**Introduction**

This project is a follow up to the research paper “Reducing Classification Cost through Strategic Annotation Assignment”. On that paper, it was found that there is a method to distinguish between easy and hard cases to label. Easy cases require less expert opinions and hence cost less to label. On the other hand, hard cases require more expert opinions and hence cost more to label.

This project goes a step further and tries to understand why some nodules are considered easy or hard to label by a data mining algorithm. The underlying features are analyzed and the goal is to identify image feature properties and decision tree algorithm analysis that explain the difficulty of labeling a nodule.

The nodules are analyzed from the following perspective:

1. Easy/Hard labels
2. Malignancy level
3. Agreement/Disagreement among radiologists
4. Features

There are 42 image features (*the GaborMean and GaborSD features were calculated as an average of their original values on the previous paper*) grouped into four types

1. **Shape Features (9):**

* Circularity, roughness, elongation, compactness, eccentricity, solidity, extent, RadialDistanceSD, SecondMoment

1. **Intensity Features(9):**

* MinIntensity, MaxIntensity, MeanIntensity, SDIntensity, MinIntensityBG, MaxIntensityBG, MeanIntensityBG, SDIntensityBG, IntensityDifference

1. **Size Features (6):**

* Area, ConvexArea, Perimeter, EquivDiameter, MajorAxisLenght, MinorAxisLenght

1. **Texture Features (18):**

* Haralick (11):Contrast, Correlation, Entropy, Energy, Homogeneity, 3rdOrderMoment, InverseVariance, SumAverage, Variance, ClusterTendency, MaximumProbability)
* Gabor (2): GaborMean, GaborSD
* Markov (5): Markov1, Markov2, Markov3, Markov4, Markov5

**Data set**

810 Cases

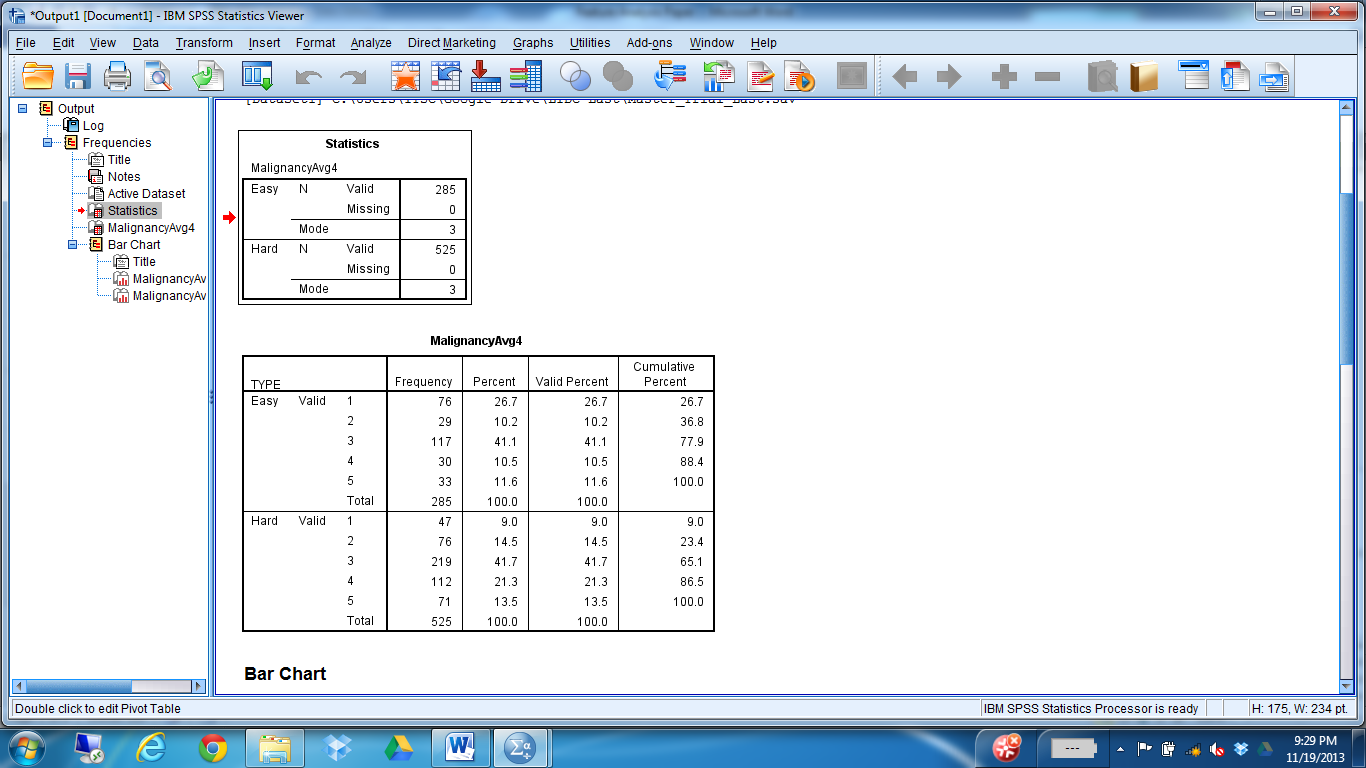
64 Image Features

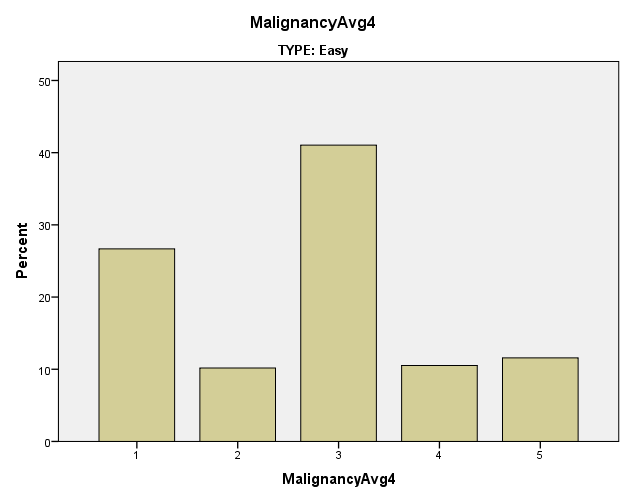
Type (0-easy or 1-hard)- LABEL

Additional Data:

* Instance ID to locate nodules
* 4 radiologists malignancy ratings
* MalAvg4 (Average rating of nodule to determine overall malignancy)

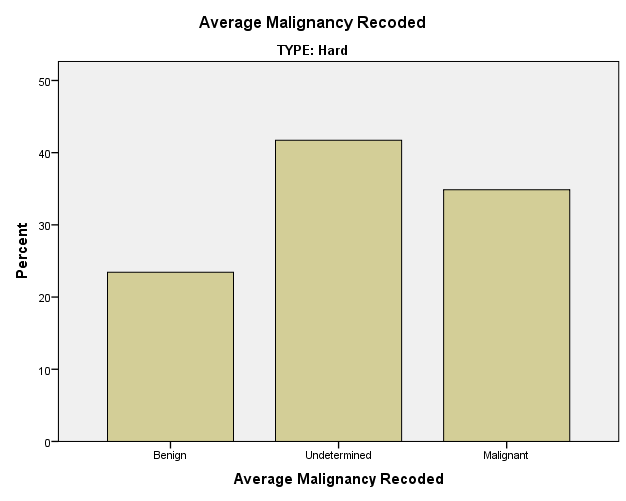
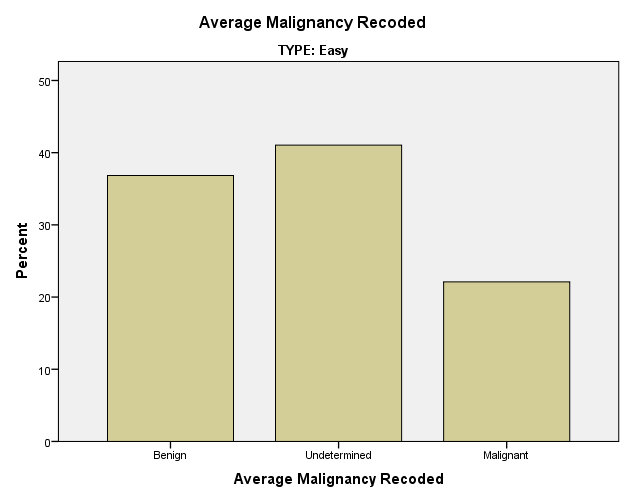
Data set description continued:





Recode MalginancyAvg4 into 3 different values: Benign (1), Undetermined (2), Malignant (3) **[IMPORTANT: MalAvg4 is being recoded into MalAvg4\_Recoded which is the average of the 4 ratings, NOT the rating of Radiologist 4. However, to measure agreement/disagreement between the radiologists further in the analysis, the individual radiologists (i.e., Malignancy\_1, Malignancy\_2,…) ratings are recoded and NOT the MalAvg ratings since we are looking for agreement/disagreement among radilogists].**

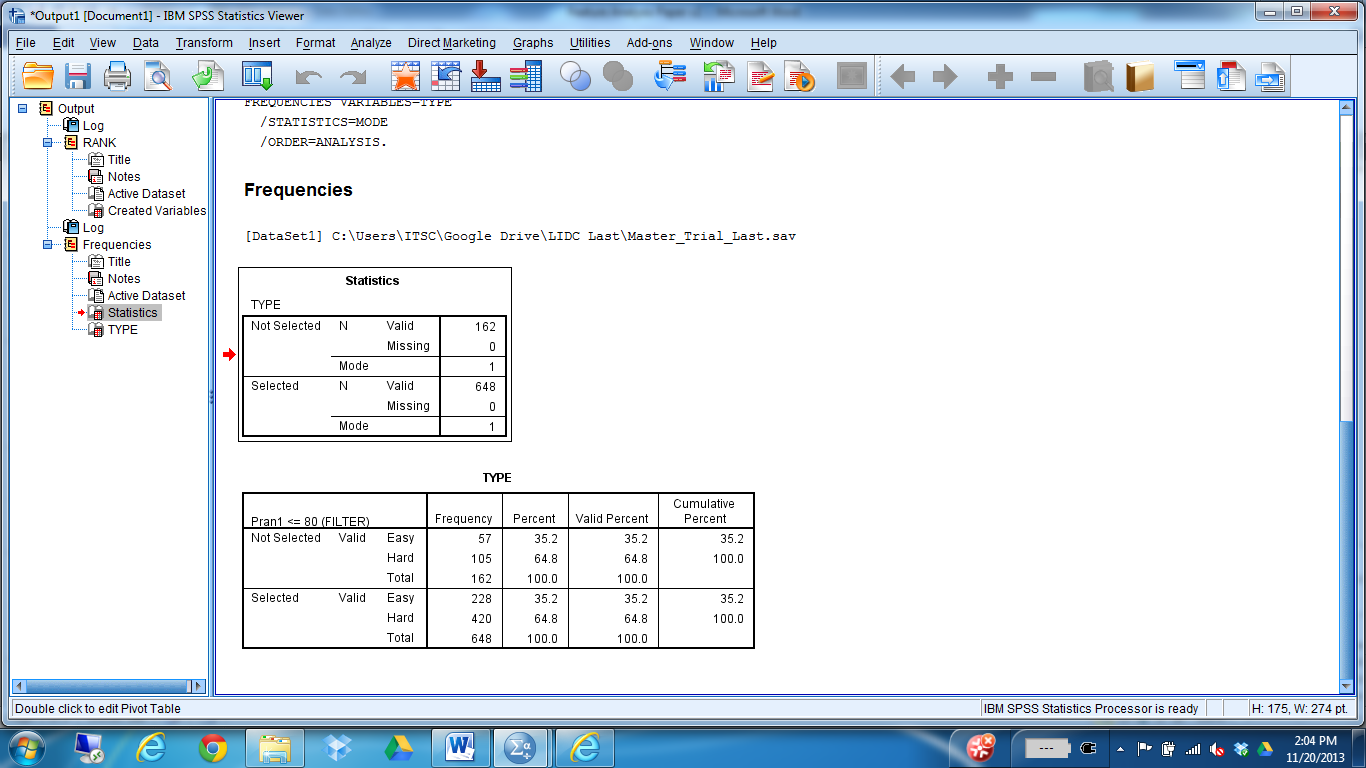




**TRAINING/TESTING SPLIT**

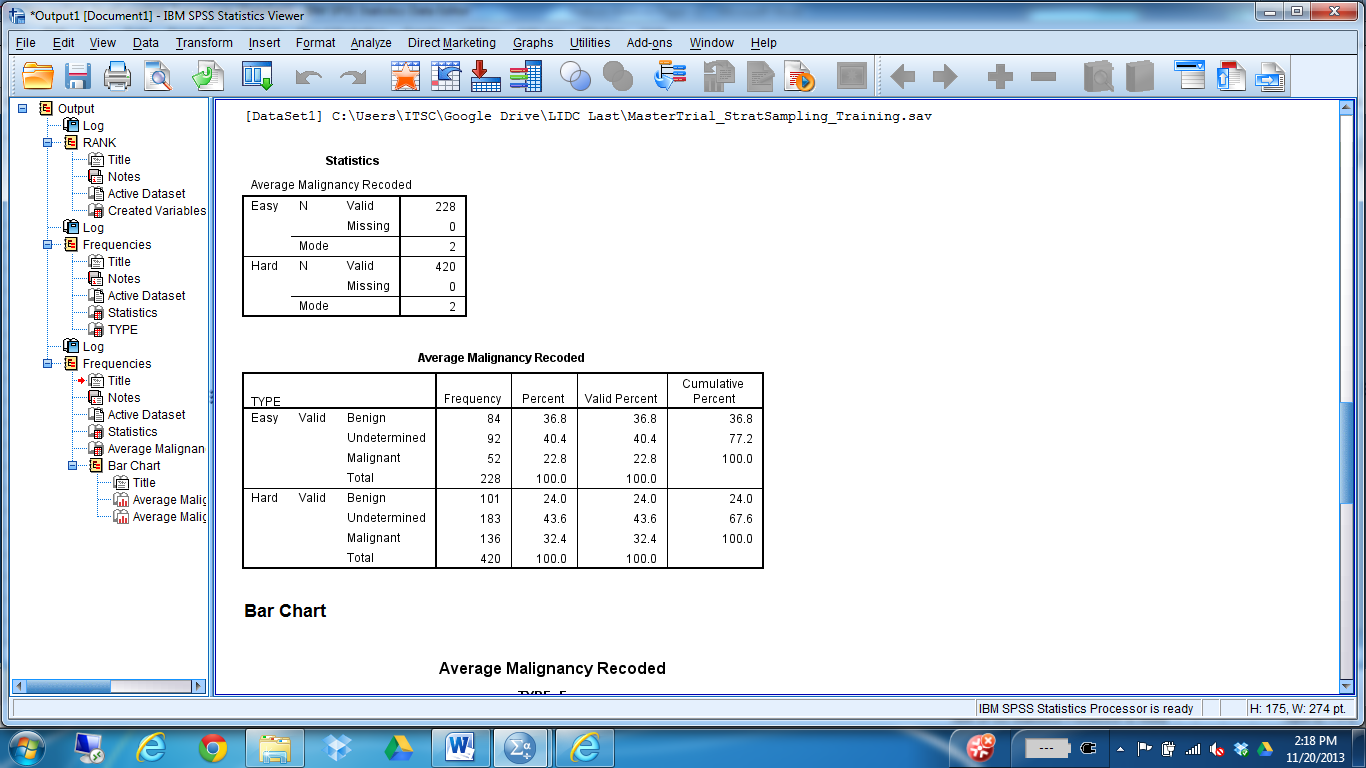
The data was then split into training (80%, 648 cases) and testing (20%, 162 cases). In order to keep the same label distribution in the training and testing sets stratified sampling was done.

