**Module 6 Critical Thinking**

**Second Draft of Final Research Paper**

Tim McCombs

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Chris den Heijer

June 26, 2022

**Contents**

[**Second Draft of Final Research Paper** 3](#_Toc105942068)

[**Introduction** 3](#_Toc105942069)

[**Objectives** 3](#_Toc105942070)

[**Overview of Study** 3](#_Toc105942071)

[**Research Questions and Hypotheses** 4](#_Toc105942072)

[**Literature Review** 5](#_Toc105942073)

[**Research Design** 8](#_Toc105942074)

[**Methodology** 8](#_Toc105942075)

[**Methods** 9](#_Toc105942076)

[**Limitations** 10](#_Toc105942077)

[**Ethical Considerations** 10](#_Toc105942078)

[**Findings** 11](#_Toc105942079)

[**Conclusion** 16](#_Toc105942080)

[**References** 17](#_Toc105942081)

# **Second Draft of Final Research Paper**

## **Introduction**

Breast cancer is the second leading cause of cancer deaths in the United States among women, second only to lung cancer (Breastcancer.org, 2022). Approximately one in eight women will get some form of invasive breast cancer during their lifespan. It is estimated that over 43,000 women will die this year from the disease. These statistics establish the urgency to combat breast cancer and the best approach to accomplish this is early diagnosis. As with most cancers, the chances for remission and survival increase the earlier the disease state is identified and treated. Once breast tumors are identified, a biopsy is usually obtained and the sampled tissue is categorized as malignant or benign through a medical examination. This research study aims to explore how data analytics could be used to aid in breast cancer diagnosis, improving tissue analysis speed and efficiency as well as decreasing patient impact and cost.

## **Objectives**

A descriptive statistical analysis of a dataset containing the breast tissue imaging measurements from fine-needle aspiration biopsy will be performed using traditional predictive models to substantiate previous literature showing these models may be effective in classifying breast tumors for malignancy. Also, a review of current literature will explore published literature supporting these techniques in two sections; prediction using traditional machine learning models (supervised, unsupervised, and semi-supervised), and research in prediction using recent trends in big data and artificial intelligence.

## **Overview of Study**

The best weapon in the fight against breast cancer death is early detection. Current clinical practice guidelines for breast cancer diagnosis start with breast lump awareness. Either the patient through self-breast examination or diagnostic scans such as mammography is essential to first locate tissue masses that may signal malignant growth. Once breast lumps are identified, tissue samples are taken to perform further diagnostic tests and a physical examination. Physical examinations must be performed by a trained pathologist to make a final determination of malignancy. As these clinical pathways have developed over time, the number of fatalities due to this disease has been decreasing year-to-year (Breastcancer.org, 2022). However, several opportunities to improve the efficiency and speed of clinical diagnosis still exist. Obtaining tissue samples for every suspect lump is invasive, expensive, and time-consuming for the patient. Also, the use of clinical examination by specially trained pathologists for the least invasive biopsy (fine-needle aspiration) adds significant cost and time to this process as well as significantly limiting access to this potentially preferred biopsy method (Bevers et al., 2018). Data mining techniques may be leveraged to improve the speed and access to diagnosis as well as possibly decrease costs.

This study explores the possible use of machine learning models to replace trained pathologist examination or at least minimize the use of these clinical pathways to only cases considered the most complicated to diagnose. The data used to determine a diagnosis is a collection of various cell nuclei measurements taken from MRI images of the fine-needle aspiration biopsy tissue. Historical data can be used to train a model to predict malignancy for future biopsy examinations.

## **Research Questions and Hypotheses**

Can tumor malignancy be predicted in women with identified breast lumps using cell nuclei measurements from fine-needle breast biopsy tissue images? Table 1 displays the null and alternative hypotheses for the research question. Two data analytics model methods will be used for the statistical tests. From these tests, does either of the prediction models explored outperform the other? Table 2 displays the null and alternate hypotheses for this second research question.

