Kari Palmier
DSC 540 Winter 2019
Final Project Report

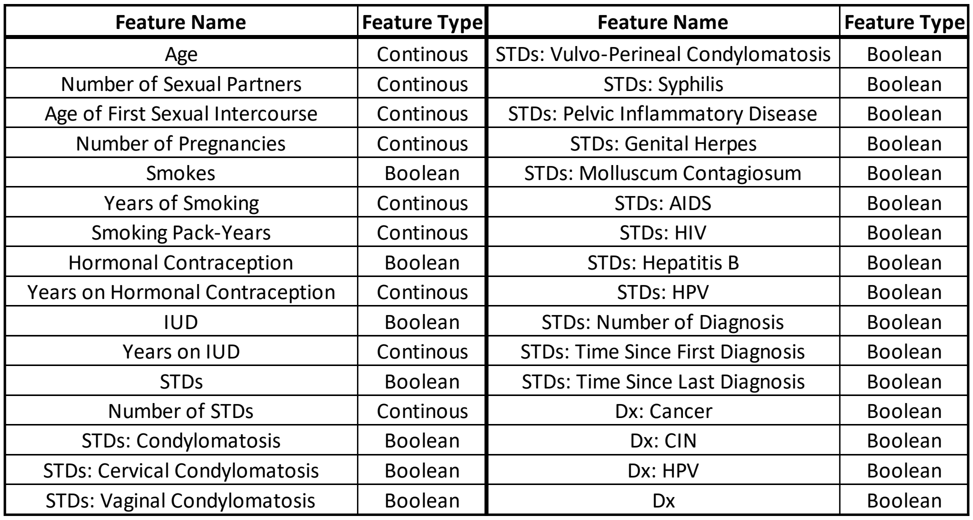
Cervical Cancer Prediction

Abstract

Cervical cancer is a disease that affects women of all age groups in the United States. The number of women diagnosed with cervical cancer has decreased over the past few decades due to better screening practices which have led to detection before cancer has formed. Unfortunately, screenings alone are not sufficient to catch all instances of cervical cancer. The diagnosis of cervical cancer is an issue due to the lack of physical symptoms until the disease reaches late stage. This paper explores using feature selection and classification methods in the prediction of cervical cancer based on demographic and health risk factors. The results of random forest, gradient boost, neural network, and support vector machine (SVM) classification models were compared to determine which had the best performance. Low variance filter, model wrapper, stepwise recursive, chi-squared univariate, and mutual information feature selection were performed on each model type in order to reduce the feature dimensionality and create more parsimonious models. All models and feature selection methods were run on the biopsy target variable and an aggregated combination target. Final results show that the random forest model performed the best for both targets.

Introduction

Cervical cancer is a difficult disease to diagnose in early stages because there are no discernable symptoms at this point. The disease is not able to be diagnosed until it reaches late stage, at which point medical options are limited. Regular check-ups are currently the only way to detect cervical cancer before it progresses too far. It is not always possible for people to have the necessary testing, especially in low income countries. This paper attempts to generate models that will determine the presence of cervical cancer based on a number of different risk factors. The dataset used was from 2017 from the Hospital Universitario de Caracas in Caracas, Venezuela. It was obtained in the UCI Machine Learning Repository. The hospital collected health and demographic information on 858 patients in a study concerning possible risk factors for cervical cancer. This dataset contains 32 risk factors and 4 different targets. Each target represents the result of a given medical test that detects cervical cancer. The four targets are Hinselmann, Schiller, cytology, and biopsy. Hinselmann refers to a colposcopy using acetic acid. Schillers, cytology and biopsy refer to colposcopy using Lugol iodine. The target values were 1 indicating detection and 0 indicating no detection. The 32 risk factors included age, sexual history, smoking history, IUD and hormonal contraception usage, STD history, previous cancer diagnosis, HPV diagnosis, CIN diagnosis, and a generic diagnosis flag. Table 1 lists all of the dataset features and their data types.



Feature Name	Feature Type	Feature Name	Feature Type
Age	Continous	STDs: Vulvo-Perineal Condylomatosis	Boolean
Number of Sexual Partners	Continous	STDs: Syphilis	Boolean
Age of First Sexual Intercourse	Continous	STDs: Pelvic Inflammatory Disease	Boolean
Number of Pregnancies	Continous	STDs: Genital Herpes	Boolean
Smokes	Boolean	STDs: Molluscum Contagiosum	Boolean
Years of Smoking	Continous	STDs: AIDS	Boolean
Smoking Pack-Years	Continous	STDs: HIV	Boolean
Hormonal Contraception	Boolean	STDs: Hepatitis B	Boolean
Years on Hormonal Contraception	Continous	STDs: HPV	Boolean
IUD	Boolean	STDs: Number of Diagnosis	Boolean
Years on IUD	Continous	STDs: Time Since First Diagnosis	Boolean
STDs	Boolean	STDs: Time Since Last Diagnosis	Boolean
Number of STDs	Continous	Dx: Cancer	Boolean
STDs: Condylomatosis	Boolean	Dx: CIN	Boolean
STDs: Cervical Condylomatosis	Boolean	Dx: HPV	Boolean
STDs: Vaginal Condylomatosis	Boolean	Dx	Boolean

Table 1. Risk Factor Features

Literature Review

Two scientific journal articles were found that used the same UCI dataset in attempt to create classification models for cervical cancer diagnosis. The first was by Wu and Zhou (1) and was on the topic of SVM classification for cervical cancer diagnosis. This article utilized basic SVM, SVM-RFE, and SVM-PCA classification models to predict all four of the dataset target variables. All missing values were removed from the dataset and the two features with mostly missing values were removed. This resulted in 30 features and 668 patients (rows). Oversampling was performed, but the type was not given. Feature selection was not performed on the basic SVM model. The SVM-RFE and SVM-PCA models using varying numbers of features (5 and 15 for SVM-RFE and 5 and 11 for SVM-PCA). The performance metrics used in this article are accuracy, sensitivity, specificity, negative prediction accuracy, and positive prediction accuracy (also called precision). This paper found that the SVM-PCA performed the best, although not by much. Most models had similar performance metrics. No data splitting was mentioned so it is assumed that all results are just from 5-fold cross-validation on the training dataset that was mentioned. This means the models may not perform well against real-world imbalanced data. This article was interesting because dealt with different types of SVM models, SVM-RFE and SVM-PCA. This article showed that SVM models may be useful in dealing with class imbalance. SVM models were added to the types of models used in the UCI cervical cancer dataset.

The second article dealing with the UCI dataset is by Unlersen, Sabanci, and Ozcan (2). This article used multilayer perceptron (MLP), BayesNet, and k-nearest neighbors (kNN) to attempt to classify cervical cancer. It only focused on one of the targets, the biopsy target. This article specifies that the data was split into training and testing, with 66% training and 33% testing. Only the training dataset was used during modeling. This paper focused on the confusion matrix for its performance statistics. The performance metrics used in model evaluation were total true classified instance (defined as true negative class + true positive class), true negative class, true positive class, false negative class, and false positive class. The article also referred to correctly classified instance percentages and false classified instance rates (also a percentage). Several values of k for kNN modeling were used, as well as several sizes of hidden layers in the MLP model. Only one BayesNet model was executed. The results discussion of this paper revolved around classification accuracy (what the article calls correctly classified instance percentage). The article used this metric along with the false negative rate (number of false negatives divided by total number of instances) to conclude that kNN was the best method. There is no mention of sampling or feature selection in the article. Because the results focused solely on the number of instances correctly classified (accuracy) as well as the false negative rate (which is also a metric that is over all samples), the results may be misleading. No metric was used to determine how the classification performed in reference to target class values (minority or majority). An imbalanced dataset will have good accuracy (and false negative rate as described here) as long as most of the majority class is correctly predicted. The classification of the minority class isn’t captured in accuracy or false negative rate. This article was useful because it covered several different types of models (BayesNet, kNN, and MLP) and how each handled class imbalance. Based on the conclusions, it appears that kNN could be a good model to try. There were issues with the performance methods used, however, so the conclusions may not be valid.

The next article reviewed was an overview of imbalanced data learning approaches by Bekkar and Alitouche (3). This article gave interesting ideas on how to approach machine learning with imbalanced data. Different types of sampling, feature selection, and ensemble methods were discussed. The article stated that the results of previous studies found that undersampling lead to better results, while oversampling produces little or no change in performance (3). The article also stated that the disadvantage to undersampling is that it may exclude potentially useful information, which could be important for the model training process (3). It said the disadvantage to oversampling is that the addition of formal copies of instances can lead to a situation of overfitting (3). It also discusses how overfitting does not introduce new data, so it does not present a solution to the fundamental issue of lack of data (3). The article proposes using weighted sampling or more intelligent sampling strategies such as Tomek Link, SMOTE, One-Sided Selection OSS, Neighborhood Cleaning Rule NCL, or Bootstrap-based Oversampling BootOS. This article proposes the usage of cost sensitive learning to model imbalanced datasets. This article had several interesting ideas on how to improve imbalanced learning. For decision trees, these ideas included using a new splitting criterion that is more sensitive to the class imbalance, adjusting the distribution reference in the tree, and using offset entropy (considering the prior distribution of classes in the partitioning criteria). This article was very helpful in understanding how to approach modeling with imbalanced datasets. The article exposed issues with traditional random under and over sampling methods and mentioned several different types of sampling that could be better. Based on this paper, SMOTE oversampling was selected for usage on the UCI cervical cancer dataset. More investigation is needed into the other types of sampling mentioned.

