 8 voting ensemble models.

Below are the scores of the 8 diverse stacking ensemble models, as well as the adaptive stacking model. The adaptive model, which consisted of KNN with QDA and Random Forests (meta-classifier: Random Forests) produced the highest score of 0.793.

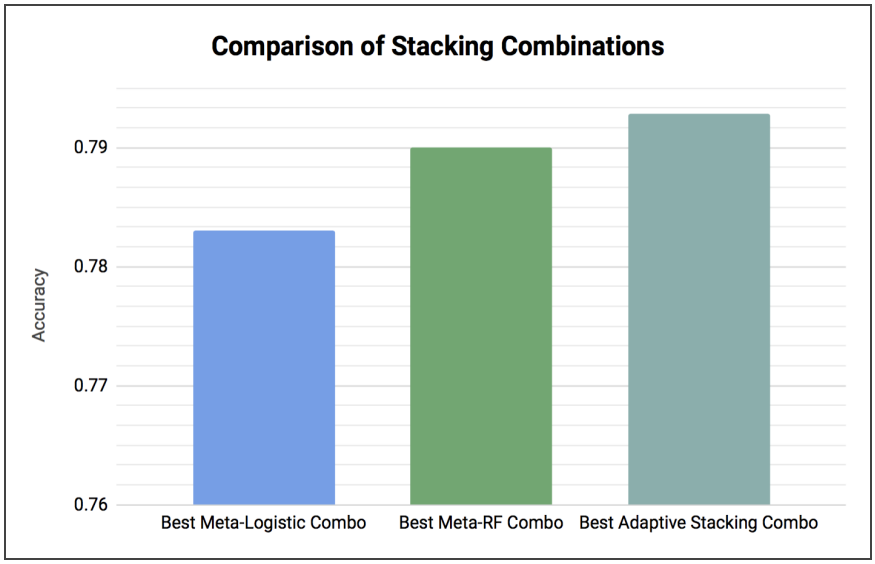



**Figure 13.** Comparison of the accuracy scores of the 3 best-performing stacking ensemble models (one with the Logistic meta-classifier and one with the Random Forests meta-classifier, as well as the adaptive stacking model).

Below is a comparison of the accuracy scores of the best individual model (Neural Networks) trained with all features with the scores of the best ensemble models (Random Forest bagging, AdaBoost boosting, KNN voting, and Random Forest stacking) trained with the 10-feature subset.

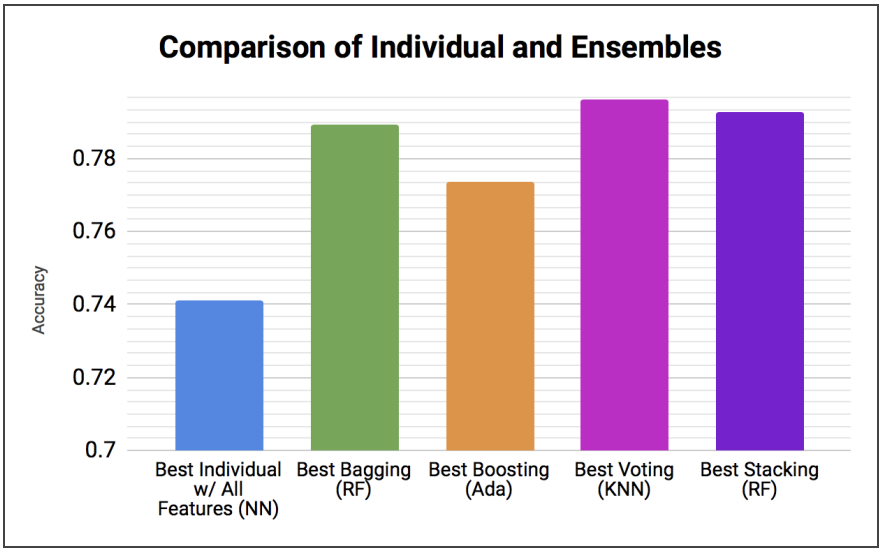



**Figure 14.** Comparison of accuracy scores of the best individual and best ensemble models.

Below are the p-values calculated from t-tests comparing the accuracy scores of the individual and ensemble models shown above. Optimized bagging, voting, and stacking were proven (using a threshold of *p-value < 0.050*) to result in statistically significant improvements in accuracy.

| | Bagging | Boosting | Voting | Stacking |
|---|---|---|---|---|
| **P-Value** (accuracy of ensembles vs. NN, the best indiv. model) | 0.010741 | 0.169670 | 0.034967 | 0.031059 |

**Table 5.** P-values of the t-test comparing the scores of the best ensemble models trained on selected features to those of the best individual model trained on all features.

