**Stats 141XP**

# Differential Features of Orbital Tumors

Natasha Vuong, Kirsten Landsiedel, Zachary Loran, Ashlyn Jew, Oscar Monroy, Emily Hou, and Aidan O'Sullivan

# Abstract

Orbital tumors tend to present themselves similarly in a clinical setting, and specific tumor type diagnosis is typically made after surgery. Thus, computed tomography (CT) and magnetic resonance imaging (MRI) imaging features are integral in differential diagnosis. We want to determine which clinical or imaging features can help doctors best differentiate between cavernous venous malformations (CVM), solitary fibrous tumors (SFT) and orbital schwannomas. For this study, we first used the Fisher's Exact Test to get a good idea as to which variables would be important in differentiating the three different tumor types. Then we employed the use of KNN (K-Nearest Neighbors) and Random Forest models to try to accurately predict orbital tumor types using the available imaging features. The models were trained using subsets of the data, and the models' performance was tested using ROC (Receiver Operating Characteristic) curves, cross-validation, and bootstrapping. We adjusted our different variable subsets according to the models' performance using various methods such as using different training/testing splits to get the most accurate results possible.

We initially identified 11 key variables that allowed us to classify CVM tumors with close to perfect accuracy and SFT/Schwannoma tumors with 70-90% accuracy (although variable) through the use of random forests. We then narrowed these variables down to the most important 5 variables, which provided a similar level of classification accuracy while simplifying our model. We believe that the difference between our model's accuracy in classifying CVM versus SFT/Schwannoma tumors is largely, if not entirely, due to the varying sample sizes between tumors. Our dataset consisted of 31 CVM tumors, 7 Schwannoma tumors, and 11 SFT tumors. Thus, it is likely that if these models were trained with data that contained more Schwannoma and SFT tumors, we might see a similarly high accuracy for those two tumors, as we do with CVMs currently. Our recommendation is to continue second generation classification of T2 variables, increase the sample size so that SFT and Schwannoma have more representation in the data, and minimize the amount of missing values in the data in key variables identified later in this report.

# 1 Introduction

## 1.1 Statement of the Problem

There are three types of orbital tumors that appear similarly in a clinical setting, the type of which is typically confirmed after surgery. Since different surgical operations work better with different tumors, we want to predict the type of tumor to aid in surgical operation selection, as well as determine which features of the data collected can help doctors best differentiate between the three different types of tumors.

## 1.2 Schematic of Variables

| Observation | Diagnosis |
|---|---|
| 1 | $y_1$ |
| 2 | $y_2$ |
| ... | ... |
| 50 | $y_{50}$ |

The outcome variable is diagnosis, and the predictors are all other variables in the dataset.

Numeric variables: age, visual acuity, proptosis, globe dystopia (mm), EOM supraduction (degrees), EOM infraduction (degrees), EOM adduction (degrees), EOM abduction (degrees), vertical strabismus, horizontal strabismus, duration of symptoms, size, ADC ratio, time early, time late, final visual acuity, follow up

Categorical variables: diagnosis (factor), sex (factor), race (factor), laterality (factor), diplopia (factor), vision change (factor), ptosis (factor), irritation (factor), pain (factor), palpable mass (factor), inflammatory signs (factor), globe dystopia direction (factor), ONH edema (factor), ONH pallor (factor), choroidal folds (factor), RAPD (factor), coronal location (factor), conal location (factor), axial location (factor), bony changes (factor), effect on globe (factor), shape (factor), fluid level (factor), T1 intensity (factor), T1 homogeneity (factor), T1 regionality (factor), T2 intensity (factor), T2 homogeneity (factor), T2 regionality (factor), MRI earlycon intensity (factor), MRI earlycon pattern (factor), MRI earlycon regionality (factor), MRI earlycon ring (factor), MRI latecon intensity (factor), MRI latecon pattern (factor), MRI latecon regionality (factor), MRI latecon ring (factor), special signs (factor), intralesional flow void (factor), presence susceptibility (factor), enhancement early (factor), enhancement late (factor), CT intensity (factor), CT homogeneity (factor), CT con intensity (factor), CT con pattern (factor), CT con regionality (factor), calcification (factor), surgical approach, capsule intact (factor), recurrence (factor)

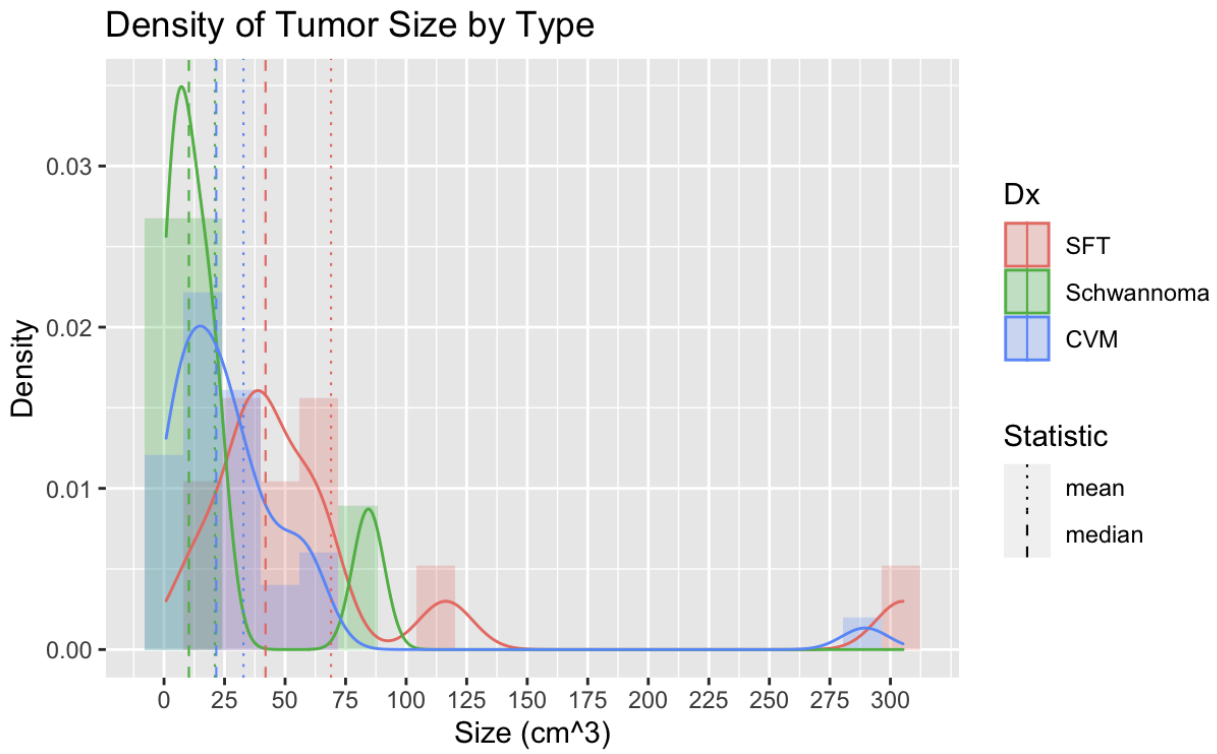
Information on how the variables were measured was not included in the dataset or the research proposal, but the data came from a multi-centered, retrospective observational case series. The variables listed as factors above were already categorical, just coded as numbers (see codebook), so we did not do any additional converting numeric variables into factors.