The fourth article by Mazurowski, Habas, Zurada, Lo, Baker, and Tourassi involved using neural network to train imbalanced datasets for medical decision making (4). This article investigated the usage of classic backpropagation and particle swarm optimization algorithms in neural networks. This study used simulated data and a real-world breast cancer dataset. The performance metric used in evaluation was area under the ROC curve (AUC) and partial AUC or pAUC where p indicates the lowest acceptable sensitivity level. This article evaluated the use of both undersampling and oversampling. This study divided both the simulated datasets and the breast cancer dataset into a training dataset and a validation dataset. Both the training and validation datasets were 50% of the original data and were split using a stratified method. Feature selection of 5 and 10 features was performed using simple forward selection based on linear discriminant performance (4). Modeling was done with each number of features and with all features. This article concluded that increasing class imbalance in the training dataset generally has a progressively detrimental effect on the classifier's test performance measured by AUC and pAUC (4). It also concluded that although undersampling was typically an inferior choice to compensate for class imbalance, there is no clear winner between oversampling and no compensation (4). The modeling process in this article was very helpful in determining how to approach imbalanced data. The usage of over and under sampling was interesting as was the finding that undersampling performed worse and that oversampling and no sampling yielded similar results. Based on this, both under and over sampling techniques were applied to the UCI cervical cancer dataset. Neural networks were also added to the list of models used on the UCI cervical cancer dataset.

The final article researched was on predicting diseases from highly imbalanced data using random forests by Khalilia, Chakraborty, and Popescu (5). This article used the National Inpatient Sample data available through the Healthcare Cost and Utilization Project. The study compared the performance of SVM, bagging, boosting, and random forest to predict the risk of eight chronic diseases (5). The performance metric used was AUC. Feature selection was performed manually. Each record had 262 features. For every record, the age, race, sex, and 15 diagnosis categories were extracted (5). To handle the class imbalance issue, repeated random sub-sampling was used. There is no mention of splitting the data in to training and validation datasets. Using the entire dataset to train could cause issues with model performance results being unrealistically high. In order to approach this issue, models were also run without sampling and the performance with and without sampling were compared. This article concluded that the random forest classification method performed the best in terms of AUC for both the no sampling and sampling cases. It also concludes that use of repeated random sub-sampling is beneficial with highly imbalanced data. This article was very helpful in determining what models to choose to handle class imbalance. Based on this article, boosting and random forest models were added to the types of models that were applied to the UCI cervical cancer dataset.

Methodology

The UCI cervical cancer dataset contained several missing values. There were two risk factor features that were almost entirely made up of missing values. These two features STD time since first diagnosis and STD time since last diagnosis. These two features were removed from the dataset. Most of the remaining features still contained missing values. In the case of categorical features, the rows that contained missing values were removed, since there is no clear way to replace a categorical variable (there is no mean or median of a category). The missing values in each continuous feature were replaced with the mean value of the feature. Once data cleaning was done, the dataset contained 726 rows and 4 target variables. The target variables were separated from the dataset dataframe into standalone variables (each target was a separate variable). Large class imbalance was present in all of the target variables. In an effort to alleviate this imbalance, a combination target was generated from the four original ones. This combination target contained a value of 1 for any row that had a value of 1 in any of the four original targets, and a value of 0 otherwise. It represents if any of the target tests detected cervical cancer. The class imbalance of this combination target was less than the individual ones, but was still large enough to cause issues. Once the targets were separated from the feature dataframe, the targets and data were split to generate training and validation datasets. Note that different training and validation splits were created per target (each target had its own training and validation datasets). The same feature dataframe was used in the generation of each of the target splits. Each split was done such that the validation dataset was 20% of the original data size and the training was 80%. All splits were stratified, ensuring that the same distribution of target class values was present in both the training and validation datasets. The validation datasets were not used during any step of the modeling process. They acted as new real-world datasets that were only used to assess the performance of the models. The original biopsy target and the new combination target were selected for use in modeling. The three remaining original targets were not used. The biopsy target was chosen because it had the largest number of 1 values of the original targets (was the least imbalanced). After splitting the data, the biopsy training target had 540 values of 0 and 40 values of 1. The combination training target had 506 values of 0 and 74 values of 1.

Models were generated during several stages of the classification process. Four types of models were used during all stages, random forest, gradient boost, neural network, and SVM. All four models were used at each stage to investigate the difference in performance between them during the given stage. A set of base models were used during the sampling and feature selection stages. The base models during these parts all used the same parameters for each model type. Next, grid search optimization and final model evaluation were performed. The primary performance metric used to evaluate all models was AUC (area under the ROC curve). This was chosen due to the class imbalance present in both targets. Accuracy values were reported but not used in model evaluation because they were found to be misleading (since majority class was significantly larger than the minority, accuracy was high even though the minority class was largely misclassified). In the remaining portion of this paper, model performance discussions refer to the AUC performance unless specified otherwise.

When each model was generated during all stages of the modeling process, two different types of evaluation were performed. The first was a 5-fold cross-validation on the training dataset. The second was generating a model with the entire training dataset, then scoring the model performance against the validation dataset (the real-world dataset). All model results contain the AUC and accuracy values reported by both the training cross-validation and validation dataset models, as well as cross-validation standard deviation values. The validation dataset AUC was used as the main performance metric, and the training cross-validation AUC with its standard deviation were used a secondary one.

Because of the large class imbalance present in the biopsy and combination targets, sampling was applied to each training dataset. SMOTE oversampling from the Python Imbalance-Learn (imblearn) package was applied with several different class sampling strategy values to determine the optimal value. SMOTE oversampling was applied to the biopsy dataset (training features and target) and combination dataset separately. Once each dataset was SMOTE oversampled, the new dataset was then used to generate base random forest, gradient boost, neural network, and SVM models. The sampling strategy value was changed until the performance of the base models showed improvement. Sampling strategy values of 1.0, 0.8, 0.6, 0.5, 0.4, 0.3, and 0.1 were initially used. It was found that there was not a discernable difference in performance from 1.0 to 0.5, nor from 0.4 to 0.1. The values of 1.0 and 0.3 were chosen for usage in the feature selection steps. The value of 1.0 resulted in same the minority and majority class sizes in the output data (540 values of 0 and 540 values of 1 for the biopsy target and 506 values of 0 and 506 values of 1 for the combination target). The value of 0.3 resulted in 540 values of 0 and 162 values of 1 for the biopsy target and 506 value of 0 and 151 values of 1 for the combination target.

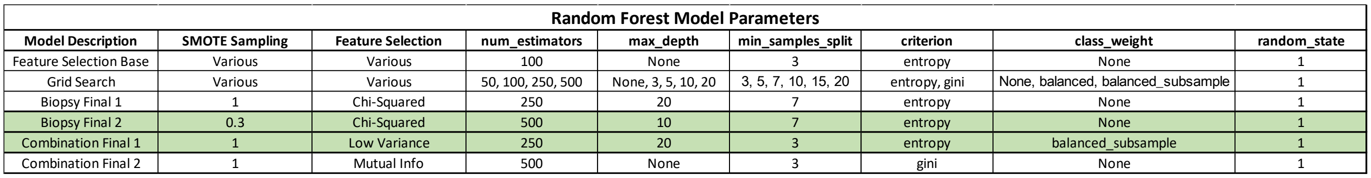
Random undersampling from the Python Imbalance-Learn package was also performed on each the same set of base models. This was also performed separately on the biopsy and combination target datasets. Sampling strategy values of 1.0, 0.3, and 0.1 were used for the biopsy target and values of 1.0, 0.3, and 0.2 were used for the combination target. It was found that random undersampling resulted in poor training cross-validation performance on all model types. This is because the number of samples used was based off of the size of the minority class of the target (minority class is the value 1 in the datasets). In the case of the biopsy target, this is 40 and for the combination target this is 74. An undersampling sampling strategy value of 1.0 resulted in the minority class and majority class both having the same number of samples, 40 for the biopsy target and 74 for the combination target. This is much lower than the total number of samples of each dataset (580 total samples in each). Even the sampling strategy value of 0.1 resulted in a biopsy dataset that was to small (biopsy target had 400 values of 0 and 40 values of 1. Likewise the sampling strategy value of 0.2 resulted in a combination dataset that was also too small (combination target had 370 values of 0 and 74 values of 1). Since the training cross-validation performance was low for all models using random undersampling, this method was not used in future modeling.

The next stage in the modeling process was applying feature selection to the biopsy and combination training datasets. Feature selection was performed for each target dataset separately. It was performed on each of the four model types separately for each target (random forest, gradient boost, neural network, and SVM). This was done so the difference in performance and selected features could be analyzed across model types. Three different types of training dataset sampling were applied prior to feature selection. They were no sampling (the original training dataset), SMOTE oversampling with a sampling strategy of 1.0, and SMOTE oversampling with a sampling strategy of 0.3. This means that all types of features selection were performed for each model type and each sampling value for each target dataset (biopsy and combination). Note feature selection was run twice in order to gauge the performance repeatability of each type.

