Paper 9420-2016

Application of Data Mining Techniques in Improving Breast Cancer Diagnosis

Josephine S. Akosa, Oklahoma State University; Shannon Kelly, Oklahoma State University

ABSTRACT

Breast cancer is the second leading cause of cancer deaths among women in the United States. Although mortality rates have been decreasing over the past decade, it is important to continue to make advances in diagnostic procedures as early detection vastly improves chances for survival.

Our goal for this study is to identify a data mining model that accurately predicts the presence of a malignant tumor using data from fine needle aspiration (FNA) with visual interpretation. Compared with other methods of diagnosis, FNA displays the highest likelihood for improvement in sensitivity. Furthermore, we aim to identify the variables most closely associated with accurate outcome prediction. We utilize the Wisconsin Breast Cancer dataset which contains 699 clinical case samples (65.52% benign and 34.48% malignant) assessing the nuclear features of the FNA.

We analyze a variety of traditional and modern models, including: logistic regression, decision tree, neural network, support vector machine, gradient boosting, and random forest. Prior to model building, we used the weights of evidence (WOE) approach to account for the high dimensionality of the categorical variables and variable selection methods were employed. Ultimately, the gradient boosting model utilizing a principal component variable reduction method was selected as the best prediction model with a 2.4% misclassification rate, 96.27% specificity, 100% sensitivity, 0.963 Kolmogorov-Smirnov statistic, 0.985 Gini coefficient, and 0.992 ROC index for the validation data. Additionally, the uniformity of cell shape and size, bare nuclei, and bland chromatin were consistently identified as the most important FNA characteristics across variable selection methods. These results suggest that future research should attempt to refine the techniques used to determine these specific model inputs. Greater accuracy in characterizing the FNA attributes will allow researchers to develop more promising models for early detection.

INTRODUCTION

Over the past few decades, public awareness and scientific research concerning breast cancer has increased immensely. Unfortunately, breast cancer continues to be the second leading cause of cancer deaths among women in the United States, second only to lung cancer. Recent studies estimate approximately 15% of newly-diagnosed cancer patients will die in 2015. [1]

Early detection is critical to reducing the mortality rate. Currently, there are 3 methods that are commonly used to diagnose cancer, including: mammography, fine needle aspiration (FNA) with visual interpretation, and surgical biopsy. Surgical biopsy has the highest sensitivity, near 100%. However, due to the high costs associated with surgical biopsy, researchers are attempting to find ways to improve the sensitivity of both mammography and FNA. Studies show the sensitivity of mammography to fluctuate between 68% and 79%.[2] Unfortunately, limitations including variation in age, breast density, and availability of technology prevent researchers from significantly improving the sensitivity of mammography.[3-5] However, these limitations are less severe for FNA with visual interpretations and studies have found sensitivity to vary between 65% and 95%.[6] Additionally, according to the University of California San Francisco (UCSF) Medical Center, FNA biopsies are only minimally invasive and can be completed within minutes. Furthermore, the results of FNA biopsies are generally available very quickly. The rate of false negatives for FNA diagnosis is approximately 5% when combined with a clinical exam and a mammogram. [7]

Our goal is thus to utilize various data mining techniques to identify a diagnostic model that most accurately predicts the presence of a malignant tumor using data from FNA with visual interpretation.

Additionally, we seek to identify the most significant characteristics (input variables) of the fine needle aspirates that aid in accurate diagnosis.

BACKGROUND & METHODOLOGY

There is considerable research regarding breast cancer diagnostics. Literature review helped to identify some of the most promising data mining models for success. After identifying the models, further research into the methods employed by these models assisted in variable selection and developing model specifications. Although breast cancer specific research was not always available for some of the newer models, other health related studies helped to lay a framework for working with the data.

LITERATURE REVIEW/BACKGROUND

Cancer is a disease caused by the uncontrolled division of abnormal cells, often presenting in the form of malignant tumors. Therefore, breast cancer is a malignant tumor that starts in the breast. In recent years, breast cancer survival rates have increased and the number of deaths has declined due to improvements across a number of factors including early detection, treatment methods, and understanding of the disease.

Fine needle aspiration (FNA) is both an accurate and cost effective method of diagnosis. Researchers have been working diligently over the past decade to improve the sensitivity of this process. In addition to advances in technology and visual interpretation, researchers have been using various data mining methods to identify the key factors that can help doctors correctly diagnose malignant tumors.