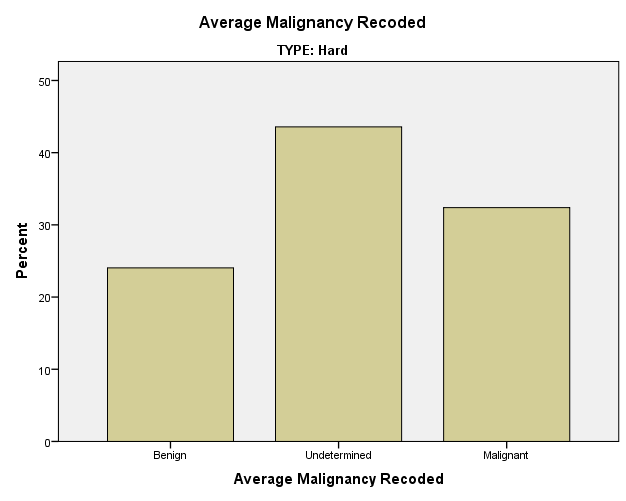
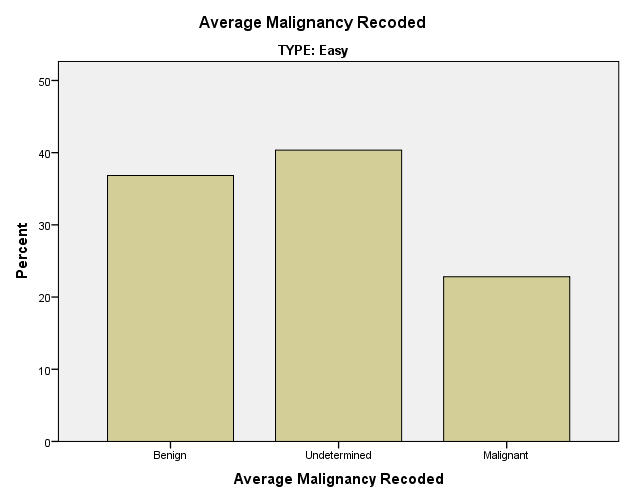
C:\Users\ITSC\Google Drive\LIDC Last\Master\_Trial\_StratSampling.spss



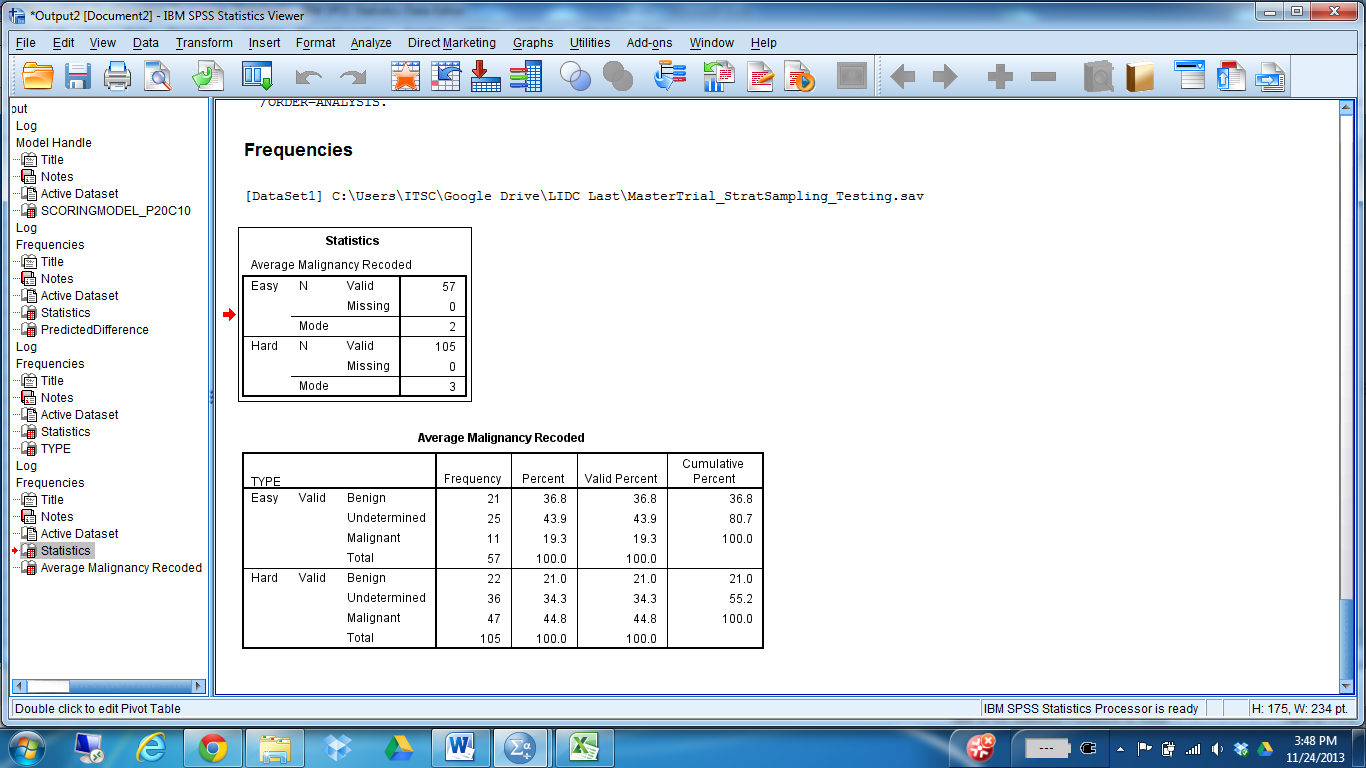
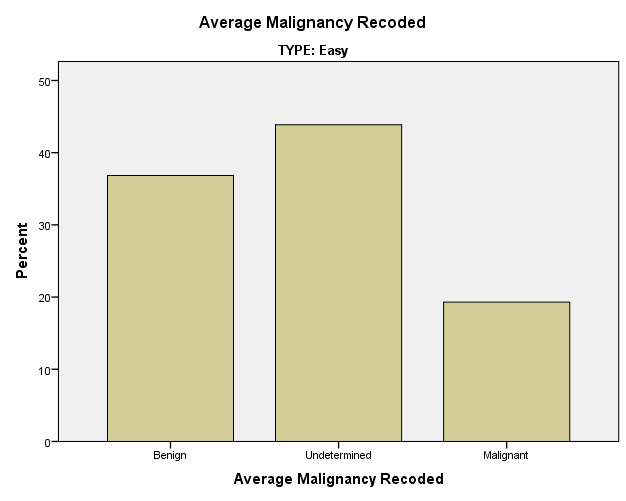
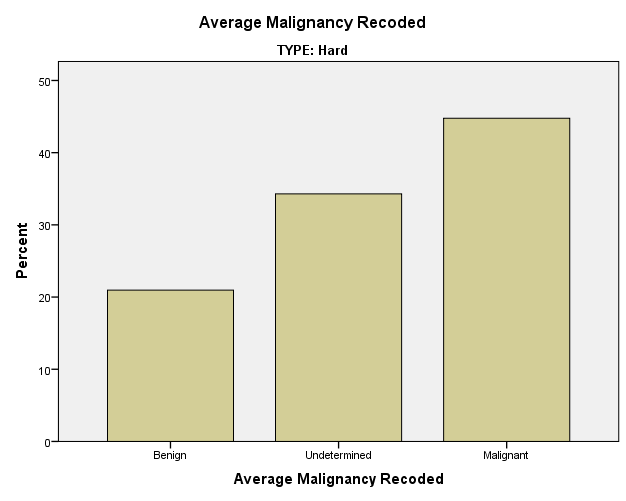
**TRAINING DATA**

There are 228 easy cases (approx. 35%) and 420 hard cases (approx. 65%), same distribution as the whole data set. Also, the malignancy labels are distributed similarly to the entire data set.





**TESTING DATA**

There are 57 easy cases (approx. 35%) and 105 hard cases (approx. 65%), same distribution as the whole data set. Also, the malignancy labels are distributed similarly to the entire data set.   
 

**RESULTS**

*Training Model Configuration*

Crossvalidation folds of 10

CRT Tree / Parent 20 Child 10 / Gini Index

\*NOTE: The same data sets (Training/Testing) were used for these 4 experiments.