Five different types of feature selection were applied to the random forest and gradient boost models. These were low variance filter, model wrapper, stepwise recursive, chi-squared univariate, and mutual information. Three of those feature selection types were also applied to the neural network and SVM models (low variance filter, chi-squared univariate, and mutual information). Model wrapper and stepwise recursive depended on a model output that is not present in the case of neural network or SVM models (the get_support method). The low variance filter method used the Python Scikit-Learn VarianceThreshold function with a variety of thresholds to perform feature selection. Models were generated for each threshold value using only the selected features at that threshold value. Training cross-validation AUC values were recorded for each model and the model with the highest 5-fold cross-validation AUC was selected for future use. The selected features that corresponded to this best performing model were returned, as was the training dataset with only the selected features present. The model wrapper method used the Python Scikit-Learn SelectFromModel function with a mean threshold to perform the feature selection and the get_support method. It returned the selected features as well as a training feature dataset with only the selected features present. The stepwise method used the Python Scikit-Learn RFE function with various k values to perform feature selection and the get_support method to return the selected features. Models were generated for each k value using only the selected features at that k value. Similar to the low variance filter method, the model with the highest 5-fold cross-validation AUC was selected for future use and its selected features and the dataset containing only those features were returned. The chi-squared univariate method used the Python Scikit-Learn SelectKBest function with the chi2 scoring function and a variety of k values to perform feature selection. As was done in stepwise recursive, models were generated for each k value using only the selected features at that k value. Similar to the low variance filter method, the model with the highest 5-fold cross-validation AUC was selected for future use and its selected features and the dataset containing only those features were returned. The mutual information method was identical to the chi-squared method except that it used the mutual_info_classif scoring method in the SelectKBest function.

The best performing two feature selection models were chosen for each model and target type. The training data from each of these feature selection models was then used in a grid search. This grid search performed models for several sets of model parameters and returned the parameter set that resulted in the best (highest) cross-validation AUC value. The Python Sckikit-Learn GridSearchCV function was used to perform the grid searches. Once the optimal parameter set was found for each feature selection model, it was used to generate a 5-fold cross-validation model on the corresponding training dataset as well as to generate an overall model from the entire training dataset. This overall model was then used to score the performance of a copy of the validation dataset (which had the features removed appropriately to match the training dataset from feature selection). This process (grid search, training cross-validation, and overall model training and validation dataset scoring) was done for each of the two selected models from feature selection for each model type and each target dataset (biopsy and combination datasets). The final model was chosen for each target dataset (biopsy and combination) based on the validation dataset AUC. Since the performance of some models were close, the training dataset cross-validation AUC and standard deviation, number of features in the final training dataset, and the model interpretability were used to narrow the selection down to the final model.

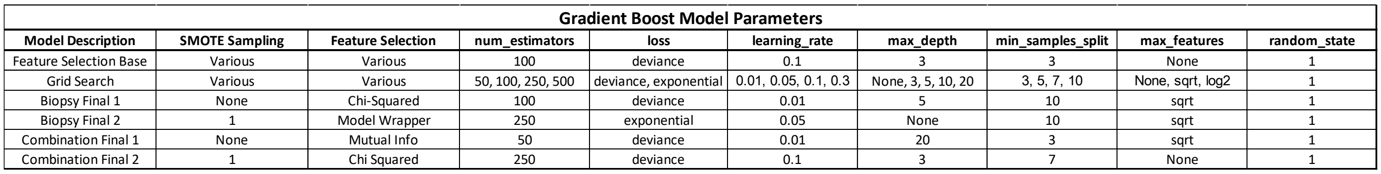
The parameters used for each set of models are given in tables 2, 3, 4, and 5 (table 2 has the random forest parameters, table 3 has the gradient boost parameters, table 4 has the neural network parameters, and table 5 has the SVM parameters). The models labeled Final 1 and Final 2 are the two models that performed the best from feature selection for each model type and target. The models highlighted in green correspond to the biopsy and combination target final models chosen. Note that the random state was set to a constant value during all modeling. This is to ensure that the model results were repeatable between runs. The random state was also set during the testing and validation splitting to ensure the splits remained the same as well to ensure the same splitting was done every time the code was executed. Repeatability between runs was important because the code was run several times during the development process. In order to be able to compare the results between all of the models generated, they had be repeatable between runs.



Random Forest Model Parameters

Model Description	SMOTE Sampling	Feature Selection	num_estimators	max_depth	min_samples_split	criterion	class_weight	random_state
Feature Selection Base	Various	Various	100	None	3	entropy	None	1
Grid Search	Various	Various	50, 100, 250, 500	None, 3, 5, 10, 20	3, 5, 7, 10, 15, 20	entropy, gini	None, balanced, balanced_subsample	1
Biopsy Final 1	1	Chi-Squared	250	20	7	entropy	None	1
Biopsy Final 2	0.3	Chi-Squared	500	10	7	entropy	None	1
Combination Final 1	1	Low Variance	250	20	3	entropy	balanced_subsample	1
Combination Final 2	1	Mutual Info	500	None	3	gini	None	1

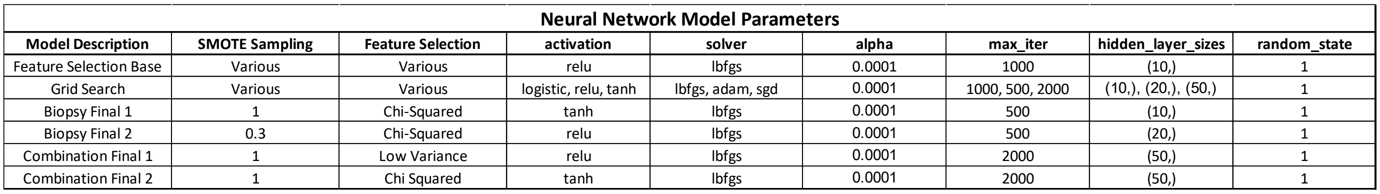
Table 2. Random Forest Model Parameters



Gradient Boost Model Parameters

Model Description	SMOTE Sampling	Feature Selection	num_estimators	loss	learning_rate	max_depth	min_samples_split	max_features	random_state
Feature Selection Base	Various	Various	100	deviance	0.1	3	3	None	1
Grid Search	Various	Various	50, 100, 250, 500	deviance, exponential	0.01, 0.05, 0.1, 0.3	None, 3, 5, 10, 20	3, 5, 7, 10	None, sqrt, log2	1
Biopsy Final 1	None	Chi-Squared	100	deviance	0.01	5	10	sqrt	1
Biopsy Final 2	1	Model Wrapper	250	exponential	0.05	None	10	sqrt	1
Combination Final 1	None	Mutual Info	50	deviance	0.01	20	3	sqrt	1
Combination Final 2	1	Chi Squared	250	deviance	0.1	3	7	None	1

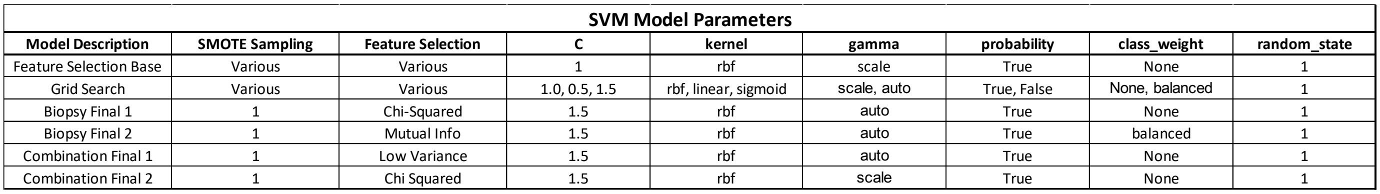
Table 3. Gradient Boost Model Parameters



Neural Network Model Parameters

Model Description	SMOTE Sampling	Feature Selection	activation	solver	alpha	max_iter	hidden_layer_sizes	random_state
Feature Selection Base	Various	Various	relu	lbfgs	0.0001	1000	(10,)	1
Grid Search	Various	Various	logistic, relu, tanh	lbfgs, adam, sgd	0.0001	1000, 500, 2000	(10,), (20,), (50,)	1
Biopsy Final 1	1	Chi-Squared	tanh	lbfgs	0.0001	500	(10,)	1
Biopsy Final 2	0.3	Chi-Squared	relu	lbfgs	0.0001	500	(20,)	1
Combination Final 1	1	Low Variance	relu	lbfgs	0.0001	2000	(50,)	1
Combination Final 2	1	Chi Squared	tanh	lbfgs	0.0001	2000	(50,)	1

Table 4. Neural Network Model Parameters



SVM Model Parameters

Model Description	SMOTE Sampling	Feature Selection	C	kernel	gamma	probability	class_weight	random_state
Feature Selection Base	Various	Various	1	rbf	scale	True	None	1
Grid Search	Various	Various	1.0, 0.5, 1.5	rbf, linear, sigmoid	scale, auto	True, False	None, balanced	1
Biopsy Final 1	1	Chi-Squared	1.5	rbf	auto	True	None	1
Biopsy Final 2	1	Mutual Info	1.5	rbf	auto	True	balanced	1
Combination Final 1	1	Low Variance	1.5	rbf	auto	True	None	1
Combination Final 2	1	Chi Squared	1.5	rbf	scale	True	None	1

Table 5. SVM Model Parameters

Results

The results for the features selection models of the biopsy target are given in tables 6, 7, 8, and 9 (table 6 is for random forest, table 7 is for gradient boost, table 8 is for neural network, and table 9 is for SVM). The results for the features selection models of the combination target are given in tables 10, 11, 12, and 13 (table 10 is for random forest, table 11 is for gradient boost, table 12 is for neural network, and table 13 is for SVM). The two models selected from each model type for each target for grid search optimization are highlighted in green. These were selected based on their validation and training AUC scores. Each of the tables represents one run. The results for the second repeatability run have not been given in this paper because they were very close to the results of the first run and would take up too much of this paper. These results are included in the feature selection analysis spreadsheet of the final project submission.