# 2 Exploratory Data Analysis

Initial exploration showed that the dataset had relatively few observations (50), and 24.27% of the data was missing (coded as NA). Of the radiological variables, the following had less than 20% of their values missing (10 or fewer NAs):

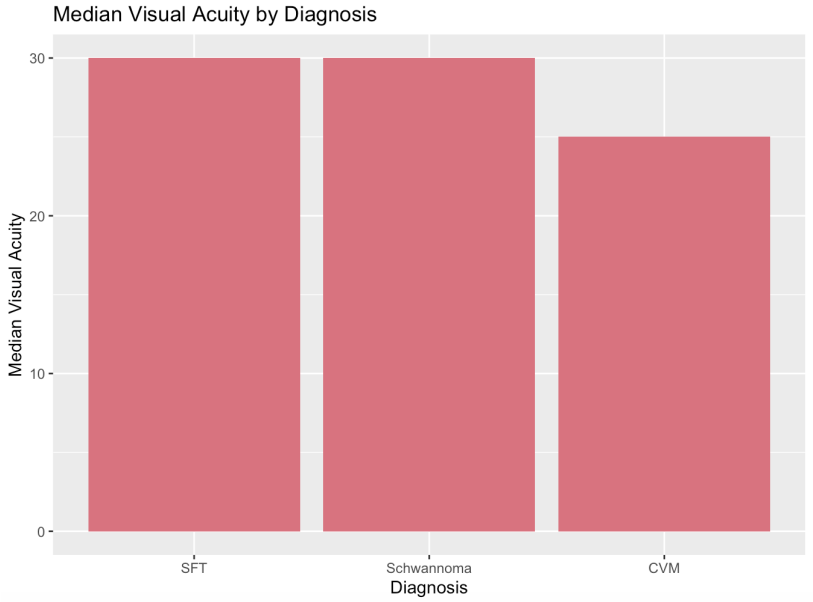
- Size
- Coronal_location
- Conal_Location
- Axial_location
- Bony_changes
- Effect_on_globe
- Shape
- Fluid_Fluid_level
- T1_Intensity
- T1_homogeneity
- T1_regionality
- T2_Intensity
- T2_homogeneity
- T2_regionality
- MRI_EarlyCon_Intensity
- MRI_EarlyCon_Pattern
- MRI_EarlyCon_Regionality
- MRI_EarlyCon_Ring
- MRI_LateCon_Intensity
- MRI_LateCon_Pattern
- MRI_LateCon_Regionality
- MRI_LateCon_Ring

We were curious about the size variable because it was one of the only radiological variables that was numeric, rather than a factor, and had few NAs. We created a density plot to show the distribution of size for each tumor. We used a density plot because it shows proportion, rather than counts, so the three tumors can be compared despite having different numbers of occurrences.

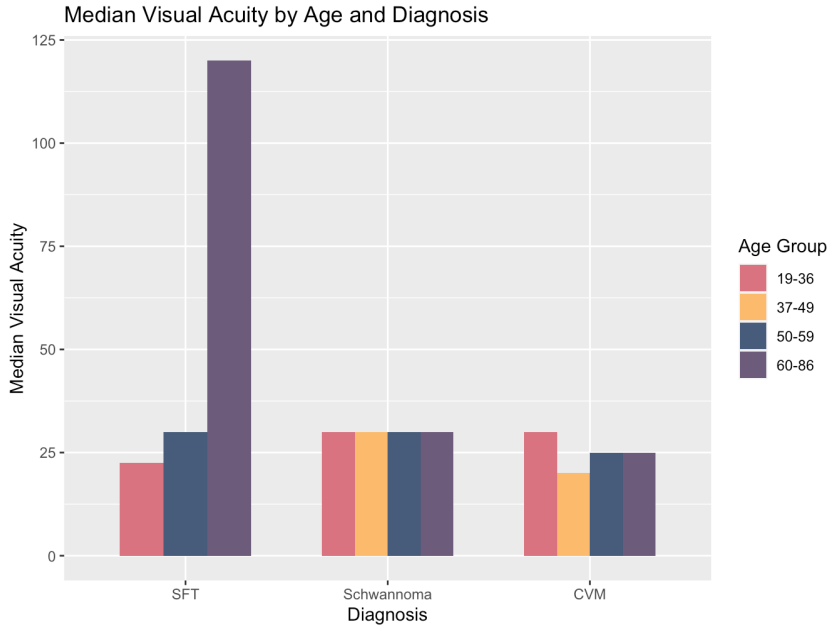



For each tumor type, the histograms in the plot show the proportion of tumors at a particular size. The solid lines are a smoothed representation of the shape of the histograms and show each tumor's size distribution. As seen in the plot, a high proportion of the Schwannoma tumors are smaller in size, while CVM and SFT are slightly larger. The means and medians of each tumor type also reflect this. CVM and SFT also had a few large outliers, which explains why their means are higher than their medians.

Next, we created some preliminary plots and tables to visualize some of the variables across the three different types of tumors.




The graph above shows the median visual acuity for each of the three tumor types, and it is evident that the median visual acuity is quite similar for all three, especially for SFT and Schwannoma, which have the same median visual acuity.




The graph above shows median visual acuity across the three tumor types, but further broken down by age. When split by age, the 60-86 age group for SFT stands out quite a bit, but the rest of the groups are still quite difficult to differentiate just from median visual acuity data alone.