In order to further verify the quality of the ensemble models, I calculated ROC AUC scores. The KNN voting model, which had produced the highest accuracy scores, had an ROC AUC score of 0.79 (shown below).

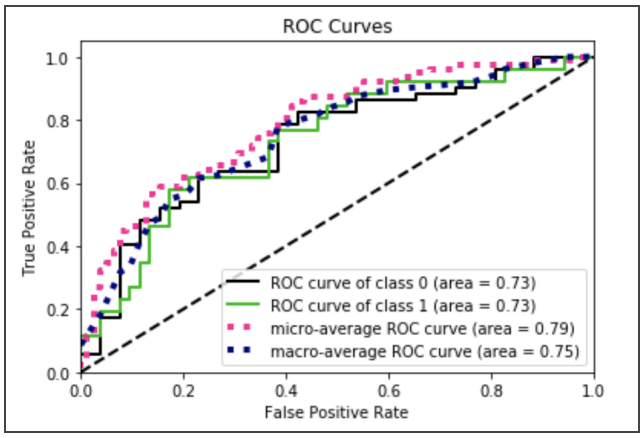
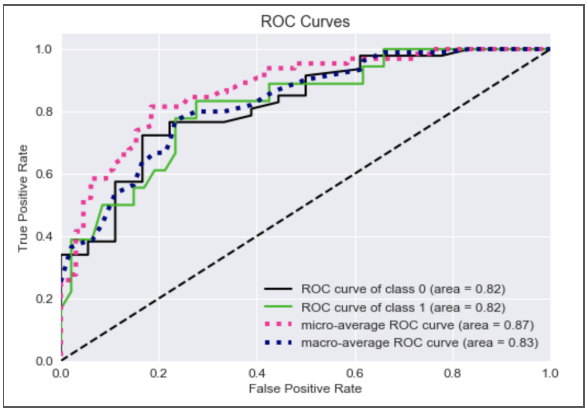
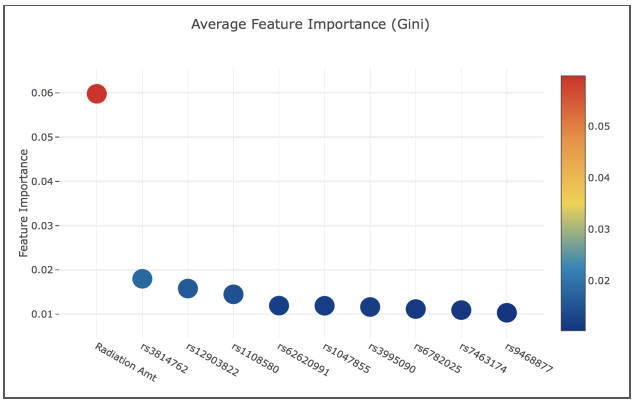



**Figure 15.** ROC curve of the highest-performing ensemble model.

Then, to demonstrate the generalizability of these models to new data, I constructed a new dataset by adding whole-exome sequencing data to the original genetic data, resulting in 1491 features and 259 samples. I then performed feature selection again and re-trained and tested the highest-performing individual and ensemble model types.

Out of the four highest-performing ensemble models, AdaBoost was the most accurate (ROC AUC score = 0.87, shown below). The 10 most important features in this dataset consisted of radiation amount as well as 9 new SNPs from genes such as MFRP, HTR4, and PRTG.




**Figures 16 and 17.** ROC curve of the highest-performing model trained on a 10-feature subset from the new dataset, and those top 10 features (selected using Gini method).