*Accuracy Results Original Data Set (285 Easy/525 Hard)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 24/Child 12** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 58.8 | 87.6 | 77.5 |
| Testing | 38.6 | 84.8 | 68.5 |
| Difference | 20.2 | 2.8 | 9.0 |

*Accuracy Results Original Data Set (285 Easy/525 Hard)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 22/Child 11** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 58.8 | 87.6 | 77.5 |
| Testing | 38.6 | 84.8 | 68.5 |
| Difference | 20.2 | 2.8 | 9.0 |

*Accuracy Results Original Data Set (285 Easy/525 Hard)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 20/Child 10** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 61.8 | 86.9 | 78.1 |
| Testing | 40.4 | 81.9 | 67.3 |
| Difference | 21.4 | 5.0 | 10.8 |

*Accuracy Results Original Data Set (285 Easy/525 Hard)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 18/Child 9** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 71.5 | 83.1 | 79.0 |
| Testing | 43.9 | 76.2 | 64.8 |
| Difference | 27.6 | 6.9 | 14.2 |

Accuracy Results using most important features TO DO: Keep this one but try different parameters and see which one gives the best accuracy (P22C11, etc….)

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 20/Child 10** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 77.6 | 83.8 | 81.6 |
| Testing | 50.9 | 70.5 | 62.8 |
| Difference | 26.7 | 13.3 | 18.8 |

**Accuracy Results Original Data Set Important features with Config 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 22/Child 11** | **Easy** | **Hard** | **Overall** |
| **Cross Validation** | **56.6** | **90.2** | **78.4** |
| **Testing** | **36.8** | **88.6** | **70.4** |
| **Difference** | **19.8** | **1.6** | **8.0** |

Accuracy Results Original Data Set Important features with Config 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 24/Child 12** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 56.6 | 90.2 | 78.4 |
| Testing | 36.8 | 88.6 | 70.4 |
| Difference | 19.8 | 1.6 | 8.0 |

Results Balanced Data Sets (285 Easy/285 Hard) TO DO: Capture results using the most important feature, optimal tree configuration.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent 24/Child 12** | **Easy** | **Hard** | **Overall** |
| Cross Validation | 72.8 | 85.5 | 79.2 |
| Testing | 54.4 | 70.2 | 62.3 |
| Difference | 18.4 | 15.3 | 16.9 |

~~Accuracy Results Balanced/Normalized Data Sets~~ VERY LOW TESTING ACCURACY, double check.

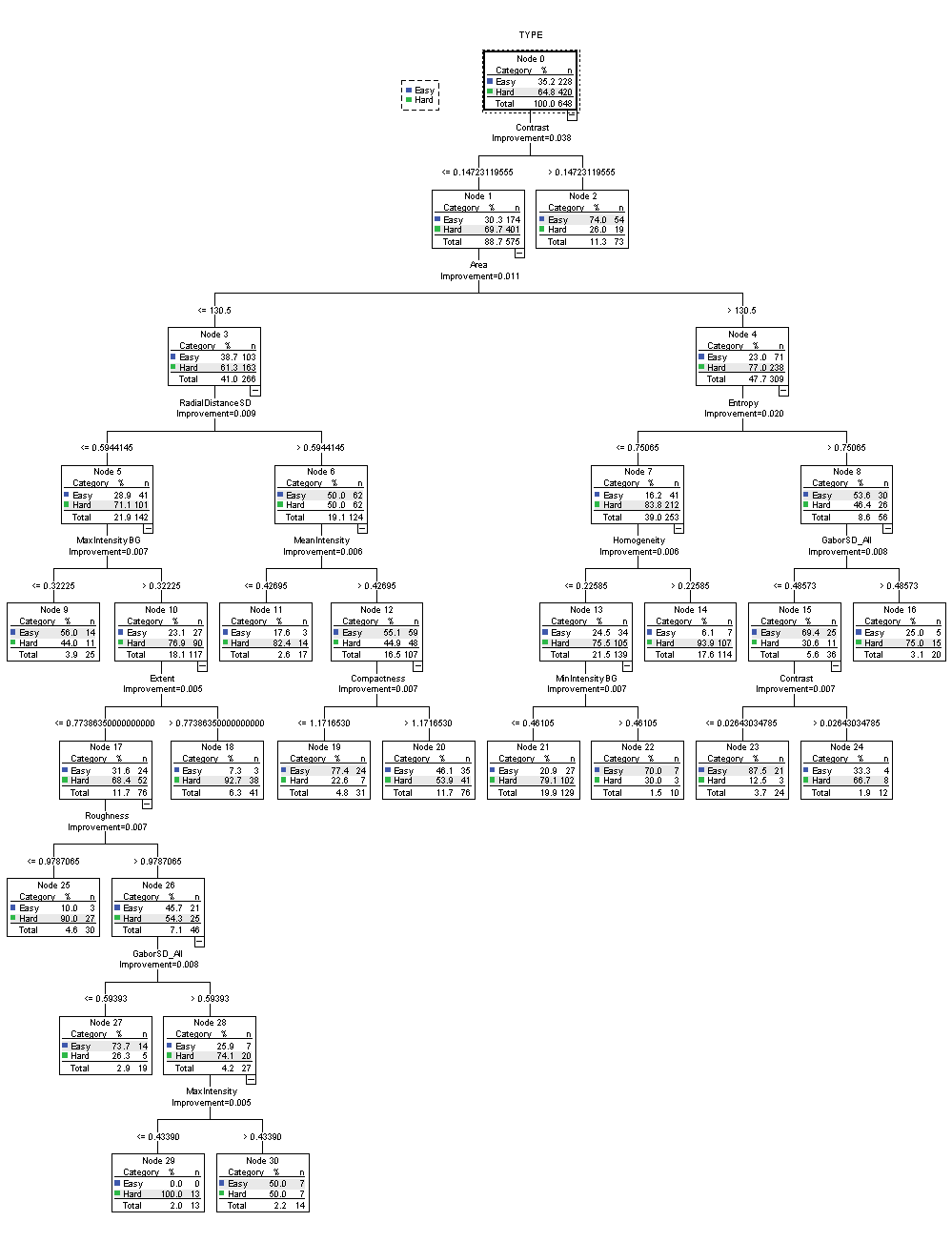
|  |  |  |  |
| --- | --- | --- | --- |
| **~~Parent 18/Child 9~~** | **~~Easy~~** | **~~Hard~~** | **~~Overall~~** |
| ~~Cross Validation~~ | ~~74.9~~ | ~~81.6~~ | ~~80.5~~ |
| ~~Testing~~ | ~~57.9~~ | ~~31.6~~ | ~~44.7~~ |
| ~~Difference~~ | ~~17.0~~ | ~~50.0~~ | ~~35.8~~ |

~~Accuracy Results Balanced/Normalized Data Sets Double check, parameters~~ **DROPPING THIS EXPERIMENT BECAUSE OF REDUNDANCY**