Studies show the reporting sensitivity of mammography and ultrasound to vary with radiologists' experience. [8] However, computer decision aids can improve the radiologists' ability to correctly diagnose the malignancy of breast tumors. As a result, researchers have attempted to develop accurate decision aids to minimize the potential for interpretation errors by radiologists. Floyd et al. developed an artificial neural network (ANN) to predict the malignancy of breast cancer tumors from mammographic findings. This study found that the ANN model was significantly more accurate than the radiologists interpreting the results. [9]

Traditionally, researchers used logistic regression models and spent considerable time refining these models. Delen et al. compared three common data mining models as they related to breast cancer survivability, including logistic regression, decision trees, and artificial neural networks (ANN). This study found the decision trees to be the most accurate, with 93.6% accuracy. [10]

More recent studies include the use of newer data mining tools such as naïve bayes, support vector machine (SVM), radial basis neural network, and classification and regression trees (CART). Aruna et al. compared these newer models in addition to decision trees and found out that with respect to sensitivity, specificity, accuracy, and precision, the SVM with radial basis function kernel outperforms other classifiers. [11]

As stated previously, research dedicated solely to breast cancer diagnosis is limited for some data mining methods when compared to the total healthcare field. In order to identify other models that could be successful, we undertook a comprehensive literature review relating specifically to data mining and diagnosis. This comprehensive review provided further clarification of the most promising modern data mining methods.

In their study, Statnikov et al. compared support vector machines and random forest data mining methods with respect to gene expression microarrays. Microarrays are often used to aid in the diagnosis and prediction of clinical outcomes for cancer. This study found that support vector machines outperform random forest classifiers.[12] Nonetheless, the findings of Diaz-Uriarte and De Andres indicate that random forest classifiers have comparable performance to the support vector machines. [13]

Doyle et al. discussed the challenges of using computerized image analysis programs to examine tissue samples and identified a boosting model to develop an algorithm for accurate diagnosis.[14] Teramoto highlighted the capabilities of balanced gradient boosting when using imbalanced data as it relates to

outcome predication. This study found the balanced gradient boosting model to be superior to other supervised learning algorithms, including random forest and support vector machines.[15]

DATA DESCRIPTION & PREPARTION

This study utilizes the 1991 Wisconsin Breast Cancer Database, originally compiled by Dr. William H. Wolberg and available within the UCI Machine Learning Repository website.

The dataset contains 699 clinical case samples assessing the nuclear features of FNAs taken from patients' breasts. There are 11 attributes per observation; including the ID and binary target variable. The target variable diagnoses whether the tumor is benign (65.52% of cases) or malignant (34.48% of cases). The remaining input variables are measured on an ordinal scale (1-10), with a value of 1 indicating a normal state and a value of 10 indicating a highly abnormal state. Display 1 shows the variable names, roles, and measurement levels.

Name	Label	Role	Level
Diagnosis		**Target**	**Binary**
ID	Sample Code Number	**ID**	**Nominal**
NN	Normal Nucleoli	**Input**	**Ordinal**
MAdh	Marginal Adhesion	**Input**	**Ordinal**
CT	Clump Thickness	**Input**	**Ordinal**
UCSh	Uniformity of Cell Shape	**Input**	**Ordinal**
UCSz	Uniformity of Cell Size	**Input**	**Ordinal**
BN	Bare Nuclei	**Input**	**Ordinal**
BC	Bland Chromatin	**Input**	**Ordinal**
SECS	Single Epithelial Cell Size	**Input**	**Ordinal**
Mit	Mitoses	**Input**	**Ordinal**

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

There were 16 missing values in this dataset. Due to the small percentage of missing values (2.3%), we excluded these cases from the analysis. Additionally, there were 54 duplicated instances. However, there were not enough information regarding these instances (whether errors, accidental duplications, or repeated measurements) to exclude them from the analysis.

Each categorical input variable has 10 levels. Although the typical modeling approach would include the use of dummy variables, this would lead to a tremendous increase in dimensionality and the possibility of overfitting in the training data. In order to address these issues, we used the weights of evidence (WOE) approach to convert the categorical variables into numerical values. The WOE approach is a quantitative method used to measure the association between an input variable and target variable. This method was originally developed for medical diagnosis, in which the magnitude of the weights depended upon the measured association of a symptom and the presence of a disease. These weights were then used to help estimate the probability of a positive diagnosis based on presence or absence of symptoms.