Below are the 8 genetic variants in the top 10 features (from the first dataset) that feature selection found to be crucial to model prediction. These SNPs have been previously validated through genetic studies to be correlated with secondary CNS cancer, which indicated that my models had accurately identified correlated SNPs (Barnes *et al*, 1998; Jensen, 2013; Lieber *et al*, 2003).

| **SNP** | rs2853676 | rs4295627 | rs828699 | rs15869 | rs1805794 | rs1805388 | rs1673041 | rs1805389 |
|---|---|---|---|---|---|---|---|---|
| **Gene** | TERT | CCDC26 | XRCC5 | BRCA2 | NBN | LIG4 | POLD1 | LIG4 |

**Table 6.** The 8 SNPs, from the first dataset, identified through feature selection to be highly correlated with secondary CNS cancer.

In the second dataset, I found 9 genetic variants to be features crucial to model training and prediction. This is the first time, to my knowledge, that these particular SNPs have been computationally demonstrated to be related to secondary CNS cancer development.

| **SNP** | rs3814762 | rs12903822 | rs1108580 | rs62620991 |
|---|---|---|---|---|
| **Gene/Chrom** | MFRP (11) | PRTG (15) | DBH (9) | LURAP1 (1) |
| **Function/results of mutation** | eye development/ retinal degeneration | abnormal tissue development | dopamine beta hydroxylase deficiency | muscle-eye-brain disease, muscular dystrophy |
| rs1047855 | rs3995090 | rs7463174 | rs6782025 | rs946877 |
| ENTPD3 (3) | HTR4 (5) | PLPBP (8) | STXBP5L (3) | CDH23 (10) |
| regulation of extracellular ATP | neurotransmitter release | tumor suppression/ epilepsy | neurodegenerative disorders | cadherin production/ breast cancer |

**Table 7.** The 9 SNPs, from the second dataset, identified through feature selection to be highly correlated with secondary CNS cancer.

**Discussion and Conclusions**

The ensemble learning models fulfilled my original project objectives: they were able to accurately predict secondary cancer development using genetic and clinical data, and data features crucial for prediction were identified. The addition of genetic data to base and clinical data improved model prediction accuracies after feature selection was performed, which indicated that patient genetics is indeed an important factor in secondary cancer development. Furthermore, the ensemble methods resulted in statistically significant improvements in ROC AUC and accuracy scores in comparison to the individual machine learning models, with KNN Voting being the highest-performing ensemble.

In addition, when those models were trained and tested on new genetic data, they performed well, with an excellent highest ROC AUC score of 0.87 from the AdaBoost ensemble model. This demonstrates that the models and feature selection methods can be generalizable to other datasets and secondary cancer types.

From the first and second datasets, my feature selection methods identified 8 and 9 specific SNP mutations as features crucial to accurate model prediction and thus correlated with secondary cancer development. I performed genetic analysis on each mutation by examining their genomic context (shown in Tables 6 and 7). The genes that the SNPs are located in, according to the NCBI dbSNP database, have been proven to be involved in a variety of cancer prevention-related functions, from tumor suppression to DNA repair, through clinical studies (Barnes *et al*, 1998; Jensen, 2013; Lieber *et al*, 2003). This showed that my models had accurately identified correlated SNPs. Additionally, my method resulted in the first instance, to my knowledge, of the 9 SNPs from the second dataset being computationally demonstrated to be related to secondary CNS cancer development.

By determining which SNPs are most correlated with secondary CNS cancer, drugs that target these specific genes and their protein products can be engineered in order to more effectively treat and halt the development of such cancer. Furthermore, testing for the presence of these specific genetic variants instead of employing whole genome sequencing procedures can dramatically decrease the amount of time and money needed to be spent during treatment recommendation procedures that determine whether a patient is susceptible to secondary cancer development, and thus whether radiation is a safe treatment option, or if they should consider alternative treatment such as chemotherapy or immunotherapy.

During my model development process, there are a few sources of error that may have affected my results. First, the datasets originally contained not a number (“NaN”) values, which I had to all set to 0 in order to prepare the data for model use. However, this may have led to decreased model accuracy as the altered values may have skewed the dataset. In addition, the case/control matching may not have been preserved during cross-validation. Due to the randomized nature of cross-validation, training may not have included both case patients who had developed secondary cancer along with matched control patients (who had similar base, clinical, and genetic features). This may have prevented the full model recognition of development versus non-development patterns.