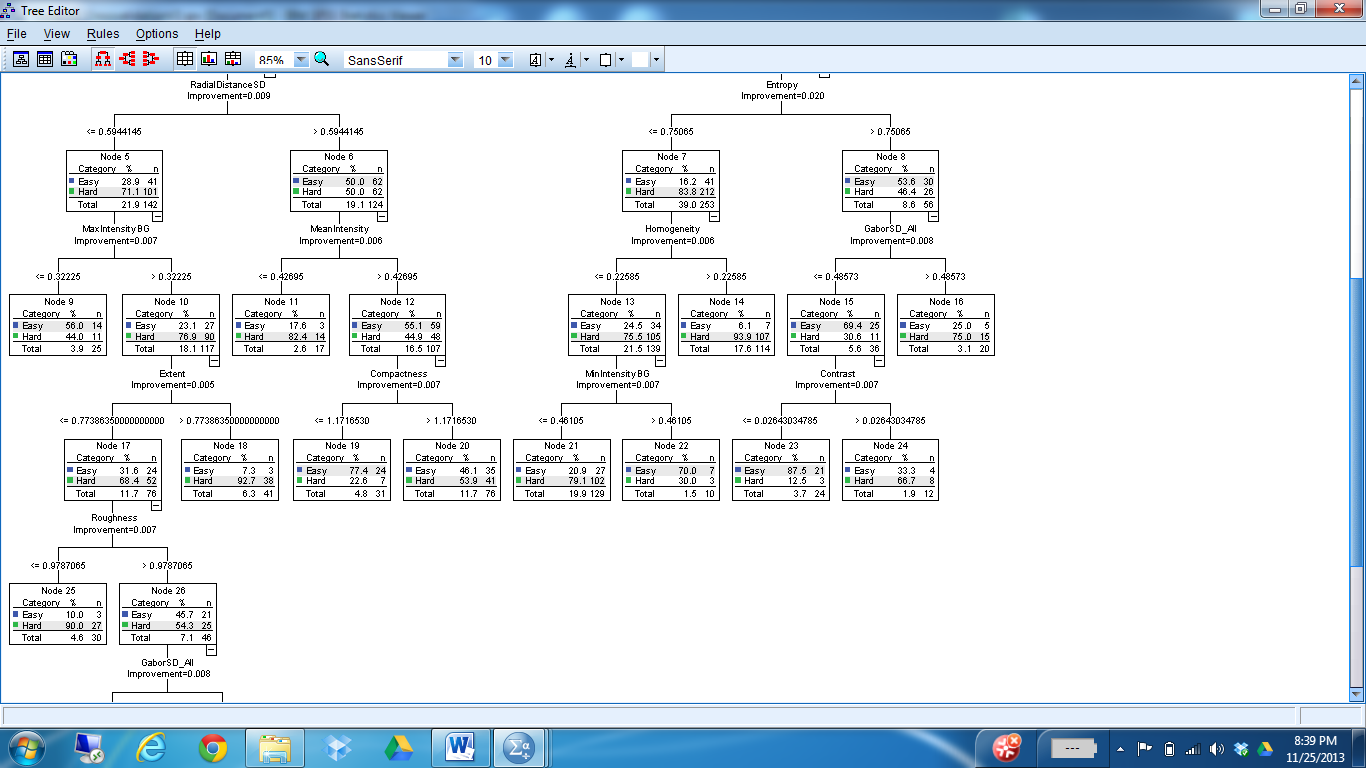
|  |  |  |  |
| --- | --- | --- | --- |
| **~~Parent 24/Child 12~~** | **~~Easy~~** | **~~Hard~~** | **~~Overall~~** |
| ~~Cross Validation~~ | ~~72.8~~ | ~~85.5~~ | ~~79.2~~ |
| ~~Testing~~ | ~~54.4~~ | ~~70.2~~ | ~~62.3~~ |
| ~~Difference~~ | ~~18.4~~ | ~~15.2~~ | ~~16.9~~ |

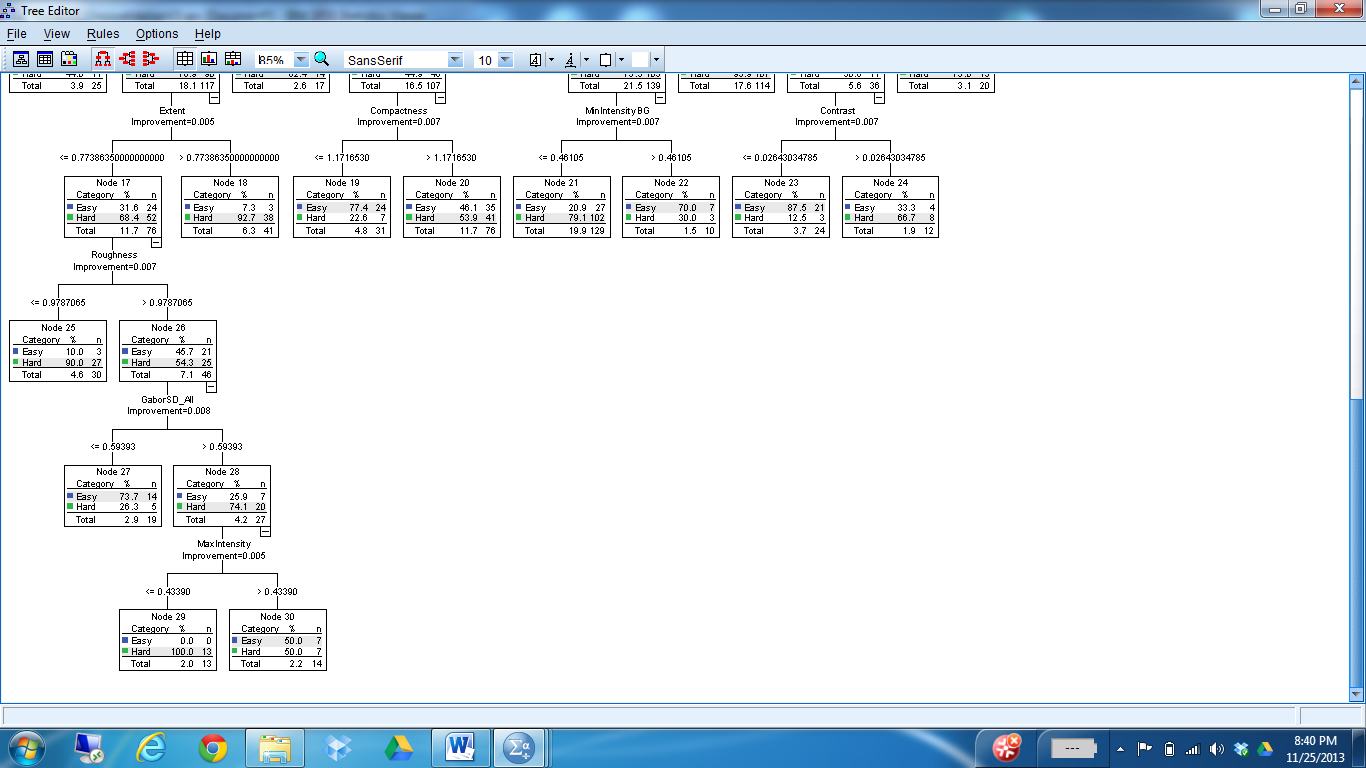
Important Tree Features across the three different experiments (the number in parenthesis represents the tree level at which the feature appears)

|  |  |  |
| --- | --- | --- |
| **Original Set (13 Feaures)** | **Balanced Set (13 Features)** | **~~Balanced/Normalized Set~~** |
| Contrast (1) | Perimeter (1) | ~~Perimeter(1)~~ |
| **Area (2)** | SDIntensity (2) | ~~SDIntensity (2)~~ |
| RadialDistanceSD (3) | **Entropy (2)** | **~~Entropy (2)~~** |
| **Entropy (3)** | SDIntensityBG (3) | ~~SDIntensityBG (3)~~ |
| **MaxIntensityBG (4)** | InverseVariance (3) | ~~InverseVariance (3)~~ |
| MeanIntensity (4) | IntensityDifference(3) | ~~IntensityDifference (3)~~ |
| Homogeneity (4) | MeanIntensityBG(4) | ~~MeanIntensityBG (4)~~ |
| **GaborSD (4)** | **MaxIntensityBG(4)** | **~~MaxIntensityBG (4)~~** |
| Extent (5) | **Area(4)** | **~~Area (4)~~** |
| Compactness (5) | Markov2 (5) | ~~Markov2 (5)~~ |
| MinIntensityBG (5) | MaxProbability (5) | ~~MaxProbability (5)~~ |
|  | **GaborSD (6)** | **~~GaborSD (6)~~** |
| Roughness (6) | Elongation(7) | ~~MaxIntensityBG (6)~~ |
|  |  | ~~Elongation (7)~~ |
| MaxIntensity (8) |  | ~~MaxIntensityBG(8)~~ |