Variable	Label	Gini Statistic	Information Value	Information Value Ordering
UCSz	Uniformity of Cell Size	95.155	6.786	1
UCSh	Uniformity of Cell Shape	94.597	6.529	2
BC	Bland Chromatin	87.958	4.77	3
BN	Bare Nuclei	86.662	4.755	4
CT	Clump Thickness	80.325	4.19	5
SECS	Single Epithelial Cell Size	85.119	4.09	6
MAdh	Marginal Adhesion	80.286	3.954	7
NN	Normal Nucleoli	78.203	3.823	8
Mit	Mitoses	42.318	1.507	9

Display 2. Output variables for Weight of Evidence

For WOE approach, consider a binary target Y, with levels; 0 and 1, where $Y = 1$ is the event of interest. Now, consider an input variable X with "m" categories. Then the weight of evidence is calculated as

$$\boldsymbol{WOE_i = \log \frac{P(X = x_i | Y = 1)}{P(X = x_i | Y = 0)} \qquad for\ i = 1, 2, \cdots, m}$$

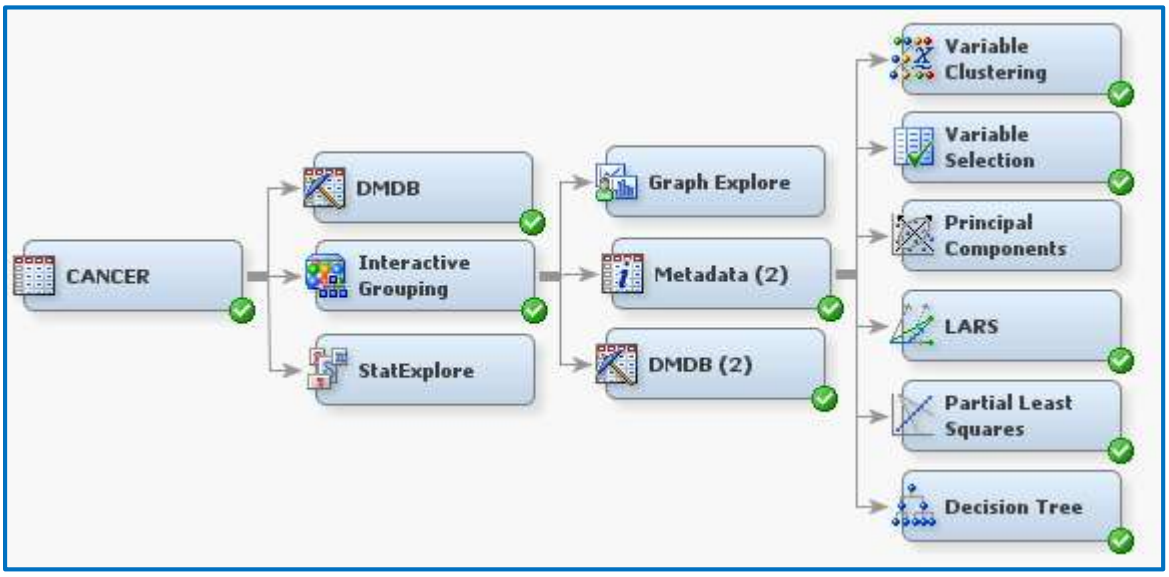
We implemented the WOE approach via the interactive grouping node in SAS® Enterprise Miner™. Display 2 provides the output variable results from the node. Further analysis of the summary statistics of the newly created WOE variables as shown in Display 3 does not give any indication that variable transformation is necessary.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
WOE_BC	INPUT	0.477707	2.890269	683	0	-5.39848	0.625434	3.684704	-0.63211	-0.82868
WOE_BN	INPUT	0.244831	2.730468	683	0	-4.38056	2.213852	2.213852	-0.82421	-1.09895
WOE_CT	INPUT	0.071218	2.774509	683	0	-5.73735	1.371444	3.194682	-0.9456	-0.13059
WOE_MAdh	INPUT	0.036366	2.542763	683	0	-5.18371	1.873844	1.873844	-1.10564	-0.22361
WOE_Mit	INPUT	-0.05134	1.366308	683	0	-3.39195	0.563945	0.563945	-1.86605	1.664913
WOE_NN	INPUT	0.054403	2.396854	683	0	-5.41515	1.635774	1.635774	-1.15138	-0.15502
WOE_SECS	INPUT	0.432055	2.424658	683	0	-4.28292	2.276049	2.276049	-0.69649	-1.25207
WOE_UCSh	INPUT	1.253832	3.621648	683	0	-3.84393	4.528133	4.528133	-0.41124	-1.58216
WOE_UCSz	INPUT	0.847369	3.674331	683	0	-4.66824	3.905141	3.905141	-0.58293	-1.41396