In future research, I aim to utilize more types of genomic data, in addition to SNP mutations, to investigate other possible genetic correlations. For instance, I will use intra-protein changes as well as gene expression and transcription data that indicate atypical gene splicing and protein isoforms. Furthermore, weighting SNPs based on their frequency in the human genome can be explored. In addition, as more patient data emerge from clinical trials, computational prediction accuracy will further improve. To determine the mechanisms of secondary-cancer development, I also aim to perform additional genetic analysis on the SNPs that I identified as being correlated with secondary cancer.

Overall, the use of these machine learning methods can greatly improve and speed up crucial cancer treatment recommendation procedures. They can provide an initial stage that can indicate which genetic characteristics patients should be tested for in order to determine which treatment option is best for them. Thus, the ability to input different clinical and genomic features and generate predictions of treatment risk has applications in precision medicine, where personalized regimens serve individual patients.

**References**

Barnes, D. E., Stamp, G., Rosewell, I., Denzel, A., & Lindahl, T. (1998). Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Current Biology*, *8*(25), 1395–1398. doi: 10.1016/s0960-9822(98)00021-9

Donin, N., Filson, C., Drakaki, A., Tan, H. J., Castillo, A., Kwan, L., … Chamie, K. (2016). Risk of second primary malignancies among cancer survivors in the United States, 1992 through 2008. *Cancer*, *122*(19), 3075–3086. doi:10.1002/cncr.30164

Jensen, R. B. (2013). BRCA2: one small step for DNA repair, one giant protein purified. *The Yale journal of biology and medicine*, *86*(4), 479–489.

Lieber, M. R., Ma, Y., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nature Reviews Molecular Cell Biology*, *4*(9), 712–720. doi: 10.1038/nrm1202

Morris, E. B., Gajjar, A., Okuma, J. O., Yasui, Y., Wallace, D., Kun, L. E., … Hudson, M. M. (2007). Survival and Late Mortality in Long-Term Survivors of Pediatric CNS Tumors. *Journal of Clinical Oncology*, *25*(12), 1532–1538. doi: 10.1200/jco.2006.09.8194

Zhao, P., Zou, P., Zhao, L., Yan, W., Kang, C., Jiang, T., & You, Y. (2013). Genetic polymorphisms of DNA double-strand break repair pathway genes and glioma susceptibility. *BMC cancer*, *13*, 234. doi:10.1186/1471-2407-13-234

**Statement on Outside Assistance**

| |
|---|
| **Name:** Julia Camacho |
| **Title of Paper:** Predicting the Development of Secondary Central Nervous System Cancer Through Ensemble Learning Methods |
| **What steps led you to formulate your hypothesis? (Where did you get the idea for your research?) Please be specific.**<br>I have been conducting research in computational biology for several years, and recently became interested in computational oncology. My research professor, Dr. Xuexia Wang, introduced me to the fact that one of the key problems that cancer patients currently face is the development of secondary cancer, and I sought to develop a method using principles of precision medicine to predict these occurrences. |
| **Where did you conduct the major part of your work? (i.e. home, school, or other institutional setting, university lab, medical center, etc.)**<br>I conducted my work at home and at the University of North Texas Department of Mathematics. |
| **If you worked in an institutional setting, did you work on your project as part of a team or group? If so, how large was the team and who was on the team (students, adult researchers, etc.)? Describe your role on the team.**<br>I worked directly with my research professor (not in a group), and led the project development. |
| **Describe what parts of the research you did on your own and what parts you received help (i.e. literature search, hypothesis, experimental design, use of special equipment, gathering data, evaluation of data, statistical analysis, conclusions, and preparation of written report (abstract and/or paper).**<br>I worked with my research professor in developing my main research goal and in making sure that my experimental design was not flawed. I conducted a literature search and read all background papers on my own, and conducted all data preparation and evaluation, model construction, and statistical analysis of results on my own. I discussed my interpretation of my results with my research professor to make sure my analysis was reasonable. I wrote my abstract and paper myself, and my father helped proofread. |
| **Is your research current or a continuation of previous research? If a continuation, please describe the current work and advancement(s) of this research in comparison to the prior work and results.**<br>My research has ended and was not a continuation of my previous research. |