This seemed to be the case across the board for the clinical presentation data, which is in line with the background information that we received from the doctors, that the clinical presentation data alone is not quite sufficient for classifying the three different types of tumors. So, we explored some of the radiological variables recommended by the doctors.

Frequency Table for T2_Intensity

| | Hypo (1) | Iso (1) | Hyper (2) |
|---|---|---|---|
| SFT | 1 | 7 | 3 |
| Schwannoma | 0 | 1 | 6 |
| CVM | 0 | 0 | 28 |

Frequency Table for T2_Regionality

| | Homogeneous (0) | Biphasic (1) | Regional (2) | Inhomogeneous (3) |
|---|---|---|---|---|
| SFT | 4 | 0 | 2 | 5 |
| Schwannoma | 1 | 2 | 0 | 4 |
| CVM | 28 | 0 | 0 | 0 |

We created frequency tables for two of the MRI specific features, T2_Intensity and T2_regionality, because these variables had 10 or fewer NAs in addition to being recommended. As shown in the tables above, all 28 of the CVM tumors have the same presentation of both the T2_Intensity and T2_regionality variables. This initial data exploration further solidified the decision to focus on the imaging features instead of the clinical features in our analysis.

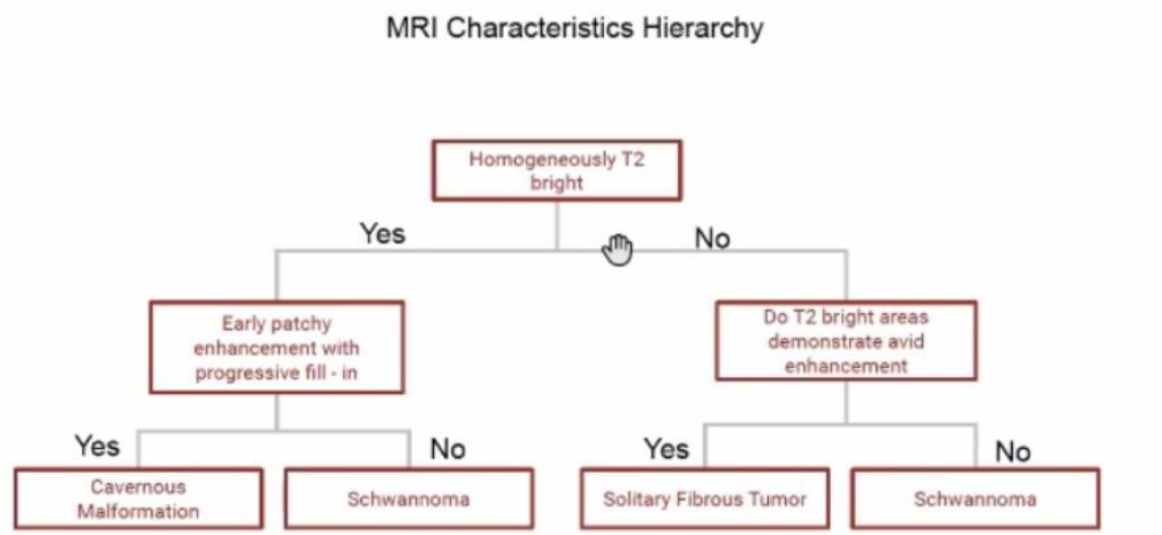
# 3 Methods

## 3.2 Statistical Models

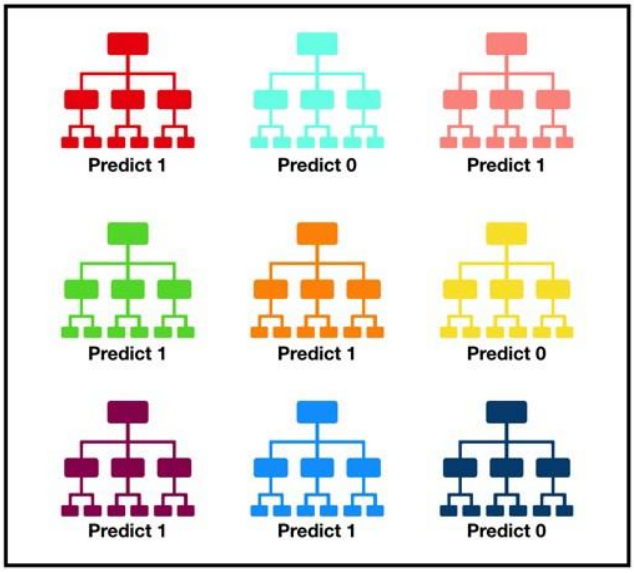
Because of the characteristics of the dataset, we were limited in the models we could work with. We decided to try working with K-Nearest Neighbors and Random Forest models.

K-Nearest Neighbors (KNN) is an algorithm that uses a similarity measure to classify observations, i.e. observations that are most similar to each other are predicted to be in the same class. We chose it because it is a non-parametric algorithm that may be better suited for a small sample size, and it handles multiclass problems well.

When the doctors showed us the chart below and explained that decision trees were already being used clinically to make diagnoses and classifications of the three tumor types, we thought that a Random Forest might be able to expand on that method in a powerful way.




A random forest is a supervised learning method, where the "forest" it builds is an aggregation of decision trees. Once the forest is built, each decision tree in the forest votes towards an overall classification (in this case) of each observation. The figure below shows a small example of a basic random forest.




In the figure, there are several different decision trees that split the data on different variables all voting towards a final diagnosis, which is democratically decided by majority rule.

Normally, we want to feed a random forest as many variables as possible, and we let the algorithm decide which are important and create trees based off of that. However, random forest models do not allow for NAs in **any** variables in **any** observations. This means that the forest will drop all observations with even a single NA. Since we were working with a small sample size, lots of variables, and lots of NAs, we could not give the random forest all of our variables (when we tried, it resulted in 0 observations in our dataset as each row had at least one NA).

## 3.3 Variable Selection Process

We used various methods to select variables for model building.

### 3.3.1 Manual Selection

In order to create random forest models, we had to reduce the number of variables we worked with. Since our dataset was relatively small to begin with, we wanted to drop as few rows as possible, while still keeping variables that could be important for determining tumor type. As such, we decided to identify variables that had both few NAs and high variation in their mean values by tumor type. The cutoff for the number of NAs and mean variation value were chosen through iterations of the model improvement process (see section 3.4.1).