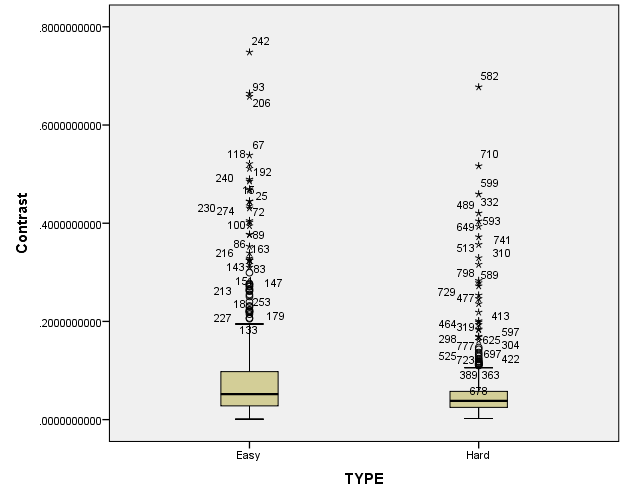




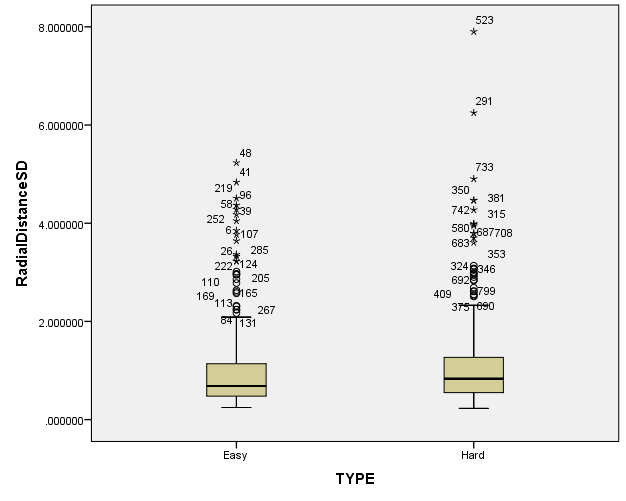
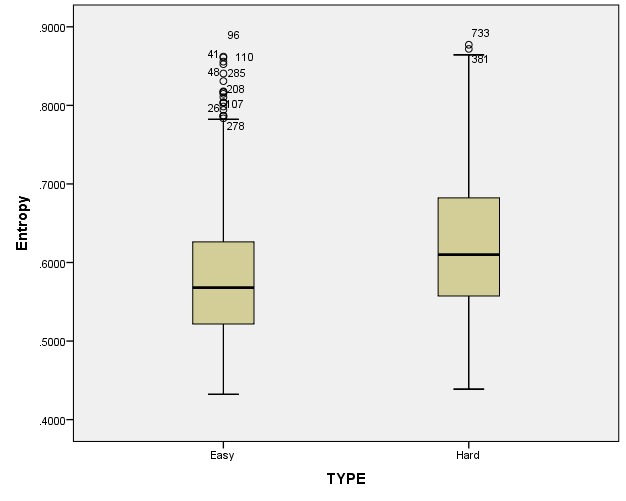




Contrast Area



Entropy RadialDistanceSD



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Terminal Node** | **Easy %** | **Hard %** | **Easy Cases** | **Hard Cases** | **Total Cases** | **Malignancy**  **Agreement** | **Malignacy Rating**  **Majority** |
| 2 | 74.0% | 26.0% | 54 | 19 | 73 | 0.20091 | Benign (95.9%) |
| 9 | 55.0% | 44.0% | 14 | 11 | 25 | 0.28667 | Undeter (68.0%) |
| 11 | 17.6% | 82.4% | 3 | 14 | 17 | 0.48529 | Benign (52.9%) |
| **14** | **6.1%** | **93.9%** | **7** | **107** | **114** | **0.60453** | **Malign (59.6%)** |
| 16 | 25.0% | 75.0% | 5 | 15 | 20 | 0.51250 | Malign (85.0%) |
| **18** | **7.3%** | **92.7%** | **3** | **38** | **41** | **0.53862** | **Undete (46.3%)** |
| 19 | 77.4% | 22.6% | 24 | 7 | 31 | 0.36559 | Undeter (71.0%) |
| 20 | 46.1% | 53.9% | 35 | 41 | 76 | 0.41996 | Undeter (68.4%) |
| 21 | 20.9% | 79.1% | 27 | 102 | 129 | 0.53036 | Undeter (45.7%) |
| 22 | 70.0% | 30.0% | 7 | 3 | 10 | 0.68333 | Malign (50.0%) |
| **23** | **87.5%** | **12.5%** | **21** | **3** | **24** | **0.34722** | **Malign (91.7%)** |
| 24 | 33.3% | 66.7% | 4 | 8 | 12 | 0.48611 | Malign (83.3%) |
| **25** | **10.0%** | **90.0%** | **3** | **27** | **30** | **0.42778** | **Undete (66.7%)** |
| 27 | 73.7% | 26.3% | 14 | 5 | 19 | 0.55263 | Undeter (57.9%) |
| **29** | **0.0%** | **100.0%** | **0** | **13** | **13** | **0.40385** | **Undete (61.5%)** |
| 30 | 50.0% | 50.0% | 7 | 7 | 14 | 0.34524 | Benign (57.1%) |

\*Malignancy agreement among a nodule is defined as the *average of the variances* across the 114 individual nodes. Each nodule has an agreement that is the variance across the 4 radiologist malignancy recoded ratings. The recoded ratings are: 1-Benign, 2-Undetermined, 3-Malignant.

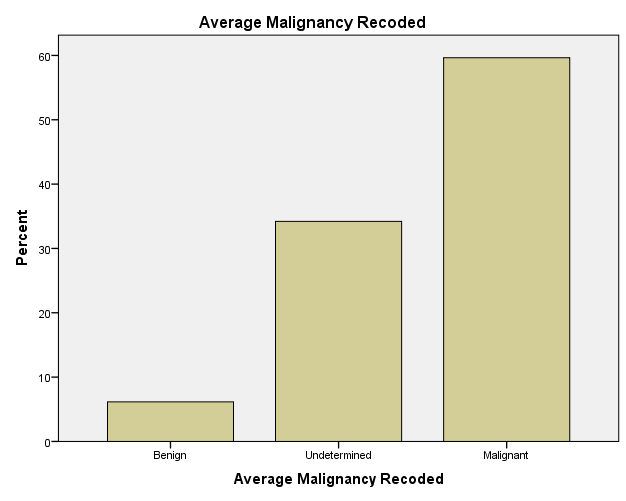
Average Agreement is 0.45898 for ALL 810 cases.

Agreement by Easy/Hard type:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | |
| TYPE | | N | Minimum | Maximum | Mean | Std. Deviation |
| Easy | Radiologists\_Agreement | 228 | .000 | 1.333 | .31579 | .335433 |
| Valid N (listwise) | 228 |  |  |  |  |
| Hard | Radiologists\_Agreement | 420 | .000 | 1.333 | .53671 | .348537 |
| Valid N (listwise) | 420 |  |  |  |  |

**NODE 14 Analysis**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Terminal Node | Easy % | Hard % | Easy Cases | Hard Cases | Total Cases | Agreement |
| 14 | 6.1% | 93.9% | 7 | 107 | 114 | 0.60453 |



**Sample Nodule Images for node 14 nodules**

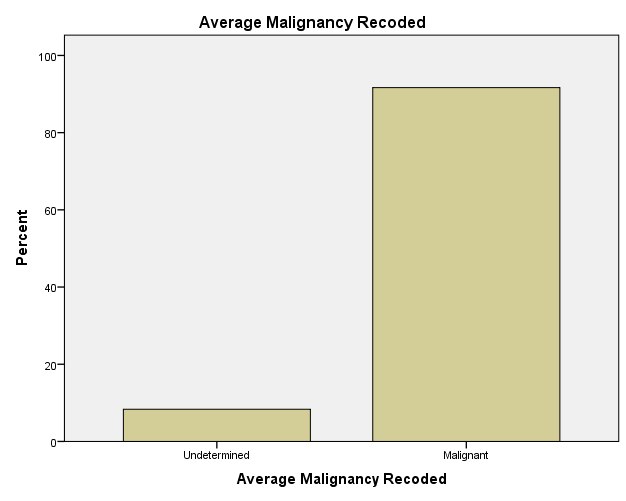
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NoduleID | Malignancy | Label | Agreement | Picture | Contrast | Area |
| 425 | Malignant | Hard | 0.917 |  | 0.0436 | 219 |
| 729 | Undetermined | Hard | 0.333 |  | 0.0238 | 278 |
| 1023 | Malignant | Hard | 0.917 |  | 0.0288 | 274 |
| 1202 | Malignant | Hard | 1.000 |  | 0.0204 | 147 |
| 1209 | Malignant | Hard | 0 |  | 0.0177 | 369 |
| 6278 | Undetermined | Easy | 0.250 |  | 0.0714 | 336 |
| 9020 | Malignant | Easy | 0 |  | 0.0260 | 228 |