**Table 1**

*Null and Alternate Hypothesis for the Research Question*

|  |  |
| --- | --- |
| H0 | Breast tumor malignancy cannot be predicted by using cell nuclei measurements from fine-needle breast biopsy tissue images |
| Ha | Breast tumor malignancy can be predicted by using cell nuclei measurements from fine-needle breast biopsy tissue images |

*Note*. H0 is the null hypothesis and Ha is the alternative hypothesis.

**Table 2**

*Null and Alternate Hypothesis for Second Research Question*

|  |  |
| --- | --- |
| H0 | The performances between the two prediction models are not observably significant. |
| Ha | The performances between the two prediction models are observably significant. |

*Note*. H0 is the null hypothesis and Ha is the alternative hypothesis.

## **Literature Review**

Using data analysis to predict malignancy from tissue examination data is already being studied. Venkatesh et al. (2020) concluded logistic regression models are an excellent model to use for prediction for this use case, outperforming several other model algorithms. The model in this study achieved a 98% accuracy score which outperformed both other supervised and unsupervised models including classification trees, support vector machines, and even neural networks. This evidence is further substantiated in work done by Mohammad (2020) who focused on optimizing feature selection with logistic regression models to boost accuracy. His results yielded a logistic regression accuracy score of 94% from his optimized model (using the area under the curve (AUC) from the receiver operating characteristic or ROC curve plot). Both of these studies suggest data analysis can produce statistically significant prediction results to detect breast cancer. Furthermore, binary logistic regression may be a good model option for the project’s research question to accurately predict malignancy using measurements from fine-needle aspiration biopsy tissue imaging.

Along with supervised and unsupervised models, semi-supervised data mining models are another possible option to use for prediction. Al-Azzam & Shatnawi (2021) used the same breast cancer dataset to compare supervised and semi-supervised model performances. These include logistic regression, Naïve Bayes, support vector machine, decision tree, random forest, Xgboost, gradient boost, and K-nearest neighbor (KNN). While the logistic regression and KNN slightly outperformed other models in accuracy, the more important conclusion was that the performance of all three model groups was not statistically significant. Therefore, there is no evidence to support choosing unsupervised or semi-supervised models over supervised methods.

There is, however, conflicting literature as to which data mining model, supervised or other, has the best accuracy and confusion matrix performance results. Naji et al. (2021) contradict the previously cited studies in their comparison of model performances from a list of algorithms. The same dataset is used once again, probably due to the scarcity of public patient data available in the industry. The authors published study results that suggest support vector machine algorithms outperform the other models. However, all models tested achieved accuracy percentages above 90% and all except the standard KNN model achieved higher than 95%. This study provides no evidence suggesting avoiding using standard models like binary logistic regression or tree classification.

Next is a literature review of recent trends happening in data science, big data, and artificial intelligence. Is there any literature supporting further exploration into these approaches for breast cancer detection? Kirola et al. (2022) perform a comprehensive review of big data and machine learning/deep learning techniques to predict breast cancer by image classification. Several separate studies were compiled and over 20 data mining models were compared. The results of this study were inconclusive. While the study acknowledges big data is beginning a new renaissance of data analysis in the healthcare industry, the mere lack of public data prevents any real evaluation of potential in image classification. The dataset that will be used for this capstone project and nearly all of the studies listed in this comprehensive review is not an instance of big data. There are big data examples with breast cancer detection via genetic sequencing but this study did not include these in their discussion. Detection from genetic data has not been discussed so far in this review as it is not within the scope of the selected dataset but will be explored next for artificial intelligence.