The results of the grid search and overall model runs for the best two feature selection models of each model type are given in table 14 for the biopsy target dataset and table 15 for the combination target dataset. Table 16 contains the features selected for each of the biopsy target final model runs. Table 17 contains the features selected for each of the combination target final model runs. The final model chosen for each target are highlighted green in these tables.

Biopsy Target Random Forest Model Feature Selection Results

Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	540	40	NA	0.93	0.01	0.72	0.10	0.92	0.49
None	None	Low Variance Filter	540	40	8	0.93	0.00	0.58	0.17	0.93	0.50
None	None	Model Wrapper	540	40	8	0.94	0.01	0.60	0.13	0.92	0.50
None	None	Stepwise Recursive	540	40	8	0.93	0.01	0.71	0.12	0.93	0.50
None	None	Chi-Squared	540	40	10	0.93	0.01	0.72	0.19	0.89	0.52
None	None	Mutual Information	540	40	9	0.92	0.04	0.62	0.07	0.92	0.59
SMOTE	1	None	540	540	NA	0.97	0.06	0.99	0.02	0.90	0.53
SMOTE	1	Low Variance Filter	540	540	8	0.96	0.08	0.99	0.02	0.88	0.47
SMOTE	1	Model Wrapper	540	540	8	0.96	0.07	1.00	0.01	0.84	0.45
SMOTE	1	Stepwise Recursive	540	540	8	0.96	0.07	1.00	0.01	0.84	0.45
SMOTE	1	Chi-Squared	540	540	9	0.90	0.06	0.94	0.05	0.88	0.57
SMOTE	1	Mutual Information	540	540	10	0.96	0.06	0.99	0.02	0.85	0.50
SMOTE	0.3	None	540	162	NA	0.94	0.10	0.97	0.06	0.89	0.48
SMOTE	0.3	Low Variance Filter	540	162	8	0.93	0.10	0.96	0.08	0.91	0.49
SMOTE	0.3	Model Wrapper	540	162	8	0.94	0.09	0.97	0.05	0.85	0.46
SMOTE	0.3	Stepwise Recursive	540	162	9	0.94	0.06	0.97	0.05	0.88	0.47
SMOTE	0.3	Chi-Squared	540	162	10	0.89	0.08	0.87	0.09	0.90	0.57
SMOTE	0.3	Mutual Information	540	162	8	0.93	0.11	0.97	0.06	0.89	0.48
Under	1	None	40	40	NA	0.65	0.15	0.66	0.18	0.54	0.43
Under	1	Low Variance Filter	40	40	9	0.56	0.29	0.53	0.30	0.59	0.50
Under	1	Model Wrapper	40	40	7	0.53	0.19	0.47	0.15	0.58	0.54
Under	1	Stepwise Recursive	40	40	7	0.60	0.17	0.63	0.16	0.58	0.54
Under	1	Chi-Squared	40	40	10	0.70	0.22	0.74	0.16	0.62	0.47
Under	1	Mutual Information	40	40	10	0.66	0.10	0.73	0.22	0.52	0.46
Under	0.3	None	133	40	NA	0.76	0.03	0.70	0.11	0.84	0.50
Under	0.3	Low Variance Filter	133	40	9	0.75	0.07	0.66	0.18	0.88	0.57
Under	0.3	Model Wrapper	133	40	8	0.76	0.04	0.62	0.17	0.85	0.50
Under	0.3	Stepwise Recursive	133	40	7	0.79	0.06	0.73	0.06	0.80	0.48
Under	0.3	Chi-Squared	133	40	10	0.76	0.05	0.71	0.11	0.81	0.48
Under	0.3	Mutual Information	133	40	9	0.77	0.04	0.68	0.12	0.88	0.52
Under	0.1	None	400	40	NA	0.91	0.01	0.69	0.11	0.90	0.49
Under	0.1	Low Variance Filter	400	40	8	0.90	0.01	0.61	0.12	0.93	0.50
Under	0.1	Model Wrapper	400	40	8	0.91	0.01	0.61	0.12	0.92	0.49
Under	0.1	Stepwise Recursive	400	40	7	0.91	0.02	0.70	0.10	0.91	0.49
Under	0.1	Chi-Squared	400	40	10	0.89	0.03	0.72	0.20	0.89	0.52
Under	0.1	Mutual Information	400	40	9	0.90	0.03	0.74	0.07	0.85	0.50

Table 6. Biopsy Target Random Forest Feature Selection Model Results

Biopsy Target Gradient Boost Model Feature Selection Results

Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	540	40	NA	0.92	0.02	0.64	0.21	0.91	0.49
None	None	Low Variance Filter	540	40	8	0.92	0.02	0.54	0.17	0.91	0.49
None	None	Model Wrapper	540	40	12	0.93	0.02	0.63	0.24	0.91	0.49
None	None	Stepwise Recursive	540	40	10	0.93	0.02	0.63	0.21	0.90	0.49
None	None	Chi-Squared	540	40	5	0.93	0.02	0.65	0.23	0.90	0.58
None	None	Mutual Information	540	40	9	0.92	0.03	0.65	0.21	0.92	0.50
SMOTE	1	None	540	540	NA	0.94	0.08	0.99	0.03	0.87	0.47
SMOTE	1	Low Variance Filter	540	540	8	0.95	0.12	0.98	0.06	0.88	0.47
SMOTE	1	Model Wrapper	540	540	10	0.94	0.08	0.98	0.04	0.86	0.56
SMOTE	1	Stepwise Recursive	540	540	8	0.95	0.09	0.99	0.04	0.86	0.51
SMOTE	1	Chi-Squared	540	540	9	0.87	0.06	0.94	0.06	0.86	0.55
SMOTE	1	Mutual Information	540	540	8	0.95	0.09	0.98	0.05	0.86	0.51
SMOTE	0.3	None	540	162	NA	0.92	0.12	0.95	0.11	0.91	0.58
SMOTE	0.3	Low Variance Filter	540	162	8	0.90	0.12	0.93	0.16	0.92	0.49
SMOTE	0.3	Model Wrapper	540	162	9	0.91	0.13	0.94	0.15	0.91	0.49
SMOTE	0.3	Stepwise Recursive	540	162	10	0.91	0.13	0.95	0.12	0.91	0.49
SMOTE	0.3	Chi-Squared	540	162	6	0.88	0.09	0.89	0.15	0.88	0.52
SMOTE	0.3	Mutual Information	540	162	9	0.91	0.15	0.94	0.14	0.88	0.47
Under	1	None	40	40	NA	0.58	0.28	0.60	0.18	0.63	0.66
Under	1	Low Variance Filter	40	40	9	0.53	0.23	0.55	0.24	0.54	0.52
Under	1	Model Wrapper	40	40	9	0.59	0.26	0.62	0.26	0.55	0.62
Under	1	Stepwise Recursive	40	40	8	0.56	0.26	0.64	0.18	0.60	0.60
Under	1	Chi-Squared	40	40	9	0.68	0.25	0.71	0.30	0.66	0.54
Under	1	Mutual Information	40	40	8	0.65	0.22	0.70	0.30	0.56	0.49
Under	0.3	None	133	40	NA	0.75	0.10	0.65	0.19	0.82	0.53
Under	0.3	Low Variance Filter	133	40	9	0.72	0.13	0.61	0.18	0.83	0.54
Under	0.3	Model Wrapper	133	40	10	0.77	0.07	0.68	0.13	0.82	0.49
Under	0.3	Stepwise Recursive	133	40	10	0.77	0.07	0.68	0.13	0.82	0.49
Under	0.3	Chi-Squared	133	40	10	0.77	0.05	0.65	0.21	0.81	0.48
Under	0.3	Mutual Information	133	40	5	0.77	0.05	0.67	0.07	0.89	0.52
Under	0.1	None	400	40	NA	0.90	0.03	0.65	0.22	0.90	0.49
Under	0.1	Low Variance Filter	400	40	8	0.89	0.04	0.56	0.09	0.90	0.49
Under	0.1	Model Wrapper	400	40	11	0.90	0.03	0.64	0.25	0.90	0.48
Under	0.1	Stepwise Recursive	400	40	10	0.90	0.05	0.65	0.27	0.90	0.49
Under	0.1	Chi-Squared	400	40	10	0.91	0.03	0.61	0.32	0.90	0.53
Under	0.1	Mutual Information	400	40	7	0.88	0.04	0.65	0.20	0.90	0.48