### 3.3.2 Fisher's Exact Test

The Fisher's Exact Test is a statistical significance test that analyzes data in a contingency table. The purpose of this test is to determine if the proportion of one categorical variable (for example, CVM) is significantly different among the values of another categorical variable (such as Schwannoma). We chose to use Fisher's Exact Tests because it gave us information in determining which predictor variables were likely important in differentiating between the types of tumors. The test is also more accurate with small sample sizes compared to other tests like the Chi-squared test. The figure below shows the general formula for the Fisher's Exact Test (right) using a 2x2 contingency table (left).

| | | |
|---|---|---|
| a | b | a+b |
| c | d | c+d |
| a+c | b+d | N=a+b+c+d |

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}}$$

The formula is much more complicated with larger contingency tables, such as a 2x4 table. The test can be performed as a one-tailed test or a two-tailed test, the main difference being that the one-tailed test is used to determine if one categorical variable is better (in the context of the research) than the other, while the two-tailed test is used to determine if the two categorical variables are significantly different in general.

For the purpose of this analysis, we used the two-tailed test as we needed to find any kind of differentiation amongst the variables. We used the significant level threshold of 0.05. To perform the tests, we used R to generate a contingency table between the variable Dx and any predictor variables we thought would be significant. We utilized two methods of the Fisher's Exact Test with R: we calculated the overall significance of the variable and the p-value of individual variable pairings. The overall significance gives us a quick glance at which variables have some form of differentiation, while the individual pair tests give us information on which specific pairs of categorical variables are significantly different. For the pairwise tests, we opted to use the adjusted p-value as that gives greater accuracy when comparing just two categorical variables than non-adjusted p-value. The overall significance value was generated in the base R package using fisher.test() and the significance levels for pairwise testing were generated with the R package "rcompanion" which contains the function pairwiseNominalIndependence().

### 3.3.3 Doctors' Recommendations

The doctors recommended that we explore T1_homogeneity, T2_Intensity, T2_homogeneity, T2_regionality, MRI_Earlycon_Pattern, MRI_Latecon_Pattern, MRI_Earlycon_Regionality, MRI_Latecon_Regionality, MRI_EarlyCon_Intensity, and MRI_LateCon_Intensity.

## 3.4 Model Performance

### 3.4.1 Model Improvement Process

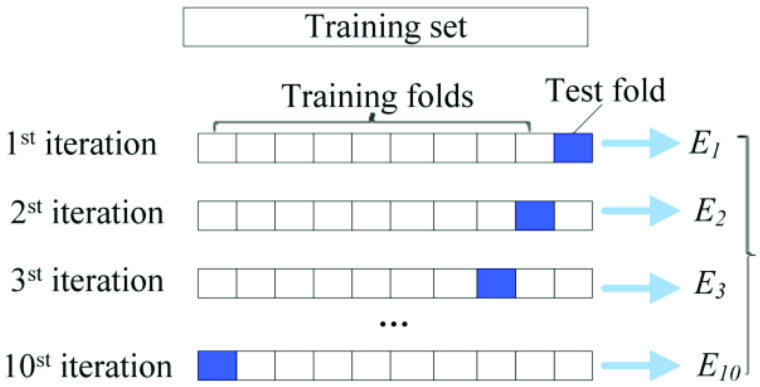
After building the models using the variable selection process, we wanted to improve them by testing their prediction accuracy. It was important to test each model's accuracy to determine whether the variables they deemed important for tumor classification were actually good predictors. Based on prediction accuracy, we chose different variable subsets by tweaking the cutoffs in the variable selection process, using different training/testing splits, and choosing different parameters. With these new variables, we built more models and tested their prediction accuracy as well. We iterated this process a few times to get the best model performance and find the best variables.

### 3.4.2 Testing Prediction Accuracy

The most straightforward way to test prediction accuracy is to split the data into a testing and a training set, fit the model on the training set, and then see how it performs on the testing set. However, the issue with a small dataset is that this prediction accuracy will vary greatly depending on which data points we put in the training set and which we put in the testing set. This problem would not be an issue in larger datasets, because 20% of the dataset could still be hundreds of observations. Therefore, much of the prediction accuracy for our models is a result of randomness - the random process of splitting the data into training and testing - and gives us only a blurry picture of the true predictive power of our models. To get around this issue, we took two approaches: cross-validation and bootstrapping. Both methods harness the power of large numbers by generating many different models with many different train-test splits, then taking the average prediction accuracy across all iterations. By taking an average of many prediction accuracies, we can minimize the influence of random variation, and get a better sense of how well our models can truly classify orbital tumors.

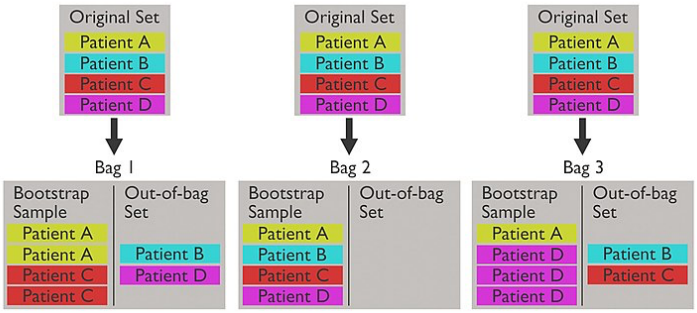
### 3.4.3 Cross-Validation

We used two different types of cross-validation: 10-fold and leave-one-out (LOOCV). Cross-validation works by splitting the full dataset into *k* folds, fitting the model on *k-1* folds, and testing it on the one "left-out" fold. This will result in *k* different estimates of the prediction accuracy, which are then averaged to generate a final aggregate estimate. We used two different values of *k*: 10 and *n*, where *n* is the number of observations in the dataset (either 37 or 38 depending on whether the old or new dataset is used). *n*-fold cross validation is LOOCV. This process is visualized in the graphic below - the model is fit on 9 out of the 10 folds and then tested on the 1 left-out fold. Then, this is repeated until each fold is left out once, for a total of 10 iterations. Each iteration leads to a prediction accuracy estimate, $E_i$, and the grand average, $E$, represents the result of the cross-validation procedure.