Display 3. Weight of Evidence variable summary statistics

METHODOLOGY

We followed the Cross Industry Standard Process for Data Mining (CRISP-DM) modeling approach. The five phases in this process include: understanding the business problem, understanding the data, data preparation, modeling, evaluation and deployment. To ensure honest assessment of the models built, we partitioned the data into training (70%) and validation (30%) subsets. Since the dataset was imbalanced, prior probabilities were set to account for oversampling.



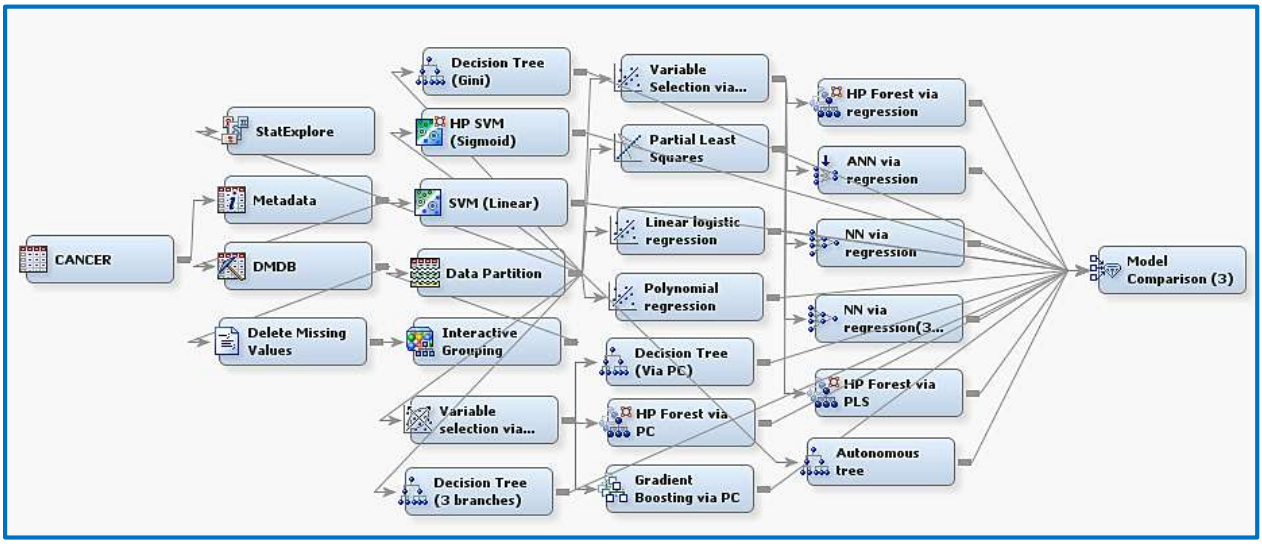
Display 4. Variable selection Process Flow

SAS® Enterprise Miner™ 13.1 provides many new models, including support vector machine (SVM), random forest, and gradient boosting. These newer models were included in the analysis along with the more traditional logistic regression, decision tree, and neural network. As previously stated, our main goal for this study is to build a diagnostic model that most accurately predicts the presence of a malignant tumor. In addition, we would expect this model to help identify the FNA characteristics with highest

importance to accurate outcome prediction. Consequently, we implemented several variable selection nodes to select the most significant input variables, including: variable selection, variable clustering, decision tree, partial least squares, principal component analysis, regression, and LARS (Display 4). Among these variable reduction techniques, the principal components were the most significant variables in reducing model assessment statistics.

Principal component analysis (PCA) is a technique used to convert a set of potentially correlated observations into sets of linearly uncorrelated variables. Typically, PCA is used to reduce the dimensionality of large multivariate data sets. The first component extracted accounts for the majority of the total variance within the variables. As a result, one can expect that this component will be correlated with at least some of the observed variables. The second component accounts for a majority of the variability not previously explained by the first component. Components continue to be generated in this fashion until all of the variability in the data is explained. Subsequently, each new component accounts for a progressively smaller amount of variance, explaining why only the first few components are typically retained for model building. Interpretation of the principal components is based upon determining which observed variables are correlated to the components. In this analysis, a correlation value of 0.5 in absolute value is deemed significant.

We considered the following models in our analysis: logistic regression with variation in variable selection criteria (default, stepwise, backward, decision tree, principal components), decision tree with variation in splitting rule target criteria (default, entropy, Gini, number of branches), neural and auto neural networks via variable selection, gradient boosting via variable selection, random forest via variable selection, support vector machine with variation in kernel function (linear, sigmoid, polynomial) via variable selection. Display 5 shows the process flow diagram for this study.



Display 5. Process flow diagram

Originally, we chose the validation misclassification rate for model selection criteria. However, it was determined that the best model could not be selected by this statistic alone as several models had equivalent values. Ultimately, we included in the model selection criterion; misclassification rate, Gini coefficient, Kolmogorov-Smirnov (KS) statistic, ROC index, sensitivity and specificity for the validation data.

In selecting the best model, we gave highest importance to the misclassification rate, or number of incorrect diagnoses, followed closely by the specificity (true negative rate) and sensitivity, also known as the true positive rate. The ROC index graphically displays the tradeoff for a higher true positive rate, plotting the true positive rate by the false positive rate. We used the KS statistic to measure goodness of fit. Lastly, although traditionally used to quantify income inequality, we included the Gini coefficient as a measure of heterogeneity, or frequency dispersion, of the data.

RESULTS & DISCUSSION

Our primary objective is to develop an accurate data mining model to diagnose the malignancy of breast cancer tumors. After comparing all of the models, we selected the gradient boosting via principal components (Boosting via PC) as the best model, exhibiting the following validation statistics: 0.024 misclassification rate, 0.985 Gini coefficient, 0.963 KS statistic, 0.992 ROC index, 100% sensitivity and 96.27% specificity. When compared to the other models, the selected model has the highest sensitivity and KS statistic, lowest misclassification rate, and the second highest Gini coefficient, ROC index and specificity as shown in Table 1.

Model Description	Misclassification rate	KS Statistic	Gini Coefficient	ROC Index	Sensitivity	Specificity
Boosting via PC	0.024	0.963	0.985	0.992	100.00%	96.27%
Decision tree via PC	0.029	0.949	0.949	0.974	98.63%	96.27%
HP Forest via PC	0.029	0.949	0.969	0.984	98.63%	96.27%
Autoneural via regression	0.029	0.943	0.979	0.990	97.26%	97.01%
Linear Logistic regression	0.034	0.940	0.982	0.991	97.26%	96.27%
HP Forest via regression	0.034	0.940	0.982	0.991	97.26%	96.27%
HP Forest via PLS	0.034	0.935	0.977	0.989	97.26%	96.27%
Autoneural (default)	0.034	0.963	0.988	0.994	95.89%	97.01%
Decision tree (3 branches)	0.043	0.920	0.937	0.968	97.26%	94.78%
SVM (Linear)	0.043	0.942	0.984	0.992	95.89%	95.52%

Table 1. Model assessment fit statistics

In general, boosting models are supervised learning ensemble models that combine the best aspects of weaker models to develop one strong model. Specifically, gradient boosting models combine predictions from a set of decision trees into a single prediction model. The ultimate goal of this technique is to increase the probability of selecting an observation that aids in predicting the target variable accurately. This technique builds a series of incrementally improved decision trees through resampling of the data set with replacement to produce results that form a weighted average of the resampled data. Typically, boosting algorithms' weighting is related to accuracy, placing greater weights on misclassified cases as the model develops.

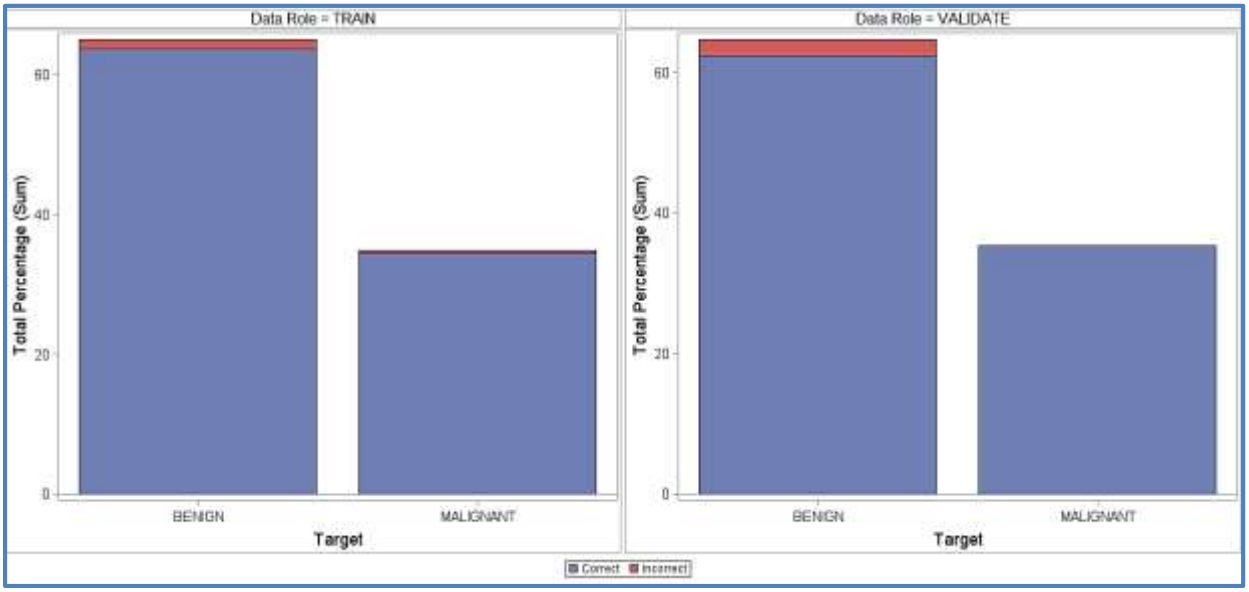
	Eigenvalue	Difference	Proportion	Cumulative
1	6.25799957	5.59225502	0.6953	0.6953
2	0.66574455	0.14459083	0.0740	0.7693
3	0.52115371	0.12056303	0.0579	0.8272
4	0.40059068	0.10255751	0.0445	0.8717
5	0.29803317	0.01056888	0.0331	0.9048
6	0.28746429	0.02630888	0.0319	0.9368
7	0.26115541	0.04287056	0.0290	0.9658
8	0.21828485	0.12871107	0.0243	0.9900
9	0.08957378		0.0100	1.0000

Display 6. Eigenvalues of correlation matrix

With respect to the selected gradient boosting model, the first five principal components were used in the model building since these components account for 90.48% of the total variability in the data as shown in the correlation matrix, Display 6. The first principal component was identified as the most important variable for accurate diagnosis; indicated by the number of splitting rules (NRULES) and the variable importance (VIMPORTANCE) in Display 7. The classification charts for both the training and validation data sets for the gradient boosting are also displayed in Display 8. These plots indicate that the model minimizes the misclassification rates and does not indicate overfitting in the training data.

NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
PC_1	Principal Component 1	20	1.00000	1.00000	1.00000
PC_3	Principal Component 3	5	0.05148	0.06616	1.28507
PC_2	Principal Component 2	5	0.03384	0.00464	0.13709
PC_5	Principal Component 5	3	0.03320	0.00794	0.23921
PC_4	Principal Component 4	2	0.01792	0.00000	0.00000

Display 7. Variable Importance using the Boosting model via principal component analysis



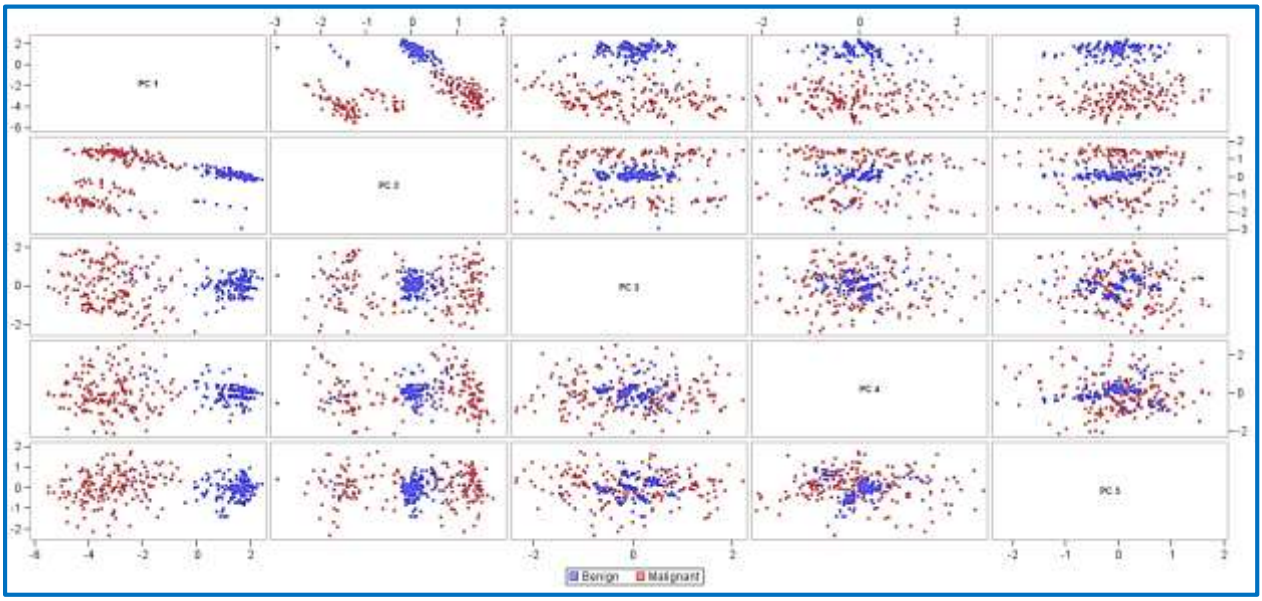
Display 8. Classification chart for boosting model via principal component analysis

Table 2 displays the correlation of the WOE of the observed variables and the principal components. As expected, the first principal component is strongly correlated with all the original variables. In this component, the value increased with increasing values in the original variables. Furthermore, this component correlates most strongly with the uniformity of cell shape and size and could be interpreted as the primary measure of these two attributes. The second and third components are respectively correlated with mitosis and clump thickness. The remaining components are not significantly correlated with any of the original observed variables.

Variable	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8
WOE_UCSz	0.939	-0.091	0.000	-0.052	-0.007	-0.098	-0.115	-0.170
WOE_UCSh	0.911	-0.150	0.042	-0.045	-0.087	-0.147	-0.105	-0.267
WOE_SECS	0.885	-0.036	-0.068	-0.106	0.022	-0.192	-0.244	0.320
WOE_BC	0.857	-0.144	-0.081	-0.050	0.021	0.463	-0.145	0.009
WOE_NN	0.834	0.074	-0.143	-0.407	-0.128	0.004	0.309	0.041
WOE_BN	0.829	-0.153	0.018	0.399	-0.317	0.003	0.139	0.097
WOE_MAdh	0.828	-0.061	-0.307	0.212	0.376	-0.055	0.163	-0.016
WOE_CT	0.750	0.017	0.626	-0.010	0.173	0.024	0.105	0.051
WOE_Mit	0.632	0.761	-0.031	0.108	-0.038	0.033	-0.065	-0.043

Table 2. Correlation between observed variables and principal components

Display 9 displays the scatterplots of the first five principal component scores, and shows that the components are uncorrelated. In addition, the pairwise scatterplots of the components and the remaining component scores indicate that a hyperplane could be constructed that significantly separates malignant and benign tumors.



Display 9. Principal Components Matrix

Our secondary objective for the study is to identify the characteristics of fine needle aspirates with the highest importance to outcome prediction. As discussed previously, we used the WOE approach to help identify the variables most closely associated with accurate outcome prediction. Display 2 showed the Gini statistic and information value for all of the input variables using the weights of evidence (WOE) approach. Variable importance is judged by the Gini statistic and information value. This approach identifies the uniformity of cell shape and size, bare nuclei, and bland chromatin as the most important input variables. These results are reinforced by other variable selection techniques, as shown in Table 3.

Variable selection node	Variable Clustering node	Decision tree	Partial least squares	LARS	Regression
BN	UCSz	UCSh	UCSz	UCSz	BN
BC	UCSh	UCSz	UCSh	UCSh	CT
CT	SECS	BC	BN	BN	UCSz
Madh	BN	BN	SECS	SECS	
NN	BC		BC	BC	
SECS	NN		NC	NC	
UCSh	Madh		Madh	Madh	
UCSz	CT		CT	CT	
				Mitosis	

Table 3. Summary of variables selected by specific nodes using weight of evidence

Across all variable selection methods, uniformity of cell size and shape, bland chromatin and bare nuclei were selected. The regression selection approach however deviated slightly from the variables selected by the other methods; selecting clump thickness instead of uniformity of cell shape and bland chromatin. The findings of this study would encourage technological advances aimed at ensuring the reliability and refinement of these measurements via FNA procedures.