Table 7. Biopsy Target Gradient Boost Feature Selection Model Results

Biopsy Target Neural Network Model Feature Selection Results

Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	540	40	NA	0.93	0.00	0.61	0.05	0.93	0.50
None	None	Low Variance Filter	540	40	8	0.93	0.01	0.49	0.15	0.92	0.54
None	None	Chi-Squared	540	40	5	0.93	0.02	0.74	0.17	0.90	0.53
None	None	Mutual Information	540	40	6	0.93	0.01	0.68	0.20	0.92	0.50
SMOTE	1	None	540	540	NA	0.70	0.16	0.73	0.24	0.77	0.46
SMOTE	1	Low Variance Filter	540	540	8	0.74	0.04	0.80	0.04	0.63	0.43
SMOTE	1	Chi-Squared	540	540	10	0.83	0.02	0.89	0.05	0.77	0.55
SMOTE	1	Mutual Information	540	540	10	0.80	0.05	0.86	0.05	0.60	0.55
SMOTE	0.3	None	540	162	NA	0.85	0.06	0.88	0.12	0.78	0.47
SMOTE	0.3	Low Variance Filter	540	162	8	0.76	0.04	0.75	0.10	0.93	0.50
SMOTE	0.3	Chi-Squared	540	162	6	0.84	0.03	0.86	0.09	0.84	0.54
SMOTE	0.3	Mutual Information	540	162	10	0.84	0.04	0.83	0.04	0.81	0.53
Under	1	None	40	40	NA	0.49	0.05	0.49	0.05	0.08	0.46
Under	1	Low Variance Filter	40	40	9	0.53	0.22	0.50	0.18	0.52	0.33
Under	1	Chi-Squared	40	40	9	0.74	0.09	0.73	0.04	0.51	0.46
Under	1	Mutual Information	40	40	7	0.73	0.13	0.70	0.15	0.71	0.52
Under	0.3	None	133	40	NA	0.71	0.06	0.72	0.17	0.66	0.58
Under	0.3	Low Variance Filter	133	40	9	0.74	0.05	0.50	0.22	0.88	0.61
Under	0.3	Chi-Squared	133	40	7	0.80	0.04	0.63	0.10	0.78	0.56
Under	0.3	Mutual Information	133	40	9	0.75	0.09	0.68	0.19	0.89	0.52
Under	0.1	None	400	40	NA	0.91	0.00	0.58	0.22	0.93	0.50
Under	0.1	Low Variance Filter	400	40	8	0.90	0.01	0.49	0.08	0.92	0.54
Under	0.1	Chi-Squared	400	40	9	0.90	0.02	0.60	0.10	0.88	0.52
Under	0.1	Mutual Information	400	40	9	0.90	0.03	0.71	0.16	0.93	0.50

Table 8. Biopsy Target Neural Network Feature Selection Model Results

Biopsy Target SVM Model Feature Selection Results											
Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	540	40	NA	0.93	0.00	0.60	0.16	0.93	0.50
None	None	Low Variance Filter	540	40	8	0.93	0.00	0.53	0.24	0.93	0.50
None	None	Chi-Squared	540	40	6	0.93	0.00	0.64	0.07	0.93	0.50
SMOTE	None	Mutual Information	540	40	5	0.93	0.00	0.60	0.21	0.93	0.50
SMOTE	1	None	540	540	NA	0.77	0.04	0.85	0.01	0.76	0.50
SMOTE	1	Low Variance Filter	540	540	8	0.75	0.04	0.85	0.02	0.66	0.40
SMOTE	1	Chi-Squared	540	540	6	0.75	0.05	0.82	0.07	0.75	0.59
SMOTE	1	Mutual Information	540	540	8	0.79	0.05	0.86	0.06	0.64	0.44
SMOTE	0.3	None	540	162	8	0.78	0.02	0.84	0.04	0.92	0.49
SMOTE	0.3	Low Variance Filter	540	162	8	0.79	0.04	0.80	0.04	0.91	0.49
SMOTE	0.3	Chi-Squared	540	162	9	0.81	0.05	0.82	0.07	0.86	0.46
SMOTE	0.3	Mutual Information	540	162	9	0.78	0.04	0.82	0.09	0.91	0.54
Under	1	None	40	40	NA	0.46	0.06	0.46	0.23	0.62	0.52
Under	1	Low Variance Filter	40	40	9	0.44	0.08	0.39	0.13	0.53	0.51
Under	1	Chi-Squared	40	40	10	0.68	0.23	0.68	0.20	0.68	0.60
Under	1	Mutual Information	40	40	6	0.68	0.32	0.70	0.26	0.42	0.41
Under	0.3	None	133	40	NA	0.76	0.02	0.63	0.15	0.91	0.49
Under	0.3	Low Variance Filter	133	40	9	0.76	0.02	0.58	0.12	0.92	0.54
Under	0.3	Chi-Squared	133	40	9	0.76	0.02	0.76	0.18	0.86	0.51
Under	0.3	Mutual Information	133	40	7	0.76	0.02	0.71	0.25	0.92	0.59
Under	0.1	None	400	40	NA	0.91	0.00	0.56	0.07	0.93	0.50
Under	0.1	Low Variance Filter	400	40	8	0.91	0.00	0.50	0.16	0.93	0.50
Under	0.1	Chi-Squared	400	40	10	0.91	0.01	0.69	0.20	0.93	0.50
Under	0.1	Mutual Information	400	40	9	0.91	0.00	0.63	0.22	0.93	0.50

Table 9. Biopsy Target SVM Feature Selection Model Results

Combination Target Random Forest Model Feature Selection Results											
Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	506	74	NA	0.87	0.04	0.65	0.14	0.85	0.49
None	None	Low Variance Filter	506	74	8	0.88	0.02	0.55	0.12	0.86	0.49
None	None	Model Wrapper	506	74	6	0.86	0.02	0.54	0.12	0.86	0.50
None	None	Stepwise Recursive	506	74	10	0.88	0.01	0.63	0.07	0.86	0.50
None	None	Chi-Squared	506	74	10	0.86	0.04	0.62	0.08	0.84	0.48
None	None	Mutual Information	506	74	7	0.85	0.02	0.65	0.08	0.86	0.51
SMOTE	1	None	506	506	NA	0.91	0.18	0.98	0.04	0.84	0.53
SMOTE	1	Low Variance Filter	506	506	8	0.90	0.16	0.97	0.07	0.84	0.55
SMOTE	1	Model Wrapper	506	506	6	0.90	0.18	0.97	0.09	0.84	0.53
SMOTE	1	Stepwise Recursive	506	506	8	0.90	0.20	0.98	0.06	0.84	0.51
SMOTE	1	Chi-Squared	506	506	10	0.80	0.17	0.86	0.15	0.82	0.52
SMOTE	1	Mutual Information	506	506	9	0.90	0.20	0.97	0.07	0.85	0.53
SMOTE	0.3	None	506	151	NA	0.86	0.12	0.87	0.18	0.84	0.48
SMOTE	0.3	Low Variance Filter	506	151	8	0.86	0.12	0.83	0.22	0.86	0.51
SMOTE	0.3	Model Wrapper	506	151	7	0.86	0.14	0.83	0.24	0.85	0.51
SMOTE	0.3	Stepwise Recursive	506	151	9	0.86	0.12	0.85	0.21	0.85	0.49
SMOTE	0.3	Chi-Squared	506	151	8	0.82	0.12	0.76	0.20	0.82	0.47
SMOTE	0.3	Mutual Information	506	151	8	0.85	0.11	0.87	0.18	0.86	0.49
Under	1	None	74	74	NA	0.50	0.12	0.49	0.19	0.68	0.64
Under	1	Low Variance Filter	74	74	9	0.46	0.17	0.47	0.20	0.64	0.66
Under	1	Model Wrapper	74	74	7	0.47	0.16	0.48	0.19	0.64	0.66
Under	1	Stepwise Recursive	74	74	7	0.47	0.16	0.48	0.19	0.64	0.66
Under	1	Chi-Squared	74	74	8	0.52	0.14	0.53	0.25	0.66	0.58
Under	1	Mutual Information	74	74	7	0.56	0.21	0.57	0.28	0.54	0.56
Under	0.3	None	246	74	NA	0.75	0.10	0.58	0.16	0.82	0.47
Under	0.3	Low Variance Filter	246	74	8	0.74	0.10	0.52	0.12	0.82	0.47
Under	0.3	Model Wrapper	246	74	6	0.73	0.08	0.52	0.12	0.83	0.48
Under	0.3	Stepwise Recursive	246	74	10	0.76	0.09	0.57	0.15	0.84	0.48
Under	0.3	Chi-Squared	246	74	6	0.72	0.08	0.66	0.09	0.84	0.51
Under	0.3	Mutual Information	246	74	8	0.75	0.06	0.60	0.13	0.86	0.56
Under	0.2	None	370	74	NA	0.82	0.06	0.63	0.15	0.84	0.48
Under	0.2	Low Variance Filter	370	74	8	0.83	0.04	0.53	0.07	0.85	0.49
Under	0.2	Model Wrapper	370	74	7	0.82	0.03	0.54	0.08	0.86	0.51
Under	0.2	Stepwise Recursive	370	74	10	0.83	0.04	0.61	0.12	0.84	0.48
Under	0.2	Chi-Squared	370	74	7	0.82	0.06	0.64	0.07	0.84	0.48
Under	0.2	Mutual Information	370	74	9	0.79	0.04	0.64	0.10	0.84	0.57