Nassif et al. (2021) perform a comprehensive review of studies utilizing artificial intelligence (AI) methods for breast cancer detection. Do AI methods show greater accuracy and therefore provide a better chance of becoming a standard of practice for diagnosis? The studies compiled were placed into two categories, detection using imaging data and detection using genetic sequencing. The group explored model performance for each detection method. For both methods, complex hybrid AI algorithm ensembles provided the best accuracy scores averaging above 99.5%. This indeed is evidence that AI methods may have a greater potential for optimum breast cancer detection. But just how feasible is AI in real practice and is one method better than the other? The authors of the study performed a comparison of breast cancer detection methods listing the advantages, disadvantages, and challenges in image data classification and gene sequencing methods. Image data classification contains more features relevant to the analysis, however, these models performed slightly less than genetic data models in accuracy on average. The mean accuracy of image classification was 99.3% versus gene sequencing at 99.8%. Imaging data also requires more processing for data cleaning and post-feature selection data transformations. This would suggest an advantage towards gene sequencing analysis methods. However, gene sequencing is complex and expensive which would affect accessibility in non-metropolitan areas. Also, the research data is not nearly as extensive with fewer studies published, all recently. This means medical confidence has yet to be established.

Published research that examines the use of data analytics to help diagnose breast cancer is extensive showing significant results. There is some debate, however, as to which traditional model algorithm is best. The best overall performers appear to be AI model ensembles using genetic sequencing data or breast biopsy imaging data for classification. AI using image classification provides the greatest potential of being universally accepted within the medical community due to more established research. Using complex hybrid ensembles is beyond the scope of the capstone project due to time, resource, and skillset constraints.

## **Research Design**

### **Methodology**

A deductive, quantitative methodology will be chosen to perform a statistical analysis of the dataset. A quantitative analysis approach focuses on numerical data and data coded as numbers. This methodology assumes scientific standards for research with the foundation being the scientific method. The scientific method process involves making observations, generating a hypothesis based on the observations, collecting numerical data (quantitative data) through experiments or surveys, analyzing the data with statistics, and then making a conclusion that either supports or rejects your hypothesis. All independent variables within the chosen dataset are measurable numerical data from tissue samples taken via experiments and the numerical representation of the patient.

### **Methods**

The dataset that will be explored in the portfolio research project originally came from patient breast biopsy image data from the University of Wisconsin (Wolberg et al., 1995) and then redistributed by the University of California School of Information and Computer Science (Dua & Graff, 2019). This dataset contains 32 variables or attributes and 569 observations. Each observation or row lists one patient’s cell biopsy examination measurements and the diagnosis of malignant (cancer identified) or benign (non-cancerous). The examination measurements were obtained using a digitalized image of cell nuclei from a fine needle aspirate biopsy. Each record or row is a patient identified with a unique ID number variable. The *diagnosis* variable is the dependent variable in the dataset. It could be looked at as a class variable but it is truly a binary class variable. The two value options are M for malignant (can also be expressed and cancer diagnosis equals true) and B for benign (cancer diagnosis equals false). All other variables are calculation results from the measurements taken, all rounded to 4 significant digits. These attributes can be grouped into three sections; the mean of the individual measurements, the standard error for the individual measurements, and the mean of the largest three values for the individual measurements (identified as *worst*). The individual measurements for each image examination are as follows; the radius of the cell nuclei, the texture or spectrum of gray-scale values when observing the sample, the perimeter of the outer surface, the area, the smoothness or how uniformly round the surface is, a compactness calculation, concavity or the severity of concave portions observed, the number of concave portions named concave points, symmetry, and a fractal dimension calculation.

The statistical analysis of the dataset using SAS Studio, SAS OnDemand will include any remarkable descriptive statistics found (SAS Institute, n.d.). Traditional predictive models will be created and used to substantiate the previously published literature performance results. The goal of the model is to effectively predict the diagnosis of malignant versus benign which can also be stated as the tumor is cancerous, true or false. Considering the literature review and observing the dependent variable is binary, a binary logistic regression would be a good choice for creating a prediction model. Another prediction model that may be effective is decision tree classification. Decision tree models also can be effective in feature selection which can be used in this model or inserted into other prediction models. To evaluate these models, a ROC plot AUC will measure accuracy performance. An AUC of 0.9 or greater is considered excellent for diagnostic accuracy (Šimundić, 2009). The null hypotheses (H0) in each model instance for this statistical test will state the ROC plot will not yield an AUC value of 0.9 or greater. The alternative hypothesis (Ha) will therefore state the ROC plot will yield an AUC value of 0.9 or greater.