### 3.4.4 Bootstrapping

We used two different types of bootstrapping: traditional and "pseudo." Bootstrapping works by resampling from the full dataset with replacement. If a dataset has *n* observations, *n* samples are taken to create the "bootstrap sample," and because the sampling is done with replacement, the "bootstrap sample" will likely contain repeat observations. This process is visualized in the graphic below:




The bootstrap sample essentially functions as the training set that the random forest model is fit on. The difference between the two types of bootstrapping that we performed is in how we construct the testing set. In traditional bootstrapping, we use the 'out-of-bag set' (see

visualization), which are the observations that were not selected by the sampling with replacement. The out-of-bag set can vary in size, depending on how many of the observations are repeated in the bootstrap sample. Because having a small out-of-bag set can negatively impact the estimated prediction accuracy, we decided to use a second bootstrapping type. Namely, we ran 'pseudo-bootstrapping,' wherein we sample a testing set of fixed size (in our case 40% of the full data set) independent of the bootstrap sample. This has the benefit of fixing the testing set size, but the drawback is that the training and testing samples can overlap in the observations they contain, which can overinflate the resulting prediction accuracy. Because one bootstrap type slightly underestimates the prediction accuracy, and the other slightly overestimates it, the true prediction accuracy is probably somewhere in between.

After generating the appropriate testing set (depending on the bootstrap type), we fit the random forest model on the bootstrap sample and calculated its prediction accuracy on the testing set. This process is repeated thousands of times; in our bootstrapping procedure we used 20,000 iterations. While this number of iterations is computationally expensive and takes time to complete, it has the advantage of reducing random variation in our estimates. Finally, one of the main benefits of bootstrapping is that, because we are iterating thousands of times, we can develop an empirical confidence interval for the prediction accuracy of our random forest models.

#### 3.4.5 ROC Curves

We also created ROC curves to get a better understanding of how our models performed. ROC curves are a visual representation of a model's false positive rate and true positive rate. Their graphs can give us an idea of how well a model is classifying each different tumor type. The area under the curve (AUC) gives us an estimate of overall accuracy for each tumor type.

# 4 Results

## 4.1 Variable Selection

For the manual selection method, after iterating through the model improvement process, we ended up with the cutoffs of fewer than 6 NAs and mean variation greater than 0.4 between at least two of the tumors. The following variables fit that criteria: Size, Coronal_location, Conal_Location, Axial_location, Effect_on_globe, T1_regionality, T2_Intensity, T2_homogeneity, and T2_regionality.

For the Fisher's Exact Tests, we got the following results.

| | Conal_Location | | |
|---|---|---|---|
| Dx | 0 | 1 | 2 |
| 1 | 0 | 2 | 10 |
| 2 | 5 | 1 | 1 |
| 3 | 20 | 4 | 3 |

Above is a contingency table between the tumor type (Dx) and Conal_location. This is one example of the contingency tables used for the tests.

| Variables | Fisher_pvalue | sft_schwannoma | sft_cvm | schwannoma_cvm |
|---|---|---|---|---|
| Shape | 0.5279 | 0.8000 | 0.7460 | 0.7460 |
| Effect_on_globe | 0.6923 | 0.8720 | 0.8720 | 0.8720 |
| Conal_location | 0.0000 | 0.0012 | 0.0000 | 1.0000 |
| Coronal_location | 0.1720 | 0.5300 | 0.3360 | 0.3360 |
| Axial_location | 0.2464 | 0.2100 | 0.2100 | 0.6540 |
| T1_Intensity | 0.0206 | 0.5280 | 0.0272 | 0.4460 |
| T2_Intensity | 0.0000 | 0.0747 | 0.0000 | 0.2000 |
| T2_Homogeneity | 0.0009 | 1.0000 | 0.0096 | 0.0096 |
| T2_Regionality | 0.0000 | 0.1970 | 0.0000 | 0.0000 |
| MRI_Earlycon_Intensity | 0.0008 | 1.0000 | 0.0086 | 0.0086 |
| MRI_Earlycon_Pattern | 0.0000 | 1.0000 | 0.0000 | 0.0001 |
| MRI_Earlycon_Regionality | 0.0014 | 1.0000 | 0.0085 | 0.0134 |

Using two methods of the Fisher's Exact Tests, we created the table above, which shows the overall significance of each variable (2nd column) and p-value of individual variable pairings (columns 3-5). We found that Conal_location, T1_Intensity, all T2 variables, and all MRI_Earlycon variables were significant, which confirmed the significance of most of the manual selection variables.

Taking the manual selection variables, we added two of the doctors' recommended variables, MRI_EarlyCon_Intensity and MRI_LateCon_Intensity. We ended up with the following list of variables to work with: Size, Coronal_location, Conal_Location, Axial_location, Effect_on_globe, T1_regionality, T2_Intensity, T2_homogeneity, T2_regionality, MRI_EarlyCon_Intensity, and MRI_LateCon_Intensity.

## 4.1 KNN

The KNN algorithm ended up not being compatible with our dataset. It is not ideal for categorical data (which most of our data is) or imbalanced data (our data has a much larger proportion of CVM tumors than SFT and schwannoma tumors). Thus, our performance with

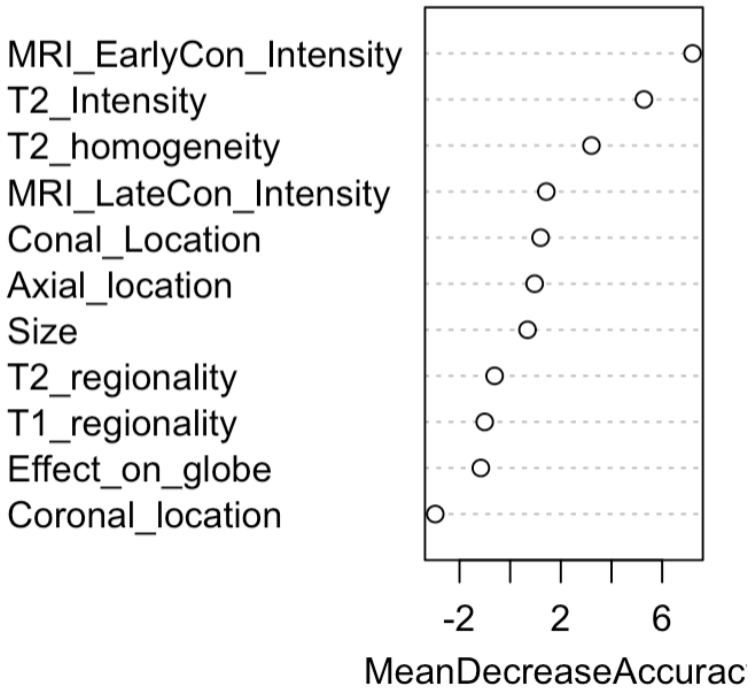
KNN models was inferior to the performance of our random forest models, and we decided not to move forward with testing prediction accuracy for them.