Reference Statistics for all 810 cases



**NODE 23 Analysis**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Terminal Node | Easy % | Hard % | Easy Cases | Hard Cases | Total Cases | Malignancy  Agreement | Malignacy Rating  Majority |
| 23 | 87.5% | 12.5% | 21 | 3 | 24 | 0.34722 | Malign (91.7%) |



**Sample Nodule Images for node 23 nodules**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NoduleID | Malignancy | Label | Agreement | Picture | Contrast | Area |
| 666 | Malignant | Easy | 0.250 |  | 0.0191 | 709 |
| 695 | Malignant | Easy | 0 |  | 0.0144 | 483 |
| 2596 | Malignant | Easy | 1.000 |  | 0.0105 | 631 |
| 2761 | Undetermined | Hard | 0.667 |  | 0.0229 | 1056 |
| 28300 | Malignant | Easy | 0 |  | 0.0154 | 901 |
| 29824 | Malignant | Easy | 1.000 |  | 0.0147 | 1512 |
| 38473 | Undetermined | Easy | 0 |  | 0.0198 | 572 |

Reference statistics for all 810 cases

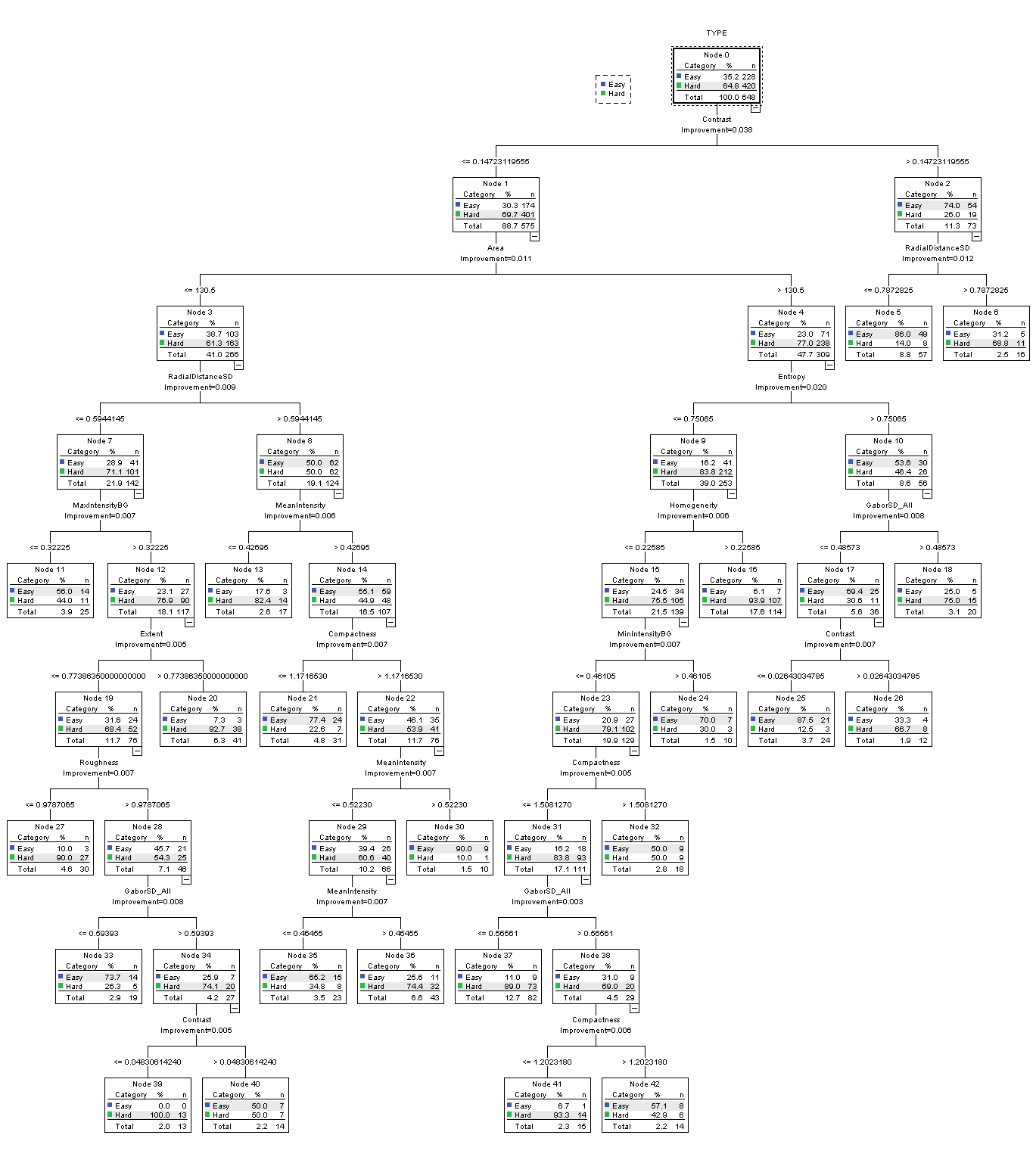


**Experiment run again with just important features identified from the tree above:**

The same experiment (same training and testing data, same tree config) was done but this time just using the important features identified from the first tree:

TREE TYPE [n] BY Contrast [s] RadialDistanceSD [s] Entropy [s] Area [s] MaxIntensityBG [s] MeanIntensity [s] Homogeneity [s] GaborSD\_All [s] Extent [s] Compactness [s] MinIntensityBG [s] Roughness [s] MaxIntensity [s]

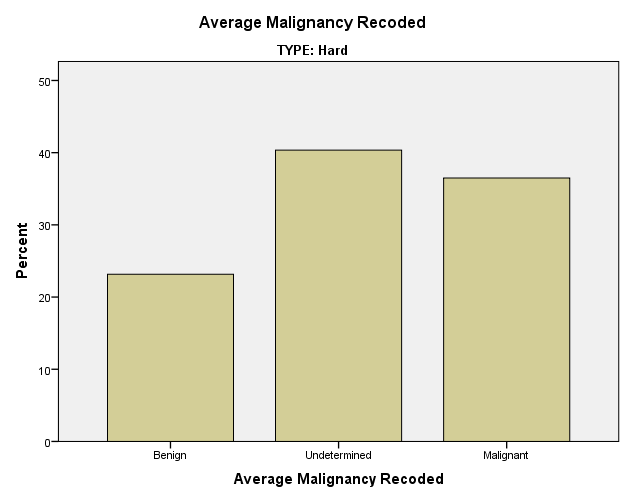
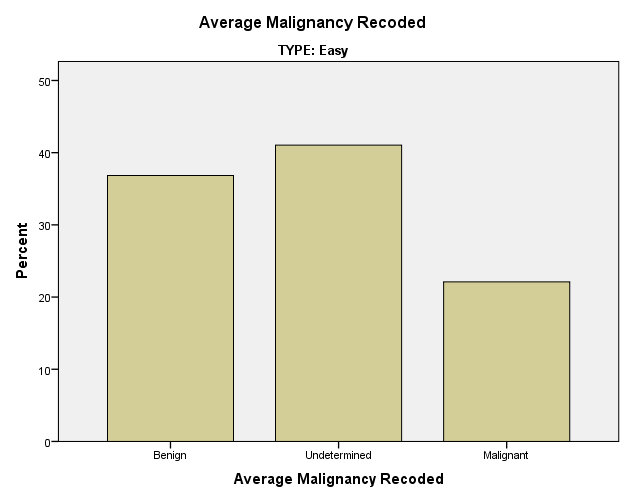
|  |  |  |  |
| --- | --- | --- | --- |
| **Classification** | | | |
| Observed | Predicted | | |
| Easy | Hard | Percent Correct |
| Easy | 177 | 51 | 77.6% |
| Hard | 68 | 352 | 83.8% |
| Overall Percentage | 37.8% | 62.2% | 81.6% |
| Growing Method: CRT  Dependent Variable: TYPE | | | |