### **Limitations**

The dataset selected is not large enough to provide an equally represented cross-section of the patient population. The patient data originates at the University of Wisconsin; therefore, the data is limited to the demographic distribution in that local geographic area. Diversity in age, race, environmental exposure, and ethnical background is unknown. The dataset is also quite old for patient health data. Societal, environmental, economic, political, and education factors most likely have changed since the collection of this data. Women are more aware of healthy lifestyle choices today and better lump detection techniques are now accessible.

### **Ethical Considerations**

As expected with handling and using patient healthcare data, ethical considerations need to be considered to ensure data security and privacy are intact. With the passing of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), patient health information is legally protected data here in the United States (Office of Civil Rights, 2015). This act has separate rules for privacy and security. In addition to providing guidelines on how to protect patient health information and creating punishments for mishandling such information, HIPAA also defines covered entities that do have rights to patient information and creates punishments for those who do not share their information with these covered entities. One section of the privacy rule sets guidelines for the de-identification of patient information when it is shared. These guidelines apply to the selected dataset for this project. Variables that can identify patients have to be excluded from any data sharing. Examples are patient name, birthdate, social security number, age, sex, and several others. The guidelines are not straightforward. Certain combinations of data elements can also identify patients which are prohibited. For the project dataset, only one patient identifier, variable *ID*, is provided which could be re-indexed and cannot be traced back to the patient without having proper institution credentials.

## **Findings**

Before either model is created, some descriptive statistics are calculated. Table 1 displays the summary statistics for the possible independent variables in the dataset. The summary statistics are reviewed to check for anomalies that may rule out the variable for analysis. If standard errors equal zero, or mean values appear disjointed from the value ranges, further evaluation would be needed to determine if the variable should be excluded. None of the variables here display any such characteristics. The breakdown of historical values for the dependent variable *diagnosis* is shown in Figure 1. These will be used to evaluate model performance. The goal of the models is to accurately predict new tissue specimens as either malignant or benign.

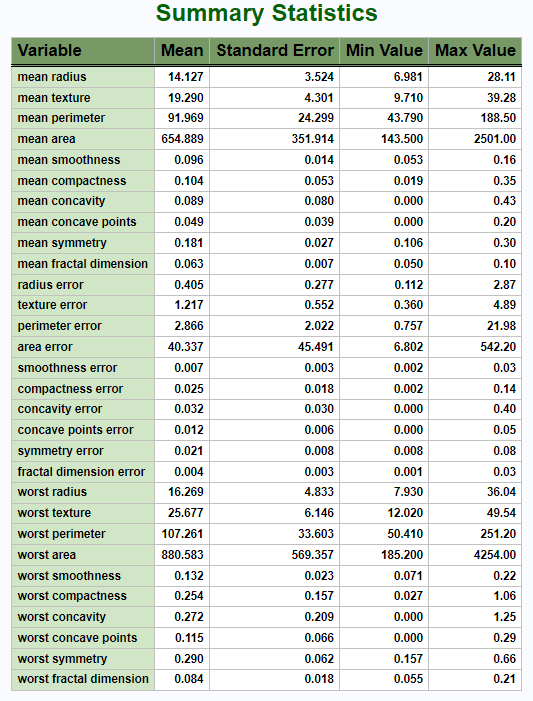
The binary logistic regression model is the first model generated for the study. This model used a stepwise variable selection process which resulted in seven variables that significantly influenced the diagnosis prediction listed in Table 2. It should be noted that the majority of the variables that most influence the model are from the worst grouping of variables that show the mean of the largest three values. The ROC curve lift for each of the steps during model generation is displayed in Figure 2. The final model yields an AUC of 0.9977. As stated in research design methods, an AUC greater than 0.9 is considered significant for diagnostic accuracy. Therefore, the logistic regression model rejects the null hypothesis.