## 4.2 Random Forest

We created an initial random forest model with the variables listed above. We trained this model on 80% of the data and then tested it on the remaining 20%.

| | **Predicted** | | |
|---|---|---|---|
| **Actual** | SFT | Schwannoma | CVM |
| SFT | 2 | 0 | 0 |
| Schwannoma | 0 | 0 | 1 |
| CVM | 0 | 0 | 4 |

The confusion matrix above shows that our first random forest model had a prediction accuracy of 85.7%. CVM and SFT tumors are correctly identified every time, but one Schwannoma tumor was classified as a CVM mistakenly.

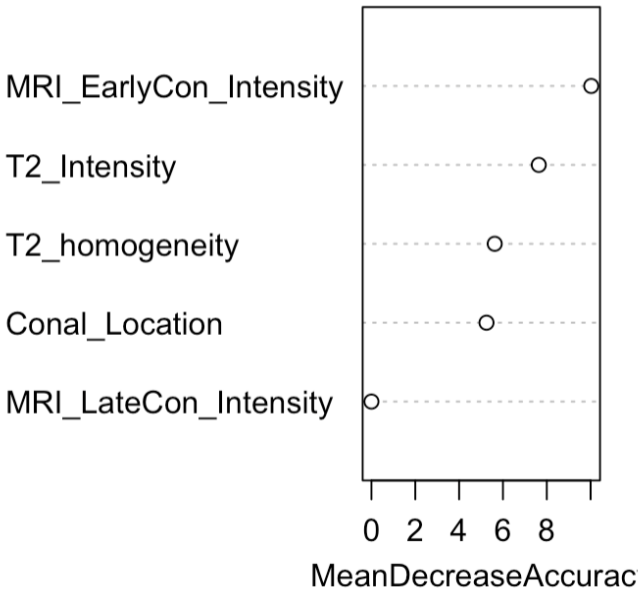



The variable importance plot from our random forest model is shown above. Variables towards the top are identified as more important by the random forest model when it comes to classifying the tumors. Important variables are those that are able to split the data into groups that lead to higher proportions of correct classifications during model training.

Next, we ran a second random forest model, which only included the top 5 most important variables from our previous model: MRI_EarlyCon_Intensity, T2_Intensity, T2_homogeneity, MRI_LateCon_Intensity, and Conal_Location. We wanted to simplify our model to include as few variables as possible while still maintaining a high level of accuracy. This makes our model computationally more efficient and conserves more data (when it comes to dropping observations due to NAs), which can give us a better shot at training our models to perform classification even better.

| | **Predicted** | | |
|---|---|---|---|
| **Actual** | SFT | Schwannoma | CVM |
| SFT | 0 | 2 | 0 |
| Schwannoma | 0 | 1 | 0 |
| CVM | 0 | 0 | 4 |

The confusion matrix above shows that our second random forest model had a prediction accuracy of 71.4%. All CVM and Schwannoma tumors are correctly identified. Although we do have two SFTs that are mistaken for Schwannomas, we still believe that this model is performing well given the small sample size. We believe that if we were able to test on a larger number of observations (as 20% of our data leaves us with only 7 observations to test on), that we would see a more stable and better classification accuracy.




The variable importance plot above is for our second model. Again, variables towards the top are identified as the most significant predictors of tumor type.

## 4.3 Cross-Validation and Bootstrapping

The tables below summarize the results from performing cross-validation and bootstrapping for the two random forest models:

| Dataset 1, Subset 1 | Accuracy | 90% CI Lower Bound | 90% CI Upper Bound |
|---|---|---|---|
| 10-Fold CV | 72.5% | – | – |
| LOOCV | 76.5% | – | – |
| Traditional Bootstrap | 72.5% | 50.0% | 91.7% |
| Pseudo-Bootstrap | 90.0% | 78.6% | 100.0% |

| Dataset 2, Subset 1 | Accuracy | 90% CI Lower Bound | 90% CI Upper Bound |
|---|---|---|---|
| 10-Fold CV | 85.8% | – | – |
| LOOCV | 83.8% | – | – |
| Traditional Bootstrap | 77.6% | 58.3% | 92.9% |
| Pseudo-Bootstrap | 91.7% | 78.6% | 100.0% |

| Dataset 1, Subset 2 | Accuracy | 90% CI Lower Bound | 90% CI Upper Bound |
|---|---|---|---|
| 10-Fold CV | 80.0% | – | – |
| LOOCV | 82.9% | – | – |
| Traditional Bootstrap | 80.9% | 58.3% | 100.0% |
| Pseudo-Bootstrap | 89.0% | 71.4% | 100.0% |

| Dataset 2, Subset 2 | Accuracy | 90% CI Lower Bound | 90% CI Upper Bound |
|---|---|---|---|
| 10-Fold CV | 78.3% | – | – |
| LOOCV | 81.6% | – | – |
| Traditional Bootstrap | 73.4% | 53.3% | 92.3% |
| Pseudo-Bootstrap | 85.5% | 71.4% | 100.0% |

Notes about the tables:

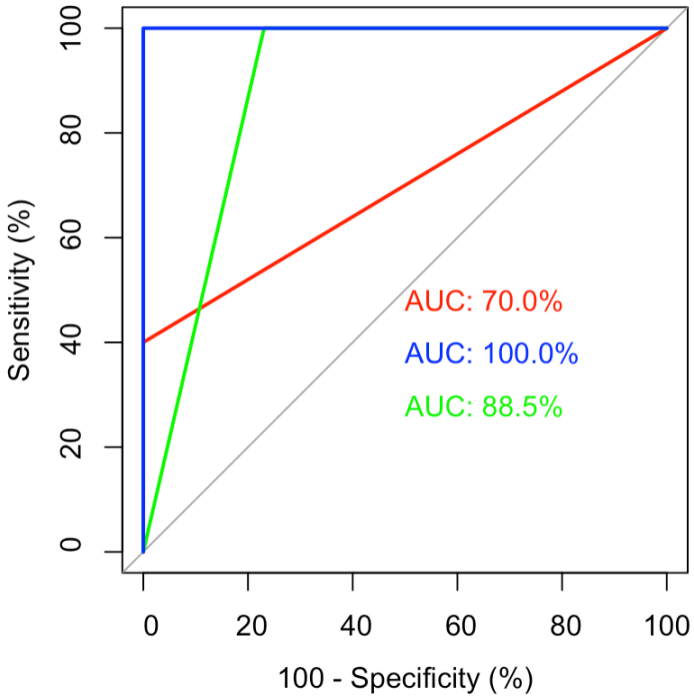
1. Dataset 1 corresponds to the old, original dataset given to us
2. Dataset 2 corresponds to the newer, reclassified dataset with a couple of new observations
3. Subset 1 contains the variables from the first random forest model: MRI_EarlyCon_Intensity, T2_Intensity, T2_homogeneity, MRI_LateCon_Intensity, Conal_Location, Axial_Location, Size, T2_regionality, T1_regionality, Effect_on_globe, and Coronal_location
4. Subset 2 contains the variables from the second random forest model: MRI_EarlyCon_Intensity, T2_Intensity, T2_homogeneity, MRI_LateCon_Intenity, and Conal_Location
5. CV = Cross-validation; LOOCV = Leave-one-out cross-validation; CI = Empirical confidence interval

Regardless of which subset of variables or which dataset is used, these results show that the random forest models classify orbital tumors well, despite the relatively small sample size. The true prediction accuracy of the models tends to be in the 70-80% range depending on which parameters are used. Perhaps the most reliable indicator of prediction accuracy is the traditional bootstrap, as it uses 20,000 iterations and has a testing set that does not overlap with the boot sample. Looking at the traditional bootstrap specifically, we can see that mean prediction accuracy ranges from 72.5% to 80.9%. It is highest for the original dataset using the second

random forest model, and lowest for the original dataset using the first random forest model. Overall, these results verify the usefulness of our random forest models for predicting orbital tumors. They also provide rough confidence intervals that suggest the range of possible performances that the models could have on future data.

## 4.4 ROC Curves

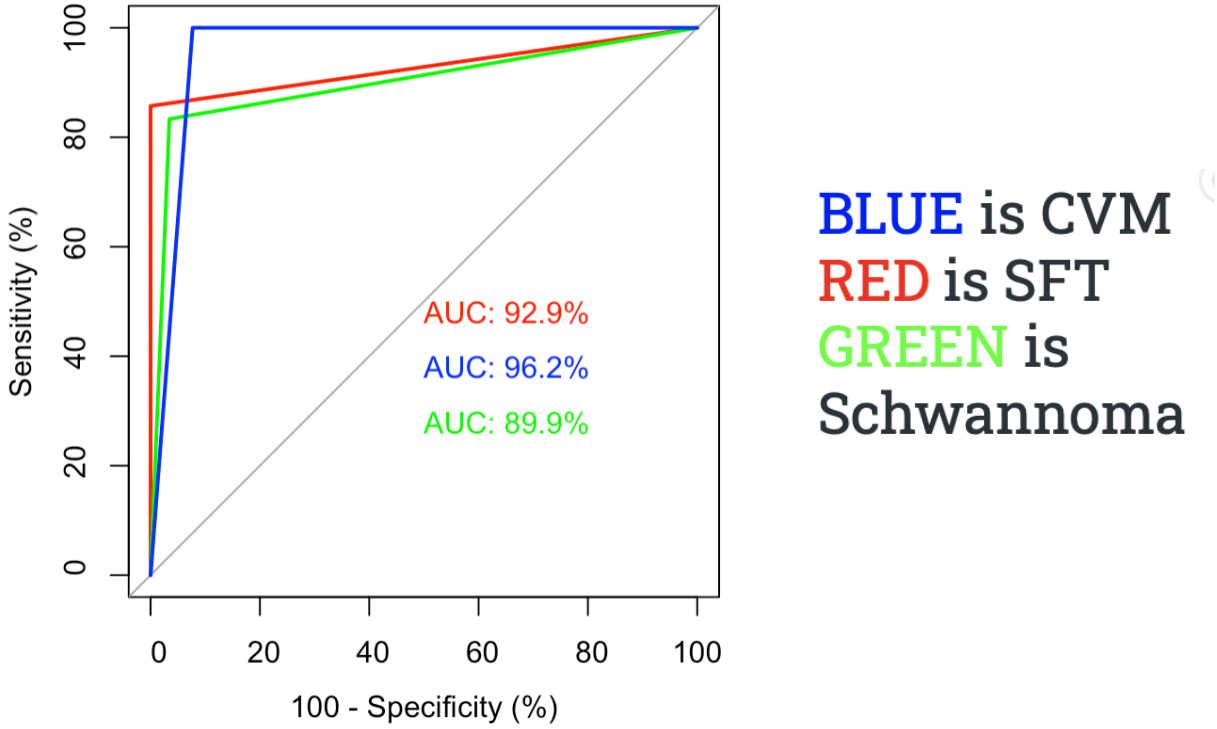
Our first ROC curve is for our random forest model with 5 variables (trained on 60% of the data and tested on the remaining 40%). This is the traditional method for testing a model — the model is trained on a portion of the data and tested on the remaining portion.




BLUE is CVM
RED is SFT
GREEN is Schwannoma

As shown above, CVM tumors are classified extremely well by our model. SFT and Schwannoma tumors are classified well but not to the same degree as the CVM (for reasons discussed previously).

Our second ROC curve is for our random forest model with 5 variables trained on the full data and testing on the full data. **NOTE** that using all of the data for both training and testing is a highly unconventional method for model testing. However, since our sample size is so small, the ROC curve looks more linear and angular with smaller sample sizes, so we decided to use the full data set. This can also give us an idea or where we might be able to go in terms of accuracy if we had more data to work with.




Overall, we see similar results as before but with a higher degree of accuracy for SFT and Schwannoma. This is likely due to the fact that all of the data was used during training and testing, so we had more observations to train and test the model on.

## 4.5 Interpretation

Overall, the results from the random forest model, cross-validation, and bootstrapping lead us to believe that the second random forest model performed well given the small sample size. We consistently saw CVM tumors being classified correctly 100% of the time (or very close to 100%). We believe that a lot of the predictive power of our model when it comes to CVMs is due to the fact that our data is almost entirely made up of CVMs. The more observations of a specific tumor type our model has to train on, the better it will perform. When it comes to SFTs and Schwannomas, the models weren't classifying as many correctly. We believe that this is due to the fact that there are fewer observations of those two tumor types in our data. This issue is heightened by the fact that we are using 80% of our data to train the model and saving the other 20% for testing. Since we are splitting the data in this way and since we have few SFT and Schwannoma tumors in our data to begin with, we are often left with a training/testing set with few/no observations of a certain tumor, making our predictive power much lower and more

variable with different splits of the data (which are randomly generated). In the future, we will see this model perform better and better as it is fed more data. Our criteria for important variables will also likely change with additional data. Overall, we can say that our random forest models are performing pretty well and could certainly be a tool in the diagnostic process, but we do not recommend that it be the *only* tool involved in the diagnostic process.

