# **Abstract**

Breast Cancer is identified as the most commonly diagnosed form of cancer, worldwide. In 2018, it is estimated that 252,710 females and 2,470 males will be diagnosed with Breast Cancer. Of these newly diagnosed patients, an estimated 38,277 (15%) will lose this battle before 2019.

The purpose of this analysis is to determine if the less costly and less invasive Fine Needle Aspirate (FNA) procedure can be used as an adequate alternative to surgical biopsy as a Breast Cancer diagnostic technique. Additionally, we aim to build a classification model using FNA features to diagnose patient breast mass malignancy and thereby diagnose patient Breast Cancer. Lastly, we look to identify the most significant characteristics of the FNA procedure that aid in classification and patient diagnostic accuracy. We utilize the Wisconsin Breast Cancer Diagnostic dataset composed of 569 patients (63% benign and 37% malignant) assessing the nuclear cell features of the FNA procedure.

In this analysis a performance comparison is conducted between a variety of different data mining techniques and machine learning algorithms: linear classification, quadratic classification, K nearest neighbor (K-NN), logistic regression, decision tree, neural network, and gradient boosting. We assess the correctness in classifying a patients breast mass as malignant with respect to efficiency and effectiveness of each algorithm/technique in terms of accuracy, sensitivity, specificity and ROC index of the validation data. Ultimately, the HP Neural Network is selected as the best predictive model with a 99% classification rate, 95% sensitivity, 100% specificity and .995 ROC index for the validation data.

# **Introduction**

Cancer is a disease caused by the uncontrolled division of abnormal cells, often presenting itself in the form of malignant tumors; thus, Breast Cancer is the presence of one or many malignant tumors found in the breast of a patient. Breast Cancer, known as the most common form of cancer diagnosed in both men and women, worldwide, represents a global health problem responsible for countless patient deaths and immeasurable human suffering. Early detection, diagnosis, and treatment of Breast Cancer is vital to patient survival. When Breast Cancer is detected in its early stages, there is an estimated 30% chance that the patient can be treated effectively, thereby increasing the patients chances of survival. However, the longer Breast Cancer goes undetected and untreated the smaller the estimated chance that treatment is effective and thus the smaller the estimated chance of patient survival.

Currently, there are three primary diagnostic techniques commonly used when screening patients for malignant breast masses: mammography, fine needle aspirate, and surgical biopsy. The mammography procedure is the least invasive, least costly procedure of the three primary diagnostic procedures, however diagnostic sensitivity is known to fluctuate greatly ranging from 68% and 80%. Thus, mammography is the least diagnostically accurate of the three procedures in detecting Breast Cancer and is primarily used in the Healthcare Industry as an initial screening tool rather than a primary diagnostic tool. Surgical biopsy is a diagnostic procedure in which surgery is used to remove all or part of a breast mass believed to potentially be cancerous, the mass is analyzed after surgery to determine if it is in fact cancerous or not. Marketed with the slogan “A biopsy is the only way to know for sure if it’s cancer”, illustrates that surgical biopsy is the primary diagnostic tool currently used in the Healthcare Industry. Surgical biopsy while offering nearly 100% sensitivity, is accompanied by many pitfalls such as high cost, highly invasive, mentally and emotionally draining, and the procedure is often accompanied by an inpatient hospital stay. Fine Needle Aspiration (FNA) is a diagnostic procedure in which a small gauge needle is used to remove fluid directly from the patients potentially cancerous breast mass. This fluid is then stained with a colored dye and placed on a glass slide to reveal the nuclei of those cells collected. Software is then used to compute various characteristics of each nuclei, these characteristics are then analyzed to produce the final diagnosis of whether the breast mass is malignant or benign. The FNA diagnostic procedure is accompanied by sensitivity that fluctuates between 65% and 95% and when used alongside mammography is estimated to have a 5% rate of false negatives.

If researchers are able to make improvements on the sensitivity of the FNA procedure and the rate of false negatives the procedure has the potential to offer patients and physicians many benefits including: low cost, minimally invasive, reduction in time between procedure and results/diagnosis, reduction in amount of time patients diagnosed with Breast Cancer go before their first treatment, etc. The purpose of this analysis is to determine if the less costly and less invasive Fine Needle Aspirate (FNA) procedure can be used as an adequate alternative to surgical biopsy as a Breast Cancer diagnostic technique. Additionally, we look to identify the most significant characteristics of the FNA procedure that aid in classification and patient diagnostic accuracy.

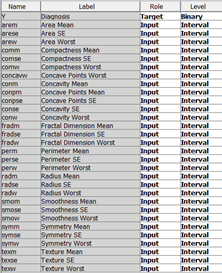
# **BACKGROUND & METHODOLOGY**

## **DATA DESCRIPTION & PREPARATION**

The dataset used in this paper is publicly available and was created by Dr. William H. Wolberg, a physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA.

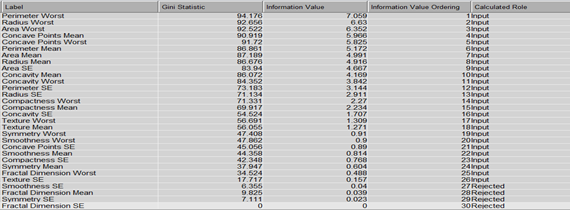
The dataset contains 569 patients along with 30 characteristics obtained through the FNA diagnostic procedure. The Fine Needle Aspirate diagnostic technique and the diagnostic process used to obtain the data analyzed in this analysis is summarized as follows. First, a small gauge needle is used to remove fluid directly from the patients potentially cancerous breast mass. This fluid is then stained with a colored dye and placed on a glass slide to reveal the nuclei of those cells collected. This slide is then scanned using a digital camera. After the majority of the nuclei have been stained and each nuclei is easily identifiable, a program calculates the mean , standard error, and worst (largest mean of the three largest nuclei) for the 10 characteristics considered relevant for accurate diagnosis: radius, texture, perimeter, area smoothness, compactness, concavity, concave points, symmetry, fractal dimension.

There are a total of 32 attributes per observation; including the ID and binary target variable. The target variable diagnoses whether the patients breast mass is malignant (37%) or benign (63%). Display 1, shows the variable names, roles and measurement levels for all variables contained within the dataset.



### **Display 1. List of variable names, roles, and measurement levels**

There are no data quality issues, such as missing values or duplicate instances, present in the dataset. The Interactive Grouping Node is used to implement the initial variable reduction technique, Weight of Evidence. The WOE procedure takes into consideration not only how the malignancy diagnostic rate (% of patients diagnosed with Breast Cancer) trends with each independent variable, but also how each independent variable is distributed across the entire spectrum of the patient portfolio. WOE is used in order to select features based on a features correlation with the dependent variable, that is it is used to determine the relative risk of a mass being malignant based on the values of a particular feature. A high negative WOE indicates high risk (Breast Cancer Diagnosis) of a feature while a high positive indicates low risk (benign mass). Then, the predictive power of each feature is assessed based on its Information Value. If the Information Value of a feature is greater than 0.1, the variable is chosen as an input variable for the next phase of the analysis. Display 2, shows the twenty six variables selected as outputs from the WOE procedure as well as the four variables rejected based on their Information Values.



**Display 2. Calculated Role of Independent Variables from Weight of Evidence Procedure.**

Next, the sample of 569 patients is split into two sub samples using a 70:30 ratio for training and validation respectively. The training data is used for model fitting while validation data is used for model validation. The sample size for the training and validation samples is depicted in Table 1.

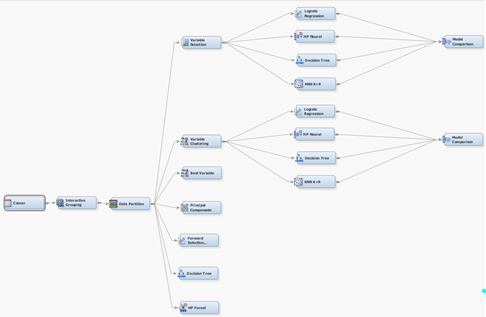
### **Table 1. The Partitioned Data for Training and Validation**

## **METHODOLOGY**

Several variable reduction techniques are implemented in order to select the most significant variables and reduce the 26 WOE variables into a smaller subset. Overall, seven variable reduction techniques are explored in Enterprise Miner, including: logistic regression (forward, backward, stepwise selection), variable selection, variable clustering, best variable clustering, principal components, random forest, and decision tree. Display 3, depicts the process flow diagram of Enterprise Miner for the variable reduction phase of this analysis.

### **Display 3. Enterprise Miner process flow of variable reduction analysis**

The following models are explored throughout this analysis: linear classification model with variation in variable subsets and priors, quadratic classification model with variation in variable subsets and priors, K nearest neighbor with variations in K and variable subsets, logistic regression with variation in variable selection criteria (default, stepwise, backward) and variable subsets as inputs, decision tree with variation in splitting rule target criteria (default, entropy, Gini, number of branches) and variable subsets used as inputs, random forest via variable selection with variation in the variable subsets used as inputs, and neural, auto neural and HP neural networks with variation in variable subsets, activation functions, hidden layers, and hidden nodes. The Enterprise Miner process flow of the modeling phase of this analysis is shown in Display 4 below. Display 4 highlights how each variable reduction technique explored is used as an input for each type of model built; this is done in an effort to achieve the highest sensitivity possible from a predictive model with the smallest number of predictors.



**Display 4. Enterprise Miner Process Flow of the Modeling Approach used**

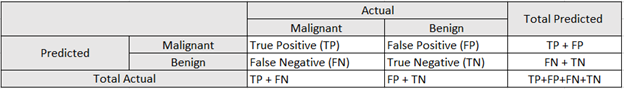
Similar to logistic regression, variable selection via a macro is used to further reduce the variable subsets used as inputs for the linear and quadratic classification models as well as the KNN models. These macros are shown in the appendix.

The type of Neural Network (NN) used for this analysis is a feedforward HP Neural Network, this type of NN is chosen due to the structures inherent stability of its behavior when learning, adjusting weights, and estimating final predictions. The feedforward structure works by allowing information (values) to flow forward through the model; information flows, from the input layer and input nodes, forward through the hidden layer and hidden nodes, eventually reaching the output layer and output node. One of the primary advantages to using the feed-forward network structure with a single hidden layer and an acceptable number of hidden neurons (3 in this analysis) is that the NN is able to mimic complex functions and relationships inherent in the data.

The input layer of the ANN in this analysis consists of 26 WOE independent variables, these variables are listed in Display 2. The input layer was used as a means of introducing the values of the independent variables into the model. The HP Neural Network executes in an iterative fashion in which variable values are placed in the input nodes, and then the NN gradually executes the hidden and output layer nodes. The hidden layer of the model calculates a weighted sum of the outputted independent variable values introduced in the input layer; the weighted sums are refined into a value for each of the hidden nodes in the hidden layer. Each of the hidden nodes output the weighted value from the hidden layer; these outputted values are then passed through the activation function. The activation function is generally defined as a function used by a node in order to transform input data from a domain of input values into a finite range of values; the activation used for the case of this analysis was the identity function which standardized values using the range. The output layer acts as the cumulative output of the entire network.

The Final Neural Network deemed to be the most predictively accurate model, of all Neural Networks build is constructed with the 26 WOE variables, one hidden layer, three hidden neurons, 82 weights, and one output node.

The performance of the models built are assessed based on the Receiver Operating Characteristic (ROC) Curve, misclassification rate, sensitivity, and specificity for the validation data. The misclassification rate is the percentage of patients who are incorrectly classified as having either a malignant or benign mass. The Receiver Operating Characteristic (ROC) Curve is used as a tool for model assessment, the larger the value for the ROC Validation Index, the higher the overall accuracy of the model is across all cutoff values. As seen in Table 2, True Positive is the number of patients who are predicted as having malignant breast masses who do in fact have malignant breast masses and who have been diagnosed with Breast Cancer. True Negative is the number of patients predicted as having benign breast masses, who have been diagnosed with benign breast masses. Sensitivity, is the probability that the model is able to correctly diagnose patients with malignant breast masses while Specificity is the models ability to correctly diagnose those patients with benign breast masses. We aim to find the best model of all those built throughout this analysis; the best model will have the lowest misclassification rate, highest sensitivity, highest specificity, and highest ROC Index for the validation data.



**Table 2. Classification Matrix of Predicted Values**

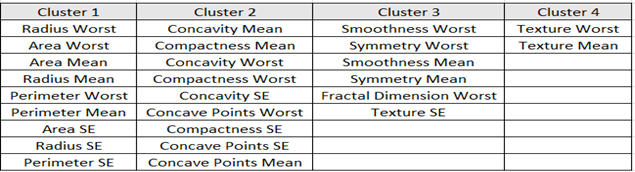
The formulas for the misclassification rate, sensitivity, and specificity are shown in the appendix. Furthermore, the settings used for each final model of its type are provided in the appendix.

**RESULTS**

Among the seven variable reduction techniques explored, the variable selection, variable clustering, and principal components methods are the three primary techniques utilized in the final models due to the high predictive accuracy achieved with these variables as inputs. Table 3 below highlights the five variables selected as the variable subset from the original twenty six WOE variables using the variable selection procedure.

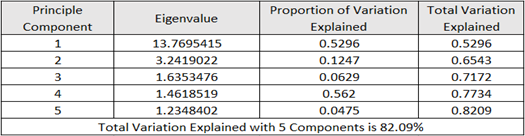
**  
Table 3. Variable Subset produced using Variable Selection**

The variable clustering procedure, reduces the 26 original WOE variables to the variable subset containing four variable clusters. These four variable clusters are able to account for an estimated 74% of the total variation present in the data. Table 4, highlights the four variables clusters and those WOE contained within each cluster.



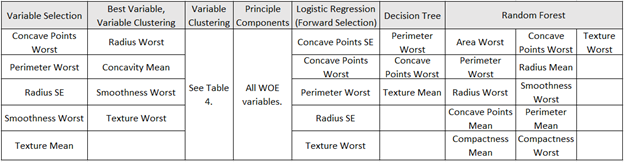
**Table 4. Variable Subset produced using Variable Clustering Procedure**

The principle components, variable reduction procedure takes all of the 26 original WOE variables as inputs and transforms all of them into uncorrelated linear combinations of the original variables. The principle components procedure produces five new variables or five principal components, these are selected using the eigenvalue greater than one criteria. These five components are able to account for nearly 82% of the variation and are used as independent variables throughout the modeling process. The eigenvalue of each principal component and the amount of variation explained by each component is seen in Table 5. Table 5 does not depict the variables contained within each principal component due to the fact that all 26 WOE variables used as inputs are contained within the components, no single variable or its variation is lost.



**Table 5. Variation Explained from Principal Components.**

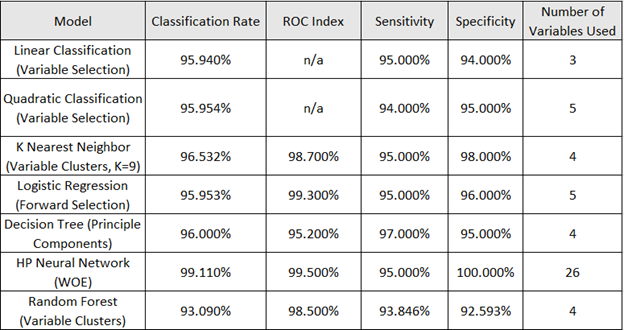
# One of the main objectives of this analysis is to determine which features of the cell nuclei obtained in the Fine Needle Aspirate diagnostic procedure are of critical importance for correctly diagnosing patient breast masses as malignant or benign. Table 6 below displays those variables that are selected in each variable reduction procedure explored in SAS Enterprise Miner.



**Table 6. Summary of WOE Variables selected in each Variable Reduction Method**

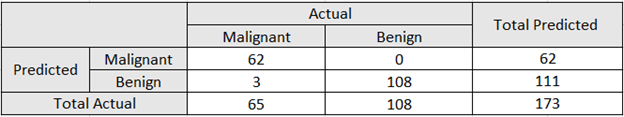
Across six of the seven variable selection methods explored, Concave Points Worst and Perimeter Worst are selected as significant features for analysis. Furthermore, across five of the seven variable selection methods explored, Smoothness Worst and Texture Worst ae selected as significant features to be used in analysis. Thus, to further improve outcome prediction and patient diagnosis, we recommend medical professionals consider refining those techniques used to measure and record the perimeter, concavity, texture, and smoothness of a cell’s nucleus within a potentially malignant breast mass.

Table 7, summarizes the Classification Rate, Sensitivity, Specificity, and ROC Index obtained for the validation data for the “best” model built for each type of model explored in this analysis along with the variable reduction technique that provides the highest predictive accuracy for that model type. Table 6, displays the model assessment statistics for the validation data. These model assessment statistics are used for model evaluation.



**Table 7. Model Assessment Fit Statistics**

After comparing all models the HP Neural Network with the 26 WOE variables as inputs is deemed to be the most diagnostically accurate model, possessing the following fit statistics : 99.11 classification rate, 99.5% ROC Index, 95% Sensitivity (True Positive), and 100% Specificity (True Negative) for the validation data. The HP Neural Network Model surpasses all other models across all cut offs of interest under the validation ROC index criteria. Furthermore, the HP Neural Network outperforms all other models in terms of misclassification with a validation misclassification error rate of .01 or 99% accuracy. Table 8, Displays the classification matrix of the validation data for the HP Neural Network.



**Table 8. HP Neural Network Classification Matrix**

# **CONCLUSION**

The best model for predictive purposes is the HP Neural Network. When compared to other models the Neural Network has the lowest misclassification rate (.01), the highest ROC index (.995), and the highest specificity of any model (100%). However, if health care professionals were to be more interested in interpretation of results rather than prediction, we would favor the Logistic Regression forward selection model. That is, with the Logistic Regression model we are able to accurately make statements such as the following: for every 1 unit increase in the radius standard error of cell nuclei the estimated odds that a patient's breast mass is malignant increase by 22%. The HP Neural Network provides physicians and patients increased levels of diagnostic accuracy with the use of the less invasive, less costly FNA procedure. Additionally, with the use of the HP Neural Network patients and physicians will see a reduction in time between accurate diagnosis of patients and medical intervention as well as a reduction in the number of patients who receive both false positives and false negatives. The Neural Network model provides patients with a diagnostically accurate alternative to the surgical biopsy procedure. In an effort to further improve outcome prediction, moving forward, we recommend medical professionals consider refining those FNA techniques used to measure and record the perimeter, concavity, texture, and smoothness of a cell’s nucleus within a potentially malignant breast mass.

\*if once this paper is trimmed down to five pages, we still have room we can include further recommendations for future analysis

# **APPENDIX**