The second model chosen classifies the binary diagnosis options using a classification decision tree. The final model fit statistic and performance is displayed in Table 3. The SAS Studio function PROC HPSPLIT also produces a confusion matrix as an added performance test shown at the top of Table 3. The error rates for false positive and false negative predictions were both below five percent. The AUC result from the ROC curve is provided at the end of the fit statistics and returned a value of 0.9795. ROC curve plot itself is shown in Figure 3. This model only lists the top five variables of importance which once again show the significance of the *worst* values. Once again, the AUC from the ROC curve is greater than 0.9, thus rejecting the null hypothesis for the statistical test.

Based on the observations that both statistical tests rejected the null hypothesis, the research question null hypothesis can also be rejected. Clinically significant prediction results were obtained from both prediction models meaning malignancy can be predicted by using cell nuclei measurements from fine-needle breast biopsy tissue images. As for the second research question asking if one model outperformed another, the comparison of the AUC values cannot be considered distinguishable. Therefore, the null hypothesis for the second research question is accepted. It would be expected that either model performance would be similar.

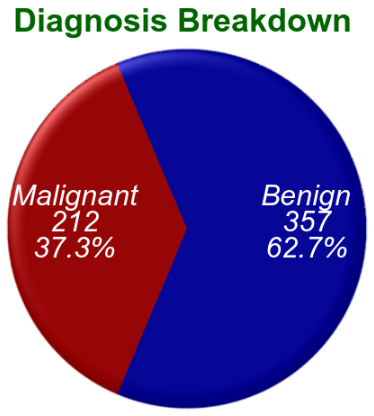
**Table 1**

*Summary Statistics for Independent Variables*



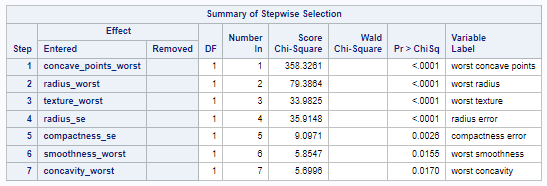
**Figure 1**

*Diagnosis Breakdown*



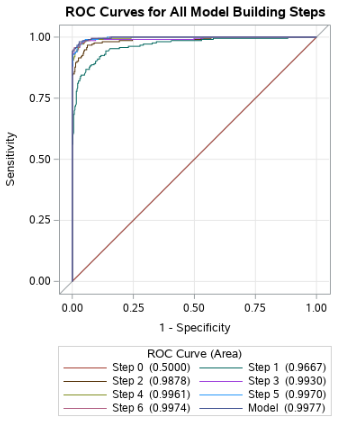
**Table 2**

*Logistic Regression Results: Variable Importance*



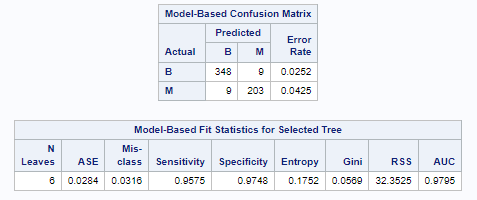
**Figure 2**

*Logistic Regression ROC Curve (All Model Building Steps)*



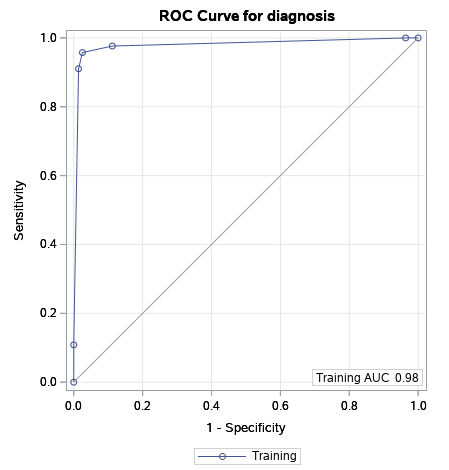
**Table 3**

*Classification Tree Model Confusion Matrix and Fit Statistics*



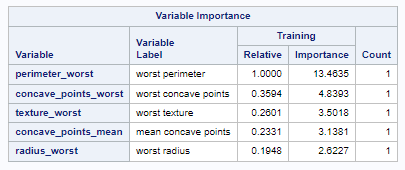
**Figure 3**

*ROC Curve for Classification Tree*



**Table 4**

*Classification Tree Variable Importance*



## **Conclusion**

When considering past literature and the results from this study’s statistical test models, one can conclude that data analytic models could indeed be used to predict and classify breast tissue biopsy malignancy thereby diagnosing breast cancer accurately. What remains to be seen is if the analytic models can achieve a high enough and consistent enough performance to be adopted and replace current medical examination standards of practice. AI hybrid models appear to have the best chance to accomplish this having the highest performances averaging above 99 percent. However, just as in other AI projects (self-driving cars are a good example), will these models gain enough human trust to be adopted even though they already outperform humans?