# 5 Conclusion

This report sought to achieve two goals: first, accurately classify orbital tumor type between cavernous venous malformations (CVM), solitary fibrous tumors (SFT), and orbital schwannomas, and second, determine the relevant features used in the prediction model. All of our models were able to accurately classify CVM tumors across both datasets, but SFT and Schwannoma classifications were less accurate due to small sample size. The best prediction accuracy was obtained with the first random forest model and the updated dataset. As such, we found that the best predictors for tumor classification are Size, Coronal_location, Conal_Location, Axial_location, Effect_on_globe, T1_regionality, T2_Intensity, T2_homogeneity, T2_regionality, MRI_EarlyCon_Intensity, and MRI_LateCon_Intensity. We constructed frequency tables to give an idea of each variable's distribution across the different tumor types; please see the Appendix section below.

## 5.1 Challenges

The data's small sample size, imbalanced classes, and abundance of missing data presented us with various challenges and limitations.

To begin with, we were not able to use subset selection procedures to identify a subset of the predictors that are related to the orbital tumor diagnosis because of the large amount of missing data. To use subset selection procedures in R, we would have to remove the observations with missing values, which would leave us with zero observations in the end. As we selected our variables for our models, we had to keep in mind that including more variables meant that we had to impute more observations with missing values. As a result, we developed our variable selection process with the intent to include as many important variables as possible without sacrificing too many observations.

In addition, our small sample size and imbalanced classes limited the number of models we could explore. We could not use logistic regression because SFT and Schwannoma had a frequency of 0 for certain levels of the variables we determined to be important. Our model performance with the KNN algorithm was also disadvantaged by the imbalanced data and the vast amount of categorical data. Hence, we focused on using random forest models as they gave us the best results.

Lastly, our prediction accuracy was highly variable due to our small sample size. When we wanted to measure our model performance by splitting our data into training and test sets, we would often find that our prediction accuracy fluctuated quite widely with different seeds. In some instances, we would find that we did not have enough SFT or schwannoma observations in the training or test set and this would consequently affect our model performance as well. Thus

with random forest, we used bootstrapping as a way to get a better estimate of our true model performance.

## 5.2 Recommendations

Based on these findings, we recommend the following three action items. Firstly, continue with the second-generation classification of T2 variables, as provided in Dataset 2. The reclassification of such variables (especially T2 Intensity and T2 Homogeneity) significantly increased the bootstrap prediction accuracy confidence intervals, as shown in section 4.3. Second, researchers should look to increase the sample size of SFT and schwannoma labels in particular. All models across both datasets predicted CVM tumors with nearly 100% accuracy, but struggled with predicting SFT and schwannoma tumors. However, when the sample sizes were increased and various bootstrapping and resampling methods were used (some more unorthodox than others), prediction accuracy of SFT and schwannoma tumors increased. This suggests that the primary limiting factor of the dataset was its size, not a lack of distinction between tumors. Finally, researchers should seek to reduce NAs in key variables.

# 6 Appendix

The distribution for the size variable was shown in the Exploratory Data Analysis section.

Coronal Location

| | Supero-temporal | Infero-temporal | Superonasal | Inferonasal | Panorbital |
|---|---|---|---|---|---|
| SFT | 22.2% | 11.1% | 44.4% | 11.1% | 11.1% |
| Schwannoma | 28.6% | 14.3% | 14.3% | 42.9% | 0% |
| CVM | 19.0% | 14.3% | 9.5% | 23.8% | 33.5% |

Conal Location

| | Intraconal | Extraconal | Both |
|---|---|---|---|
| SFT | 0% | 11.1% | 88.9% |
| Schwannoma | 71.4% | 14.3% | 14.3% |
| CVM | 76.2% | 14.3% | 9.5% |

Axial Location

| | Middle | Posterior | Anterior | Lid | Panorbital |
|---|---|---|---|---|---|
| SFT | 0% | 55.6% | 0% | 33.3% | 11.1% |
| Schwannoma | 0% | 42.9% | 42.9% | 14.3% | 0% |
| CVM | 9.5% | 23.8% | 28.6% | 28.6% | 9.5% |

Effect on Globe

| | Indentation | Moulding | No Effect |
|---|---|---|---|
| SFT | 33.3% | 0% | 66.7% |
| Schwannoma | 14.3% | 0% | 85.7% |
| CVM | 33.5% | 4.8% | 61.8% |

T1 Regionality

| | Homogenous | Biphasic | Regional | Inhomogeneous |
|---|---|---|---|---|
| SFT | 66.7% | 22.2% | 0% | 11.1% |
| Schwannoma | 71.4% | 28.6% | 0% | 0% |
| CVM | 100% | 0% | 0% | 0% |

T2 Intensity

| | Hypo | Iso | Hyper |
|---|---|---|---|
| SFT | 11.1% | 55.6% | 33.3% |
| Schwannoma | 0% | 14.3% | 85.7% |
| CVM | 0% | 0% | 100% |

T2 Homogeneity

| | Homogenous | Heterogeneous |
|---|---|---|
| SFT | 22.2% | 77.8% |
| Schwannoma | 28.6% | 71.4% |
| CVM | 85.7% | 14.3% |

T2 Regionality

| | Homogenous | Biphasic | Regional | Inhomogeneous |
|---|---|---|---|---|
| SFT | 22.2% | 0% | 22.2% | 55.6% |
| Schwannoma | 14.3% | 28.6% | 0% | 57.1% |
| CVM | 100% | 0% | 0% | 0% |

MRI EarlyCon Intensity

| | Moderate | Maximal |
|---|---|---|
| SFT | 22.2% | 77.8% |
| Schwannoma | 14.3% | 85.7% |
| CVM | 81.0% | 19.0% |

MRI LateCon Intensity

| | Moderate | Maximal |
|---|---|---|
| SFT | 22.2% | 77.8% |
| Schwannoma | 14.3% | 85.7% |
| CVM | 19.0% | 81.0% |