CONCLUSION

This study aimed to identify a diagnostic model using data mining techniques that most accurately predicts the presence of a malignant tumor using data from fine needle aspiration (FNA) with visual interpretation. Furthermore, this study sought to identify the characteristics most closely associated with accurate outcome predictions. A variety of data mining models were considered, but the gradient boosting model using principal components variable selection was ultimately selected as the best model, based on a number of validation data statistics. In addition, four input variables were identified as significant to outcome prediction, including: uniformity of cell shape and size, bare nuclei, and bland chromatin. These results indicate that outcome prediction can be further improved by refining the methods used to identify and measure these characteristics. For example, technological advances that improve the reliability of uniformity estimates could improve the results of the data mining models. Finally, utilizing this model would help decrease interpretation errors by radiologists. In order to validate these findings, it is important for further research to be conducted; including applying this method to other types of malignant tumor diagnosis.

REFERENCES

[1] American Cancer Society. (2015). Cancer Facts & Figures 2015. Atlanta: American Cancer Society

[2] Fletcher, S. W., Black, W., Harris, R., Rimer, B. K., & Shapiro, S. (1993). Report of the international workshop on screening for breast cancer. Journal of the National Cancer Institute, 85(20), 1644-1656.

[3] Machida, Y., Tozaki, M., Shimauchi, A., & Yoshida, T. (2015). Breast density: the trend in breast cancer screening. Breast Cancer, 22(3), 253-261.

[4] Kolb, T. M., Lichy, J., & Newhouse, J. H. (2002). Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations. Radiology, 225(1), f65-175.

[5] Saarenmaa, I., Salminen, T., Geiger, U., Heikkinen, P., Hyvärinen, S., Isola, J., ... & Hakama, M. (2001). The effect of age and density of the breast on the sensitivity of breast cancer diagnostic by mammography and ultasonography. Breast cancer research and treatment, 67(2), 117-123.

[6] Giard, R. W., & Hermans, J. O. (1992). The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer, 69(8), 2104-2110.

[7] Learn more about fine needle aspiration (fine needle biopsy) (http://ww5.komen.org/BreastCancer/FineNeedleBiopsy.html)

[8] Britton, P., Warwick, J., Wallis, M. G., O'Keeffe, S., Taylor, K., Sinnatamby, R., ... & Wishart, G. C. (2014). Measuring the accuracy of diagnostic imaging in symptomatic breast patients: team and individual performance. The British journal of radiology.

[9] Floyd, C. E., Lo, J. Y., Yun, A. J., Sullivan, D. C., & Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. Cancer, 74(11), 2944-2948.

[10] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

[11] Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2, 37-45.

[12] Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics, 9(1), 319.

[13] Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 3.

[14] Doyle, S., Feldman, M., Tomaszewski, J., & Madabhushi, A. (2012). A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. Biomedical Engineering, IEEE Transactions on, 59(5), 1205-1218.

[15] Teramoto, R. (2009). Balanced gradient boosting from imbalanced data for clinical outcome prediction. Statistical applications in genetics and molecular biology, 8(1), 1-19.

[16] UCI machine learning repository: Breast cancer wisconsin (original) data set (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))

[17] Biopsy for Breast Cancer Diagnosis: Fine Needle Aspiration Biopsy (http://www.ucsfhealth.org/education/biopsy_for_breast_cancer_diagnosis/fine_needle_aspiration_biopsy/)

[18] Principal Component Analysis (http://support.sas.com/publishing/pubcat/chaps/55129.pdf)

ACKNOWLEDGMENTS

We wish to express our sincere gratitude to Dr. Goutam Chakraborty, Department of Marketing and founder of SAS and OSU Data Mining Certificate program – Oklahoma State University for his support and guidance throughout this study.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Josephine Sarpong Akosa
Oklahoma State University
Department of Statistics
320C Math Sciences (MSCS)
Stillwater, OK 74078
Email: josephine.akosa@okstate.edu
Web: www.statistician.wix.com/josephine-akosa