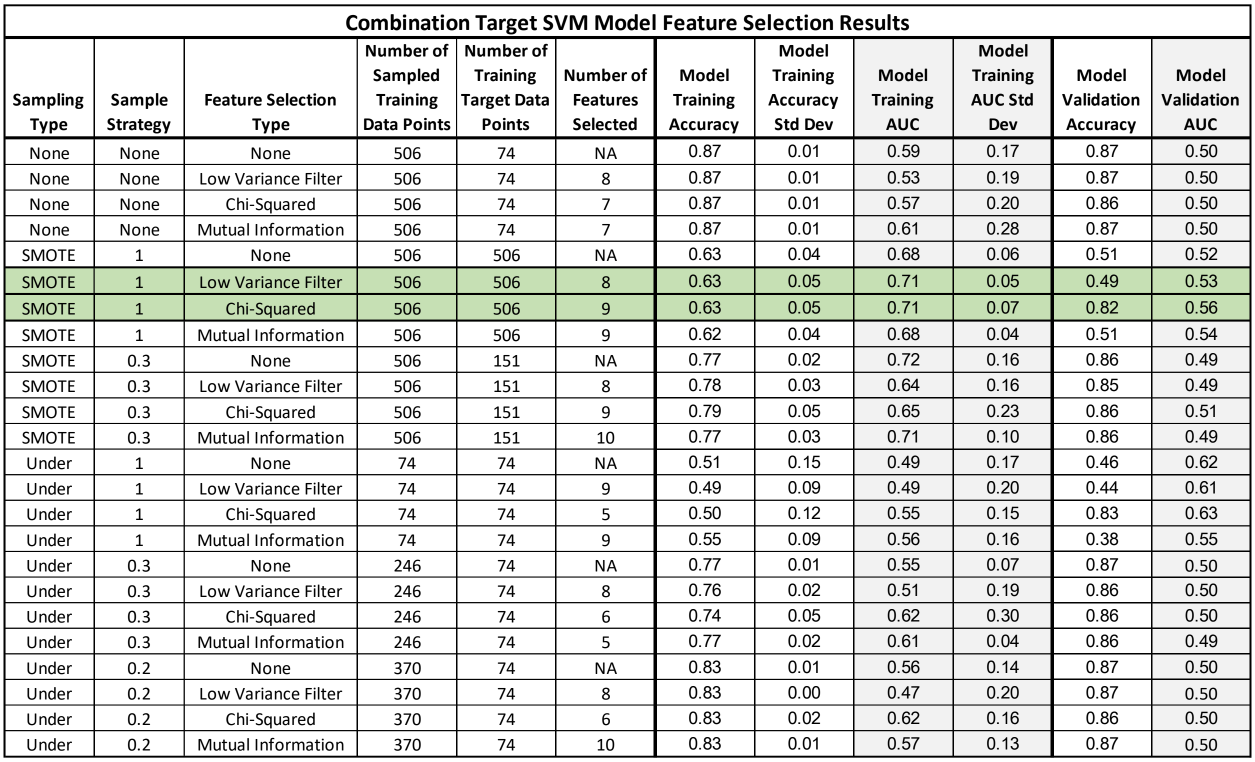
Table 10. Combination Target Random Forest Feature Selection Model Results

Combination Target Gradient Boost Model Feature Selection Results											
Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	506	74	NA	0.86	0.01	0.61	0.19	0.84	0.48
None	None	Low Variance Filter	506	74	8	0.86	0.02	0.55	0.13	0.86	0.49
None	None	Model Wrapper	506	74	10	0.87	0.04	0.61	0.17	0.83	0.48
None	None	Stepwise Recursive	506	74	9	0.86	0.03	0.63	0.18	0.84	0.48
None	None	Chi-Squared	506	74	6	0.87	0.01	0.61	0.14	0.85	0.49
None	None	Mutual Information	506	74	9	0.86	0.03	0.62	0.13	0.87	0.50
SMOTE	1	None	506	506	NA	0.89	0.25	0.96	0.12	0.85	0.49
SMOTE	1	Low Variance Filter	506	506	8	0.88	0.23	0.96	0.13	0.81	0.46
SMOTE	1	Model Wrapper	506	506	9	0.89	0.24	0.96	0.11	0.84	0.48
SMOTE	1	Stepwise Recursive	506	506	10	0.89	0.25	0.96	0.11	0.85	0.49
SMOTE	1	Chi-Squared	506	506	5	0.77	0.16	0.85	0.19	0.83	0.50
SMOTE	1	Mutual Information	506	506	10	0.89	0.24	0.96	0.12	0.84	0.48
SMOTE	0.3	None	506	151	NA	0.84	0.13	0.81	0.24	0.86	0.49
SMOTE	0.3	Low Variance Filter	506	151	8	0.82	0.07	0.77	0.25	0.86	0.49
SMOTE	0.3	Model Wrapper	506	151	10	0.84	0.14	0.82	0.26	0.84	0.48
SMOTE	0.3	Stepwise Recursive	506	151	9	0.84	0.12	0.81	0.25	0.86	0.49
SMOTE	0.3	Chi-Squared	506	151	10	0.83	0.10	0.74	0.20	0.85	0.49
SMOTE	0.3	Mutual Information	506	151	8	0.85	0.14	0.82	0.25	0.86	0.49
Under	1	None	74	74	NA	0.47	0.17	0.47	0.23	0.62	0.62
Under	1	Low Variance Filter	74	74	8	0.47	0.16	0.52	0.29	0.53	0.53
Under	1	Model Wrapper	74	74	10	0.47	0.12	0.47	0.22	0.59	0.63
Under	1	Stepwise Recursive	74	74	7	0.48	0.18	0.52	0.32	0.60	0.62
Under	1	Chi-Squared	74	74	5	0.49	0.10	0.52	0.26	0.73	0.62
Under	1	Mutual Information	74	74	7	0.56	0.08	0.57	0.12	0.84	0.57
Under	0.3	None	246	74	NA	0.72	0.11	0.56	0.22	0.83	0.50
Under	0.3	Low Variance Filter	246	74	8	0.71	0.11	0.47	0.12	0.80	0.48
Under	0.3	Model Wrapper	246	74	9	0.73	0.11	0.56	0.22	0.82	0.49
Under	0.3	Stepwise Recursive	246	74	7	0.73	0.08	0.58	0.17	0.82	0.47
Under	0.3	Chi-Squared	246	74	6	0.74	0.11	0.60	0.19	0.84	0.51
Under	0.3	Mutual Information	246	74	5	0.75	0.08	0.60	0.17	0.80	0.46
Under	0.2	None	370	74	NA	0.81	0.07	0.58	0.18	0.84	0.50
Under	0.2	Low Variance Filter	370	74	8	0.82	0.07	0.53	0.15	0.82	0.51
Under	0.2	Model Wrapper	370	74	9	0.82	0.09	0.61	0.16	0.83	0.52
Under	0.2	Stepwise Recursive	370	74	9	0.82	0.09	0.61	0.16	0.83	0.52
Under	0.2	Chi-Squared	370	74	10	0.81	0.05	0.57	0.16	0.84	0.48
Under	0.2	Mutual Information	370	74	5	0.83	0.03	0.60	0.16	0.86	0.50

Table 11. Combination Target Gradient Boost Feature Selection Model Results

Combination Target Neural Network Model Feature Selection Results											
Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	506	74	NA	0.87	0.01	0.61	0.18	0.87	0.50
None	None	Low Variance Filter	506	74	8	0.87	0.01	0.50	0.11	0.87	0.50
None	None	Chi-Squared	506	74	5	0.86	0.02	0.60	0.06	0.86	0.50
None	None	Mutual Information	506	74	8	0.88	0.02	0.65	0.19	0.85	0.49
SMOTE	1	None	506	506	NA	0.52	0.02	0.53	0.03	0.30	0.58
SMOTE	1	Low Variance Filter	506	506	8	0.61	0.05	0.63	0.05	0.57	0.60
SMOTE	1	Chi-Squared	506	506	8	0.71	0.06	0.75	0.13	0.68	0.57
SMOTE	1	Mutual Information	506	506	8	0.67	0.08	0.74	0.08	0.51	0.60
SMOTE	0.3	None	506	151	NA	0.76	0.06	0.68	0.10	0.83	0.54
SMOTE	0.3	Low Variance Filter	506	151	8	0.75	0.04	0.59	0.14	0.86	0.50
SMOTE	0.3	Chi-Squared	506	151	10	0.80	0.05	0.70	0.12	0.84	0.53
SMOTE	0.3	Mutual Information	506	151	7	0.78	0.05	0.65	0.12	0.87	0.50
Under	1	None	74	74	NA	0.48	0.05	0.49	0.07	0.16	0.52
Under	1	Low Variance Filter	74	74	9	0.49	0.22	0.48	0.19	0.81	0.58
Under	1	Chi-Squared	74	74	5	0.56	0.20	0.51	0.27	0.82	0.56
Under	1	Mutual Information	74	74	8	0.60	0.15	0.61	0.12	0.82	0.56
Under	0.3	None	246	74	NA	0.75	0.09	0.56	0.17	0.82	0.54
Under	0.3	Low Variance Filter	246	74	8	0.76	0.03	0.47	0.12	0.86	0.50
Under	0.3	Chi-Squared	246	74	5	0.75	0.09	0.63	0.16	0.84	0.53
Under	0.3	Mutual Information	246	74	9	0.75	0.10	0.61	0.19	0.86	0.52
Under	0.2	None	370	74	NA	0.82	0.09	0.60	0.17	0.88	0.53
Under	0.2	Low Variance Filter	370	74	8	0.82	0.03	0.50	0.04	0.88	0.53
Under	0.2	Chi-Squared	370	74	7	0.82	0.04	0.62	0.15	0.86	0.54
Under	0.2	Mutual Information	370	74	5	0.83	0.04	0.60	0.20	0.86	0.50

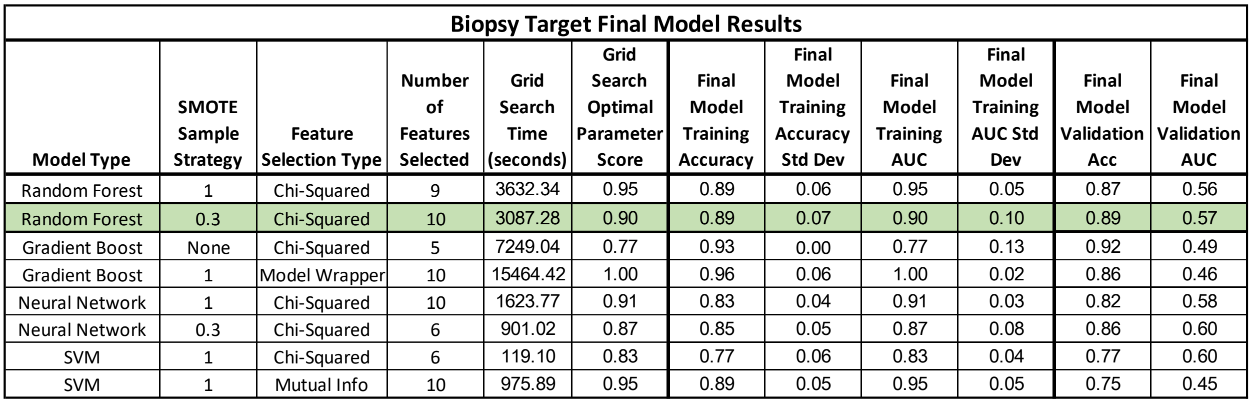
Table 12. Combination Target Neural Network Feature Selection Model Results



Combination Target SVM Model Feature Selection Results

Sampling Type	Sample Strategy	Feature Selection Type	Number of Sampled Training Data Points	Number of Training Target Data Points	Number of Features Selected	Model Training Accuracy	Model Training Accuracy Std Dev	Model Training AUC	Model Training AUC Std Dev	Model Validation Accuracy	Model Validation AUC
None	None	None	506	74	NA	0.87	0.01	0.59	0.17	0.87	0.50
None	None	Low Variance Filter	506	74	8	0.87	0.01	0.53	0.19	0.87	0.50
None	None	Chi-Squared	506	74	7	0.87	0.01	0.57	0.20	0.86	0.50
None	None	Mutual Information	506	74	7	0.87	0.01	0.61	0.28	0.87	0.50
SMOTE	1	None	506	506	NA	0.63	0.04	0.68	0.06	0.51	0.52
SMOTE	1	Low Variance Filter	506	506	8	0.63	0.05	0.71	0.05	0.49	0.53
SMOTE	1	Chi-Squared	506	506	9	0.63	0.05	0.71	0.07	0.82	0.56
SMOTE	1	Mutual Information	506	506	9	0.62	0.04	0.68	0.04	0.51	0.54
SMOTE	0.3	None	506	151	NA	0.77	0.02	0.72	0.16	0.86	0.49
SMOTE	0.3	Low Variance Filter	506	151	8	0.78	0.03	0.64	0.16	0.85	0.49
SMOTE	0.3	Chi-Squared	506	151	9	0.79	0.05	0.65	0.23	0.86	0.51
SMOTE	0.3	Mutual Information	506	151	10	0.77	0.03	0.71	0.10	0.86	0.49
Under	1	None	74	74	NA	0.51	0.15	0.49	0.17	0.46	0.62
Under	1	Low Variance Filter	74	74	9	0.49	0.09	0.49	0.20	0.44	0.61
Under	1	Chi-Squared	74	74	5	0.50	0.12	0.55	0.15	0.83	0.63
Under	1	Mutual Information	74	74	9	0.55	0.09	0.56	0.16	0.38	0.55
Under	0.3	None	246	74	NA	0.77	0.01	0.55	0.07	0.87	0.50
Under	0.3	Low Variance Filter	246	74	8	0.76	0.02	0.51	0.19	0.86	0.50
Under	0.3	Chi-Squared	246	74	6	0.74	0.05	0.62	0.30	0.86	0.50
Under	0.3	Mutual Information	246	74	5	0.77	0.02	0.61	0.04	0.86	0.49
Under	0.2	None	370	74	NA	0.83	0.01	0.56	0.14	0.87	0.50
Under	0.2	Low Variance Filter	370	74	8	0.83	0.00	0.47	0.20	0.87	0.50
Under	0.2	Chi-Squared	370	74	6	0.83	0.02	0.62	0.16	0.86	0.50
Under	0.2	Mutual Information	370	74	10	0.83	0.01	0.57	0.13	0.87	0.50

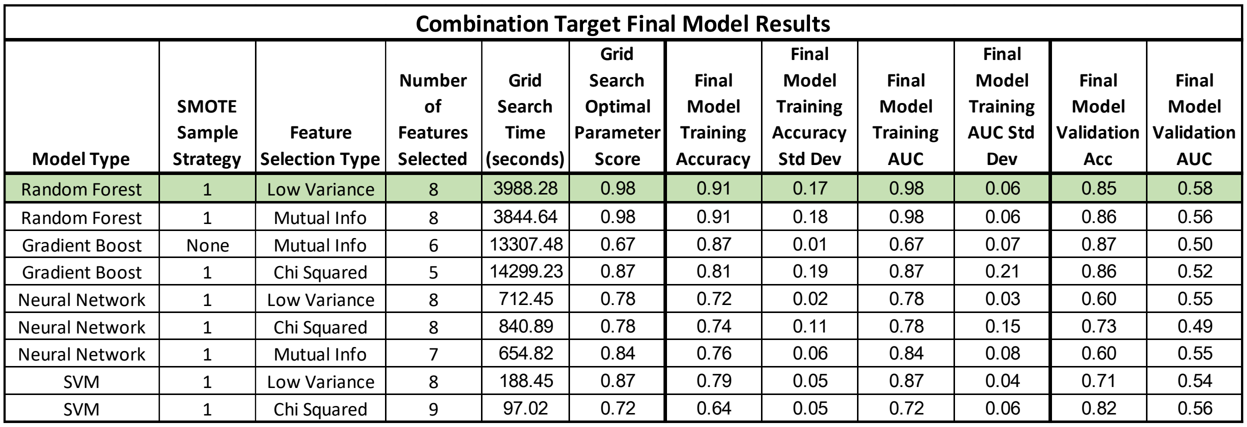
Table 13. Combination Target SVM Feature Selection Model Results



Biopsy Target Final Model Results

Model Type	SMOTE Sample Strategy	Feature Selection Type	Number of Features Selected	Grid Search Time (seconds)	Grid Search Optimal Parameter Score	Final Model Training Accuracy	Final Model Training Accuracy Std Dev	Final Model Training AUC	Final Model Training AUC Std Dev	Final Model Validation Acc	Final Model Validation AUC
Random Forest	1	Chi-Squared	9	3632.34	0.95	0.89	0.06	0.95	0.05	0.87	0.56
Random Forest	0.3	Chi-Squared	10	3087.28	0.90	0.89	0.07	0.90	0.10	0.89	0.57
Gradient Boost	None	Chi-Squared	5	7249.04	0.77	0.93	0.00	0.77	0.13	0.92	0.49
Gradient Boost	1	Model Wrapper	10	15464.42	1.00	0.96	0.06	1.00	0.02	0.86	0.46
Neural Network	1	Chi-Squared	10	1623.77	0.91	0.83	0.04	0.91	0.03	0.82	0.58
Neural Network	0.3	Chi-Squared	6	901.02	0.87	0.85	0.05	0.87	0.08	0.86	0.60
SVM	1	Chi-Squared	6	119.10	0.83	0.77	0.06	0.83	0.04	0.77	0.60
SVM	1	Mutual Info	10	975.89	0.95	0.89	0.05	0.95	0.05	0.75	0.45

Table 14. Biopsy Target Final Model Results



Combination Target Final Model Results

Model Type	SMOTE Sample Strategy	Feature Selection Type	Number of Features Selected	Grid Search Time (seconds)	Grid Search Optimal Parameter Score	Final Model Training Accuracy	Final Model Training Accuracy Std Dev	Final Model Training AUC	Final Model Training AUC Std Dev	Final Model Validation Acc	Final Model Validation AUC
Random Forest	1	Low Variance	8	3988.28	0.98	0.91	0.17	0.98	0.06	0.85	0.58
Random Forest	1	Mutual Info	8	3844.64	0.98	0.91	0.18	0.98	0.06	0.86	0.56
Gradient Boost	None	Mutual Info	6	13307.48	0.67	0.87	0.01	0.67	0.07	0.87	0.50
Gradient Boost	1	Chi Squared	5	14299.23	0.87	0.81	0.19	0.87	0.21	0.86	0.52
Neural Network	1	Low Variance	8	712.45	0.78	0.72	0.02	0.78	0.03	0.60	0.55
Neural Network	1	Chi Squared	8	840.89	0.78	0.74	0.11	0.78	0.15	0.73	0.49
Neural Network	1	Mutual Info	7	654.82	0.84	0.76	0.06	0.84	0.08	0.60	0.55
SVM	1	Low Variance	8	188.45	0.87	0.79	0.05	0.87	0.04	0.71	0.54
SVM	1	Chi Squared	9	97.02	0.72	0.64	0.05	0.72	0.06	0.82	0.56

Table 15. Combination Target Final Model Results

Biopsy Target Final Model Features and Feature Importances

Model Type	SMOTE Sample Strategy	Feature Selection Type	age	num_sex_partners	first_sex_int	num_pregnancies	smokes_yrs	smokes_pk_yrs	hormonal_contr	hormonal_contr_yrs	iud_yrs	stds	stds_num	stds_vp_condylomatosis	stds_hiv	stds_num_dx	dx_cancer	dx_hpv	dx
Random Forest	1	Chi-Squared					X			X		X	X		X	X	X	X	X
Random Forest	0.3	Chi-Squared					X	X		X	X	X	X		X		X	X	X
							0.06	0.07		0.30	0.08	0.12	0.09		0.04		0.05	0.06	0.11
Gradient Boost	None	Chi-Squared								X			X				X	X	X
Gradient Boost	1	Model Wrapper	X	X	X	X			X	X		X	X	X					X
Neural Network	1	Chi-Squared					X			X		X	X	X	X	X	X	X	X
Neural Network	0.3	Chi-Squared					X			X			X				X	X	X
SVM	1	Chi-Squared					X			X		X	X					X	X
SVM	1	Mutual Info	X	X	X	X			X	X		X	X	X					X

Table 16. Biopsy Target Final Model Selected Features

Combination Target Final Model Features and Feature Importances

Model Type	SMOTE Sample Strategy	Feature Selection Type	age	num_sex_partners	first_sex_int	num_pregnancies	smokes_yrs	smokes_pk_yrs	hormonal_contr	hormonal_contr_yrs	iud_yrs	stds	stds_num	stds_condylomatosis	stds_vp_condylomatosis	stds_num_dx	dx_cancer	dx_hpv	dx
Random Forest	1	Low Variance	X	X	X	X	X	X		X	X								
			0.17	0.21	0.17	0.17	0.03	0.04		0.17	0.04								
Random Forest	1	Mutual Info	X	X	X	X			X	X			X					X	
Gradient Boost	None	Mutual Info	X		X									X	X	X			X
Gradient Boost	1	Chi Squared								X			X				X	X	X
Neural Network	1	Low Variance	X	X	X	X	X	X		X	X								
Neural Network	1	Chi Squared								X		X	X		X	X	X	X	X
Neural Network	1	Mutual Info	X	X	X	X			X	X		X							
SVM	1	Low Variance	X	X	X	X	X	X		X	X								
SVM	1	Chi Squared								X		X	X	X	X	X	X	X	X

Table 17. Combination Target Final Model Selected Features

Discussion

The results for the random under sampling method for all models given in tables 6 through 13 show poor training cross-validation AUC (from 0.5 to 0.7) compared to the SMOTE oversampled training AUC values (0.65 to 1.0). This poor AUC performance was most likely due to the undersampled training datasets being too small. Valuable data rows were most likely removed from the datasets during the undersampling process, leading to a lack of important information in the resulting undersampled data.

The feature selection models with the highest training cross-validation and validation AUC values all employed SMOTE oversampling. Models without oversampling resulted in low training cross-validation and validation AUC values, most likely due to the presence of the target variable class imbalance. For the biopsy target dataset, most of the top two feature selection models for each model type contained one SMOTE oversampled dataset with a sampling strategy of 1.0 and another with a sampling strategy of 0.3. The only exclusion to this was the gradient boost model, for which a no sampling model was among the top two. The top two feature selection models chosen for the combination biopsy target dataset all used SMOTE oversampling with a sample strategy of 1.0. Since the top models for both targets involved SMOTE oversampled datasets, this shows that oversampling was beneficial.

All of the top models chosen from feature selection for all model types involved a form of feature selection. The method of feature selection varied across model types, but the fact that all the models employed some form of feature selection means that the removal of features was required to obtain the best performance. It is possible that some of the removed features were adding noise to the model, or were redundant.

There were several feature selection models with similar validation AUC values for all of the model types per target. The results of each model were grouped together by validation AUC values, low performance ones (AUC values between 0.4 and 0.5) and higher performing ones (AUC values between 0.51 and 0.6). The training cross-validation AUC values for the higher performing models were then taken into account. A final set of feature selection models was obtained, which had both high validation AUC and high training cross-validation AUC. If any of the final feature selection models had similar performance but did not use feature selection where the others did, they were not used. This is because other models were able to capture the same performance with a smaller set of features (were more parsimonious). This is how the top two models ended up being chosen from the feature selection results for all model types and targets.

The models that were built from the optimal parameters selected from grid search showed that the random forest models had the highest training cross-validation and validation AUC values for both target datasets (see tables 14 and 15). The training cross-validation AUC values were similar for most of the biopsy target models, although the gradient boost model that did not use sampling performed significantly worse than the rest. Similarly, the gradient boost model with no sampling for the combination target also had the worse training cross-validation AUC. This backs the assumption that sampling was beneficial for both targets.

The gradient boost models had lower validation AUC for both the biopsy and combination targets, which means that this model is not the best for either target dataset. The SVM models for the biopsy target showed inconsistent training cross-validation and validation AUC results (low cross-validation with high validation and high cross-validation with low validation). This inconsistency was the reason the SVM model was not chosen for the biopsy target. The biopsy neural network model results were comparable to those of the random forest models. The reason the random forest model was chosen over the neural network came down to model interpretability. Since the models had similar performance, the random forest was chosen because of the ability to discern the importance of the features used in modeling. There were two random forest models present for the biopsy target. The random forest model with the highest validation AUC was chosen as the final model for the biopsy target dataset. The combination neural network model results had the lowest training cross-validation AUC values and also had lower validation AUC values, making the neural network models less desirable for this target. The combination target SVM models had higher validation AUC, but still had significantly lower training cross-validation AUC values when compared to the random forest models. For this reason, the random forest model was chosen for the combination target.

Since there were two random forest models present, the random forest model with the highest validation AUC value was chosen as the final combination target model. Note that the two random forest models for both the biopsy and combination targets had comparable performance. Either of the two random forest models per target could have been have been chosen and would have been acceptable as final models. It should be noted that the features selected for the random forest models per target were very similar.

The final biopsy random forest model used SMOTE oversampling with a sampling strategy of 0.3 and chi-squared univariate features selection with 9 features selected. The final combination random forest model used SMOTE oversampling with a sampling strategy of 1.0 and low variance filter feature selection with 8 features selected. The final validation AUC values of 0.57 for the biopsy random forest model and 0.58 for the combination random forest model were still low. An AUC value of 0.5 is considered to be a random classifier for a balanced binomial target dataset. This means that there is an equal likelihood of both target classes (is a 50% chance of either a 0 or a 1). The validation AUC values of the final models are slightly over 0.5, which means that they do perform better than a random classifier, but still struggle classifying the minority class of the validation set. This shows that the class imbalance of the validation set is an issue in model performance.

The features selected by the final biopsy random forest model were (in order of from most important to least important) the number of years on hormonal contraception, the flag specifying if STDs are present, the general diagnosis flag, the number of STDs, the number of years using an IUD, the number of smoking pack-years, the number of smoking years, the flag specifying an HPV diagnosis, the flag specifying a previous cancer diagnosis, and the flag specifying an HIV diagnosis (refer table 16). The general diagnosis flag, HPV diagnosis flag, previous cancer flag, number of STDs, STDs present flag, and number of years smoking were chosen during feature selection for most overall biopsy models. This implies these variables were dominant in modeling the biopsy target. The features selected by the combination random forest model were (in order from most important to least important) the number of sexual partners, patient age, the age of first sexual encounter, the number of pregnancies, the number of years on hormonal contraception, the number of years using an IUD, the number of smoking pack-years, and the number of years smoking (refer to table 17). The number hormonal contraception years, age, age of firs sexual encounter, number of sexual partners, and number of pregnancies were selected for most overall combination models.

Conclusion

The random forest model was chosen as the final model for both the biopsy and combination target datasets due to performance and interpretability. The final validation AUC values were greater than 0.5, which shows that they are higher than a random classifier, but are still low. This is due to the class imbalance issue still being present in both the validation datasets. SMOTE oversampling did improve the AUC performance of both the biopsy and combination training cross-validation datasets (from around 0.5 to around 0.8 or higher). The change in distribution allowed the models to better fit to the training datasets.

The features selected varied fairly significantly between the final biopsy and combination models. The biopsy features were focused on presence of diseases, including HPV, cancer, HIV, and STDs, where the combination features were more focused on sexual history and age. Both targets included features related to smoking, hormonal contraception, and IUD usage.

The model results in this paper were obtained through a limited range of parameters during both feature selection and grid search optimization. In order to improve model performance, more parameter values should be investigated. Expanding feature selection number of features to 15 or 20 for example could include features that may improve model performance. Opening up the ranges of the model parameters used in the grid searches may also allow more optimal models to be found. Another possibility for improvement is to vary the training and validation data split, then perform all feature selection and model optimization with the new splits. This would investigate the effect of the split on model performance. Similarly, varying the random state used in the generation of the models could also result in improved model performance. These states were held constant in the models mentioned in this paper because model repeatability was important between runs. Now that repeatability has been established, it would be advantageous to test different random states for each model type to investigate their impact. Using other sampling and cost-based learning approaches that were mentioned in the scientific articles reviewed could also provide better performance.

References

(1) Wu, Wen, and Hao Zhou. “Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches.” Ieee Access, vol. 5, 2017, doi:10.1109/ACCESS.2017.2763984.

(2) Muhammed Fahir Unlersen, Kadir Sabanci, and Muciz Ozcan, “Determining Cervical Cancer by Using Machine Learning Methods”, International Journal of Latest Research in Engineering and Technology Volume 3, Issue 12, 2017, pp. 65 – 71

(3) Mohamed Bekkar and Dr. Taklit Akrouf Alitouche, “Imbalanced Data Learning Approaches Review”, International Journal of Data Mining and Knowledge Management Process, Volume 3, No. 4, 2013, pp. 15 – 33

(4) Mazurowski, Maciej A, et al. “Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance.” Neural Networks, vol. 21, no. 2-3, 2008, pp. 427–436., doi:10.1016/j.neunet. 2007.12.031.

(5) Khalilia, Mohammed, et al. “Predicting Disease Risks from Highly Imbalanced Data Using Random Forest.” Bmc Medical Informatics and Decision Making, vol. 11, no. 1, 2011, doi:10.1186/1472-6947-11-51.