**Recommendations**

## **References**

Al-Azzam, N., & Shatnawi, I. (2021). Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. *Annals of Medicine and Surgery, 62*, 53–64. <https://doi.org/10.1016/j.amsu.2020.12.043>.

Bevers, T. B., Helvie, M., Bonaccio, E., Calhoun, K. E., Daly, M. B., Farrar, W. B., Garber, J. E., Gray, R., Greenberg, C. C., Greenup, R., Hansen, N. M., Harris, R. E., Heerdt, A. S., Helsten, T., Hodgkiss, L., Hoyt, T. L., Huff, J. G., Jacobs, L., Lehman, C. D., Monsees, B., Niell, B. L., Parker, C. C., Pearlman, M., Philpotts, L., Shepardson, L. B., Smith, M. L., Stein, M., Tumyan, L., Williams, C., Bergman, M. A., & Kumar, R. (2018). Breast Cancer Screening and Diagnosis, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology, *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw, 16*(11), 1362-1389. Retrieved May 21, 2022, from <https://jnccn.org/view/journals/jnccn/16/11/article-p1362.xml>.

Breastcancer.org. (2022, March 10). *Breast cancer facts and statistics.* Breastcancer.org. <https://www.breastcancer.org/facts-statistics>.

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. <http://archive.ics.uci.edu/ml>.

Kirola, M., Memoria, M., Dumka, A., Tripathi, A., & Joshi, K. (2022). A Comprehensive Review Study on: Optimized Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction in Big Data Context. *Biomedical and Pharmacology Journal, 15*(1), 13. <https://doi.org/10.13005/bpj/2339>.

Mohammed, Y. A. (2020). Identifying the Most Influential Variables in Breast Cancer Using Logistic Regression. *Indian Journal of Public Health Research & Development, 11*(2), 1968–1972. <https://doi.org/10.37506/v11/i2/2020/ijphrd/195119>.

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science, 191*, 487–492. <https://doi.org/10.1016/j.procs.2021.07.062>.

Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine, 127*, 102276. <https://doi.org/10.1016/j.artmed.2022.102276>.

Office of Civil Rights. (2015, November 6). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. HHS.gov. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>.

SAS Institute. (n.d.). *SAS studio, SAS OnDemand for academics* (Release 9.04.01M6P11072018 for Linux LIN X64 3.10.0-1062.9.1.el7.x86\_64). SAS.com.

<https://www.sas.com/en_us/home.html>.

Šimundić A. M. (2009). Measures of diagnostic accuracy: Basic definitions. *EJIFCC, 19*(4), 203–211. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975285//>.

Venkatesh, V., Anishin Raj, M. M., Mohamed Sajith, K., Anushiadevi, R., Suriya Praba, T., Varadarajan, V., Kommers, P., Piuri, V., & Subramaniyaswamy, V. (2020). A precision-based diagnostic model ADOBE-accurate detection of breast cancer using logistic regression approach. *Journal of Intelligent & Fuzzy Systems, 39*(6), 8419–8426. <https://doi.org/10.3233/JIFS-189160>.

Wolberg, W., Street, N., & Mangasarian, O. (1995). *Wisconsin Diagnostic Breast Cancer (WDBC)* [data set]. University of Wisconsin – Madison. [ftp.cs.wisc.edu](ftp://ftp.cs.wisc.edu).