**BALANCED DATA SET APPROACH**

The data was balanced in terms of the label variable ‘TYPE’. Easy cases were the minority label at 285, therefore 285 random hard cases were selected to balance the data set. Stratified sampling was done to split the data into training and testing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Average Malignancy Recoded** | | | | | | |
| TYPE | | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Easy | Valid | Benign | 105 | 36.8 | 36.8 | 36.8 |
| Undetermined | 117 | 41.1 | 41.1 | 77.9 |
| Malignant | 63 | 22.1 | 22.1 | 100.0 |
| Total | 285 | 100.0 | 100.0 |  |
| Hard | Valid | Benign | 66 | 23.2 | 23.2 | 23.2 |
| Undetermined | 115 | 40.4 | 40.4 | 63.5 |
| Malignant | 104 | 36.5 | 36.5 | 100.0 |
| Total | 285 | 100.0 | 100.0 |  |



Training (Selected) and Testing (Not Selected) split:

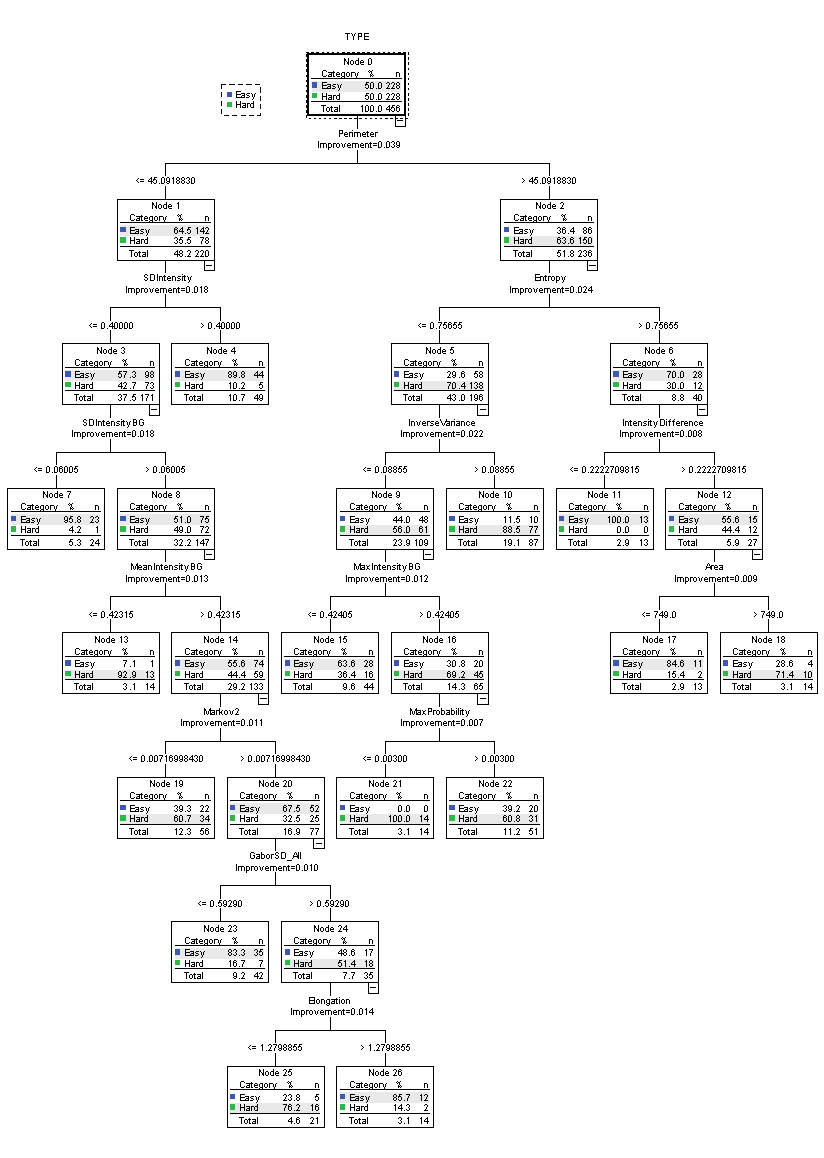
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TYPE** | | | | | | |
| Pran1 <= 80 (FILTER) | | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Not Selected | Valid | Easy | 57 | 50.0 | 50.0 | 50.0 |
| Hard | 57 | 50.0 | 50.0 | 100.0 |
| Total | 114 | 100.0 | 100.0 |  |
| Selected | Valid | Easy | 228 | 50.0 | 50.0 | 50.0 |
| Hard | 228 | 50.0 | 50.0 | 100.0 |
| Total | 456 | 100.0 | 100.0 |  |

**RESULTS:**

Cross validation results

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification** | | | |
| Observed | Predicted | | |
| Easy | Hard | Percent Correct |
| Easy | 166 | 62 | 72.8% |
| Hard | 33 | 195 | 85.5% |
| Overall Percentage | 43.6% | 56.4% | 79.2% |
| Growing Method: CRT  Dependent Variable: TYPE | | | |

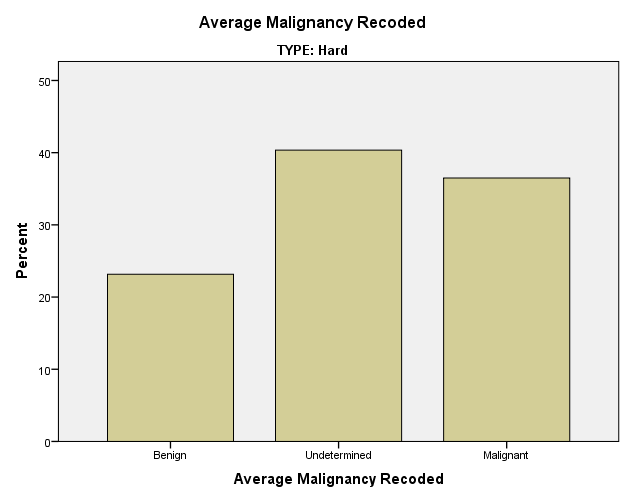
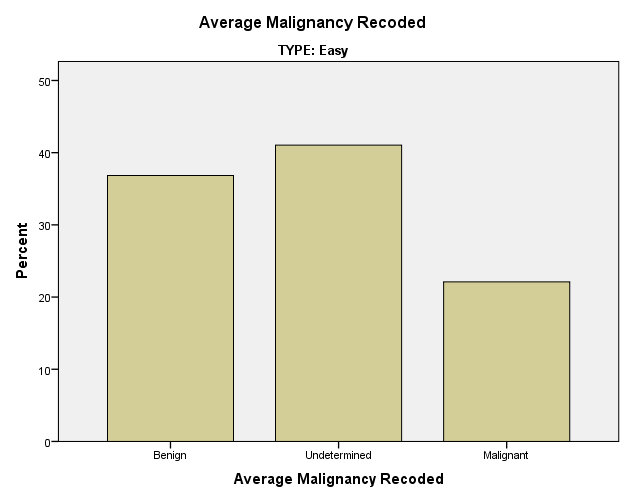
Tree shown below:



**BALANCED AND NORMALIZED DATA SET APPROACH**

All 810 cases were first normalized (min-max 0:1). The data was then balanced in terms of the label variable ‘TYPE’. Easy cases were the minority label at 285, therefore 285 random hard cases were selected to balance the data set (different 285 cases than the cases selected in the experiment above). Stratified sampling was done to split the data into training and testing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Average Malignancy Recoded** | | | | | | |
| TYPE | | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Easy | Valid | Benign | 105 | 36.8 | 36.8 | 36.8 |
| Undetermined | 117 | 41.1 | 41.1 | 77.9 |
| Malignant | 63 | 22.1 | 22.1 | 100.0 |
| Total | 285 | 100.0 | 100.0 |  |
| Hard | Valid | Benign | 66 | 23.2 | 23.2 | 23.2 |
| Undetermined | 115 | 40.4 | 40.4 | 63.5 |
| Malignant | 104 | 36.5 | 36.5 | 100.0 |
| Total | 285 | 100.0 | 100.0 |  |
|  |  |  |  |  |  |  |



**RESULTS:**

Cross validation results

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification** | | | |
| Observed | Predicted | | |
| Easy | Hard | Percent Correct |
| Easy | 181 | 47 | 79.4% |
| Hard | 42 | 186 | 81.6% |
| Overall Percentage | 48.9% | 51.1% | 80.5% |
| Growing Method: CRT  Dependent Variable: TYPE | | | |

Tree shown below:

