Applications of Machine Learning Models on the Original Wisconsin Breast Cancer Dataset.

PRYANGKA RAO BATUMALAY

501001811

CKME 136

CAPSTONE PROJECT

TABLE OF CONTENTS

INTRODUCTION

Breast cancer is the second leading cause of cancer related death worldwide. The American Cancer Society has estimated 276,480 new breast cancer cases for the year of 2020 (The American Cancer Society, 2020) while The Canadian Cancer Society predicts 27, 400 impending breast cancer cases will heavily impact both the American and Canadian hospitals, especially with the current status of coronavirus (The Canadian Cancer Society, 2020). Hospitals and clinics are experiencing an influx of COVID patients which trumps over all cancer care patients (Grant, 2020). Cancer patients are also more susceptible to COVID due to their compromised immune systems (The Canadian Cancer Society, 2020). Therefore, it is vital to attain quick and accurate diagnosis immediately to alleviate the burdens on healthcare staff and breast cancer patients.

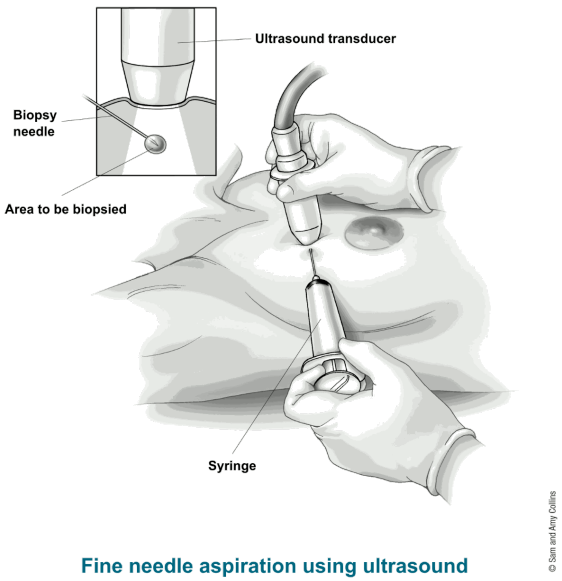


Figure 1: Image of fine needle aspiration (FNA) carried out with an ultrasound (The American Cancer Society, 2017)

Needle Biopsy Types

There are two needle biopsies which are commonly done to detect breast cancer. Fine Needle Aspiration (FNA) is an important biopsy used in the pre diagnosis of breast cancer. In FNA, the sample is collected through a hollow needle which is connected to a syringe. Ultrasounds may be used as a guide in the event the sample area is deep or difficult to locate (The American Cancer Society, 2020). Core Needle Biopsy (CNB) on the other hand, begins with an incision of about a quarter inch on the breast area. A hollow needle which is much wider than the FNA is used to allow more samples to pass and is connected to a spring loaded tool or suction device (The American Cancer Society, 2020) . This method is often repeated multiple times to get more samples.

FNA is non-invasive, fast and a more cost effective tool when compared to CNB in detecting breast cancer. It is a simple step to determine if the sample is a cyst or cancerous within 15 seconds (The American Cancer Society, 2020). However, only a small amount of tissue samples can be obtained via this method, affecting its accuracy resulting in the invasive counterpart, the CNB to be the preferred biopsy choice (Casaubon and Regan, 2019).

With the projected increase in breast cancer incidence during current pressing times with COVID-19, it is crucial to distinguish the incidence of benign and malignant cases by implementing a biopsy which is more cost effective, quick and ideally non-invasive to attain accurate diagnosis at an early stage before carrying out more detailed testings. FNA would have been an ideal fit with the exception of attaining a small amount of samples.

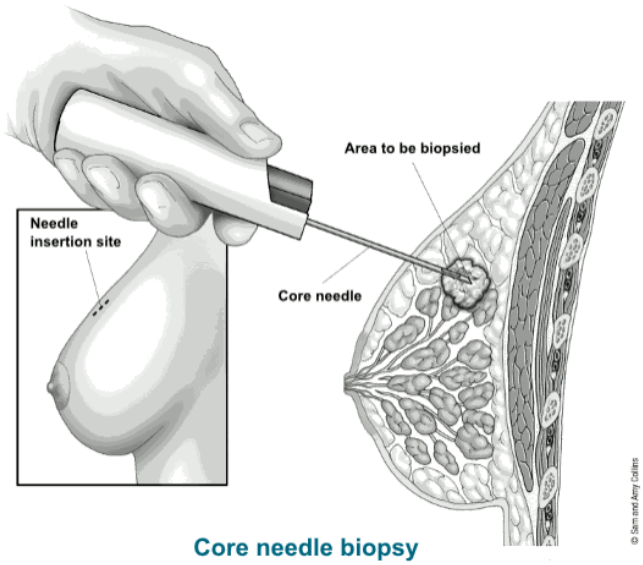


Figure 2: Image of core needle biopsy (CNB). Notice the three invasive incisions made to extract the samples. (The American Cancer Society, 2017)

The goal of this project is to ease the oncology departments burden by assisting in the diagnosis of benign and malignant breast cancer using the economical, quick and non-invasive FNA biopsy. Machine learning models are applied on the Wisconsin Breast Cancer Original (WBCO) dataset where the breast cell samples were obtained through FNA. As time is of the essence, relevant cytological characteristics which affect the diagnosis need to be determined to reduce unnecessary data which will impact the model's performance. The models will be tested with these features and the accuracy in diagnosing the breast cancer types will be used to select the best performing model. By training the model to detect breast cancer types with small samples, FNA can be used as the main form of biopsy for detecting breast cancer type.

LITERATURE REVIEW

Fine-Needle Aspiration Cytology Can Play a Role in Neoadjuvant Chemotherapy in Operable Breast Cancer (Garbar, C, and H Curé, 2013).

This literature distinguishes the advantages and disadvantages for FNA and CNB. The following selected considerations table was obtained from the literature. The literature credits FNA as a safe and effective method to evaluate breast cancer. As noted, FNA is more time efficient but produces unsatisfactory samples. It is also noted that it fails to distinguish benign and malignant cancer cells directly. With proper machine learning tools, this con can be overturned.

Table 1: The Advantages and Disadvantages of FNA and CNB

Considerations	**FNA**	**CNB**
Quick diagnosis	Yes	No
Pain or any discomfort	Very low	Low
Distinction of benign and malignant cells	No	Yes
Complication Rate	Very low	Low
Satisfactory samples	Low	High

Table 1: Differences between FNA and CNB. While FNA is unable to distinguish between benign and malignant cells directly, the diagnosis can be made with proper machine learning tools.

Fine Needle Aspiration Of Breast Masses (Casaubon and Regan, 2019).

This paper a comparison between FNA and core needle biopsy (CNB) is done. It is determined that CNB produces better results but is invasive and expensive. Therefore, FNA, which is non-invasive, cost effective and fast, tends to be the method of choice. The down side of FNA however is that only a small sample can be obtained which may affect the accuracy of the diagnosis.

Machine learning applications in cancer prognosis and prediction (Kourou et al., 2015).

This is an extremely useful overview which discusses multiple machine learning concepts which can be used in the capstone project. While focusing on the supervised methods, preprocessing steps such as dimensionality reduction, feature selection and feature extraction were explained followed by multiple ways to train and test the model. Model performance can be measured using accuracy, sensitivity, specificity, and area under the curve (AUC). In this overview, machine learning techniques from multiple literature were compared and analyzed such as artificial neural networks, support vector mechanisms, bayesian network and decision trees were considered to detect cancer susceptibility, recurrence and survival to model the cancer risks outcomes. The literature acknowledges that although many machine learning studies are carried out for the detection of cancer, it does not permeate clinical practices as more accurate validated results are required. This is good insight that further validations of the results must be carried out.

Machine learning techniques for personalized breast cancer risk prediction: comparison with BCRAT and BOADICEA models (Ming et al., 2019).

Although this study does not use WBCO, it utilizes ML methods and techniques which can be used in the capstone project. This study compares eight machine learning techniques with pre existing methods called the Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models. Of the eight methods, five are considered to be model based (logistic regression, generalized linear models, linear discriminant analysis, Markov Chain Monte Carlo generalized linear model and quadratic discriminant analysis) and three were model-free based machine learning techniques (adaptive boosting, random forest and k-nearest neighbours). R program is used to rebalance the data using the SMOTE and “unbalanced” method. The study finds that the ML models proved to be more accurate than the BCRAT and BOADICIEA. The literature carried out variable rankings with the models will affect its performance. It will be a good idea to see how features selections will affect the accuracy of each model.

Breast Cancer Detection Using Machine Learning Algorithms (Sharma et al., 2018).

This study utilizes random forest, naive bayes and k-nearest neighbours (kNN) to determine the best model using WBCD (Diagnostic Version) on Python. Each model was trained and tested using the 10-fold cross validation and the performance was analysed using the confusion matrix. In this study, kNN was determined to be the best model. This is another literature which utilizes cross validation to analyze the results as well.

DATASET

The Wisconsin Breast Cancer Original (WBCO) Dataset was obtained from the UCI Machine Learning Repository. It provides 699 FNA samples with 11 cytological features to which 458 (65.5%) are classified as benign and 241 (34.5%) as malignant samples.

The initial review of the dataset shows the following list of variables and its description is noted in the table below (Table 2). The following table also gives a general description of the variables and how it may impact the decision if the cell is cancerous or not (Karaa and Dey, 2015). All the variables data types, range and descriptions are listed as well. As it has labeled data, a supervised binary classification approach will be implemented.

Table 2 : Dataset Variable Descriptions

Variables	**Description**	**Data Type**	**Range**
Sample code number	The unique ID number for each samples	Continuous	ID
Clump Thickness	Benign cells tend to be mono-layered while malignant cells tend to be multilayered	Discrete	1-10
Uniformity of Cell Size	Malignant cells tend to have abnormal cell size	Discrete	1-10
Uniformity of Cell Shape	Malignant cells tend to have abnormal cell shape	Discrete	1-10
Marginal Adhesion	Malignant cells do not bond well with other cells compared to normal cells	Discrete	1-10
Single Epithelial Cell Size	Malignant epithelial cells tend to be inflamed	Discrete	1-10
Bare Nuclei	Nuclei not enclosed in cytoplasm tend to be benign.	Discrete	1-10
Bland Chromatin	Chromatin texture are coarser with malignant cells	Discrete	1-10

Normal Nucleoli	Cancer cell have larger and more nucleoli compared to normal cells	Discrete	1-10
Mitoses	Process of cell division. Cancer cells are known to have abnormal division rates.	Discrete	1-10
Class	Benign cells were categorized as 2 Malignant cells are 4	Discrete	2 - Benign 4 - Malignant

Table 2: The description of the datasets from the Wisconsin Breast Cancer Original Data. The description of each variable with the exception of ID Number and Class. The characteristics of each variable is noted as well (Karaa and Dey, 2015)

Using Python on the Jupyter Notebook, the following analysis on the WBCO data was obtained. There are 16 missing values which need to be addressed in the Bare Nuclei variable.

```
WBC.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 699 entries, 0 to 698
Data columns (total 11 columns):
ID                             699 non-null int64
Clump Thickness                699 non-null int64
Uniformity of Cell Size        699 non-null int64
Uniformity of Cell Shape       699 non-null int64
Marginal Adhesion              699 non-null int64
Single Epithelial Cell Size    699 non-null int64
Bare Nuclei                    683 non-null float64
Bland Chromatin                699 non-null int64
Normal Nucleoli                699 non-null int64
Mitoses                        699 non-null int64
Class                          699 non-null int64
dtypes: float64(1), int64(10)
memory usage: 60.1 KB
```

```
#Determine where the missing values based on variables
WBC.isnull().sum()
```

```
ID                             0
Clump Thickness                0
Uniformity of Cell Size        0
Uniformity of Cell Shape       0
Marginal Adhesion              0
Single Epithelial Cell Size    0
Bare Nuclei                   16
Bland Chromatin                0
Normal Nucleoli                0
Mitoses                        0
Class                          0
```

Figure 3: Snippet of Jupyter Notebook to determine missing values.

An initial simple statistical pythonic code is applied to describe the dataset is shown below.

```
# Describe the dataset.
WBC.describe()
```

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
count	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000	683.000000
mean	4.442167	3.150805	3.215227	2.830161	3.234261	3.544656	3.445095	2.869693	1.603221	0.349927
std	2.820761	3.065145	2.988581	2.864562	2.223085	3.643857	2.449697	3.052666	1.732674	0.477296
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	2.000000	1.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	1.000000	0.000000
50%	4.000000	1.000000	1.000000	1.000000	2.000000	1.000000	3.000000	1.000000	1.000000	0.000000
75%	6.000000	5.000000	5.000000	4.000000	4.000000	6.000000	5.000000	4.000000	1.000000	1.000000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	1.000000

Figure 4: Snippet of Jupyter Notebook listing statistical findings.

More information on the dataset will be seen in the coding and findings section of the project.

APPROACH

The following approach was carried out for the project:-

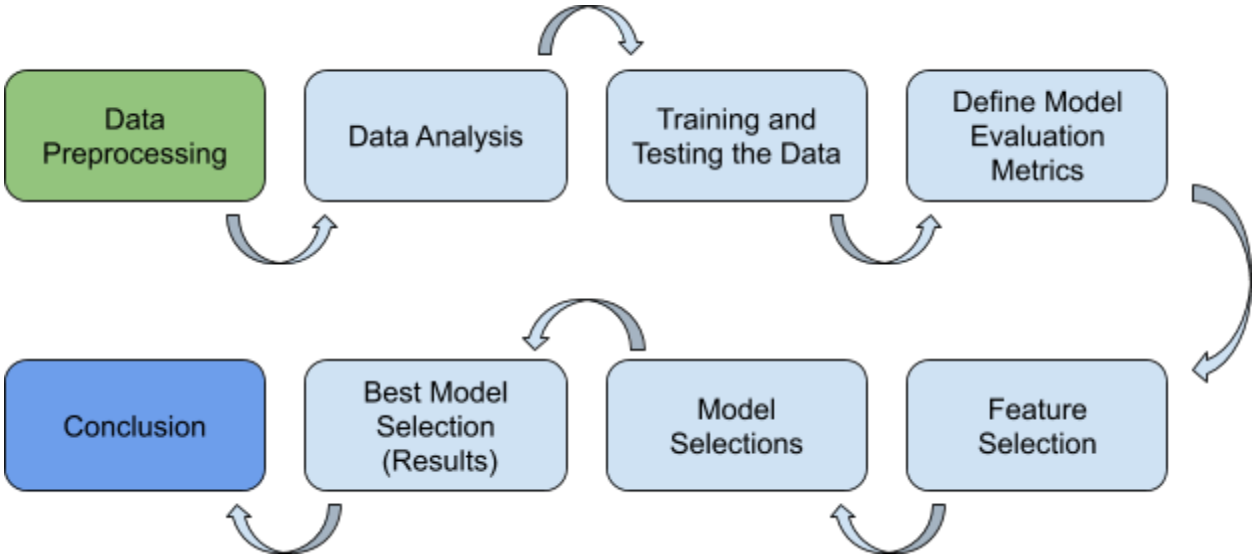


Figure 5 shows the approach taken for the overall project.

The entire project was carried out with Python on Jupyter Notebook for simplicity.

Step 1: Data Preprocessing

The following steps were carried out to initially

1. The data was uploaded into the Jupter notebook. All missing values which were denoted as “?” were converted to NaN.

```
#Upload dataset on pandas
WBC = pd.read_csv('/Users/pryangkarao/Desktop/Wisconsin Breast Cancer Python/breast-cancer-wisconsin.data',
                  header=None, na_values="?")

#We ensure that the header is set as None and all missing values '?' were converted to NaN.
```

Figure 5: Snippet on the Jupyter Notebook on data import

2. Appropriate headers are assigned to the dataset.

3. Remove missing values - Missing values in the dataset were denoted as “?”. This needs to be addressed when importing the dataset into the environment. Once the missing dataset is determined, we will drop it as there are only 16 of them.

4. Drop unnecessary variables such as ID variable has no effect on our model

5. Recategorize Benign as 0 and Malignant as 1 for simplicity

Once all the steps have been carried out, we will have the following dimensions.

```
#Check on new dimensions of the dataset
WBC.shape
```

```
(683, 10)
```

```
# That we removed all missing values, we can allocate 'Bare Nuclei' as an integers instead of float.
WBC['Bare Nuclei'] = WBC['Bare Nuclei'].astype('int')
WBC.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 683 entries, 0 to 698
Data columns (total 10 columns):
Clump Thickness               683 non-null int64
Uniformity of Cell Size       683 non-null int64
Uniformity of Cell Shape      683 non-null int64
Marginal Adhesion             683 non-null int64
Single Epithelial Cell Size   683 non-null int64
Bare Nuclei                   683 non-null int64
Bland Chromatin               683 non-null int64
Normal Nucleoli               683 non-null int64
Mitoses                       683 non-null int64
Class                         683 non-null int64
dtypes: int64(10)
memory usage: 58.7 KB
```

Figure 6: Snippet on the Jupyter Notebook on the data dimensions

There are now 444 Benign cases (65%) and 239 Malignant cases (35%) in the new dataset.

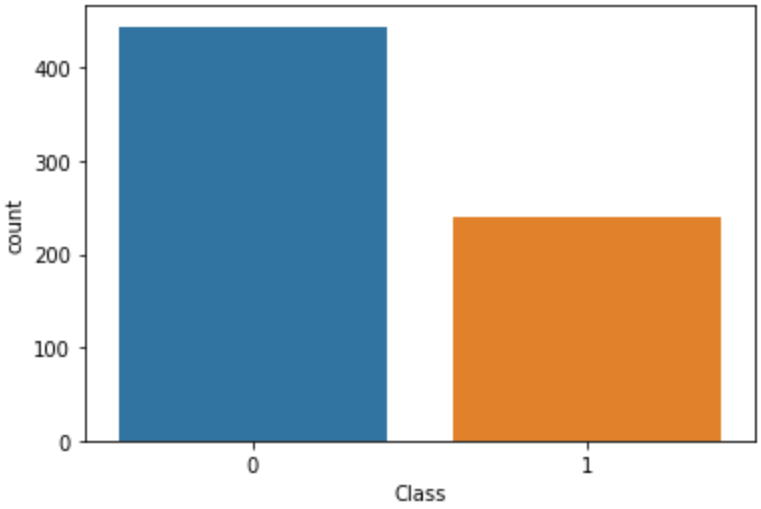


Figure 7: Histogram on class distribution. There are 444 Benign cases and 239 Malignant cases.

Step 2: Data Analysis

Statistical Analysis

First we will carry out statistical analysis using histograms

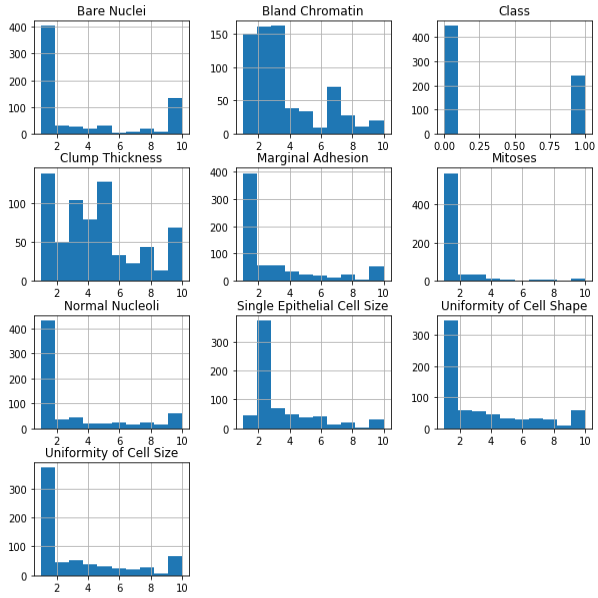


Figure 8: Histograms on all variables.

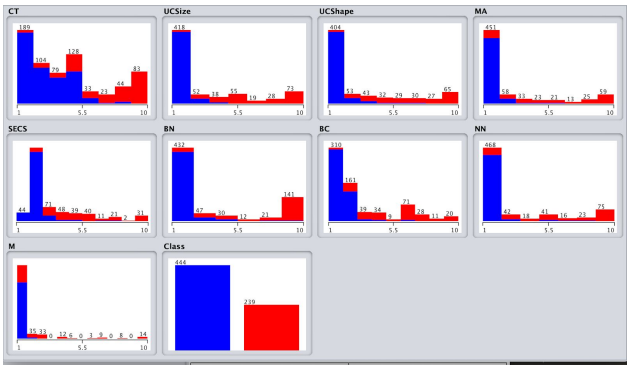
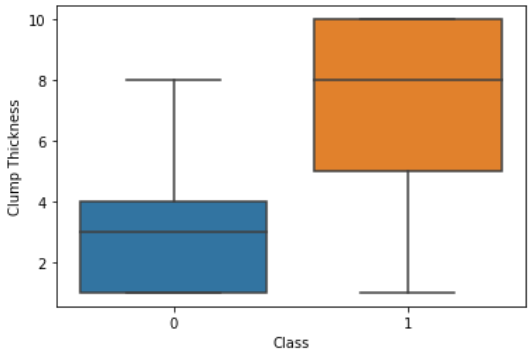


Figure 9: Histograms on all variables seen on WEKA by Class.

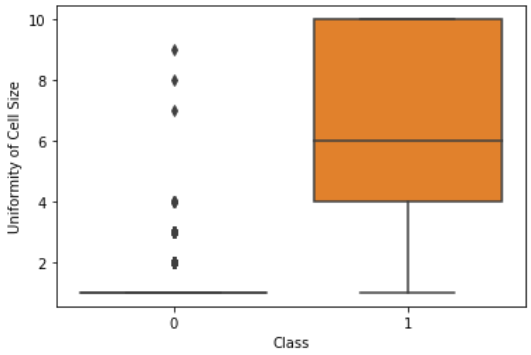
Looking at the histograms obtained through WEKA which shows the distribution of each class, Benign (Blue) and Malignant (Red), we can see that the distribution for each class is somewhat skewed, especially the benign cases.

The boxplots for each variable shown below. We can see the outliers if it exists.

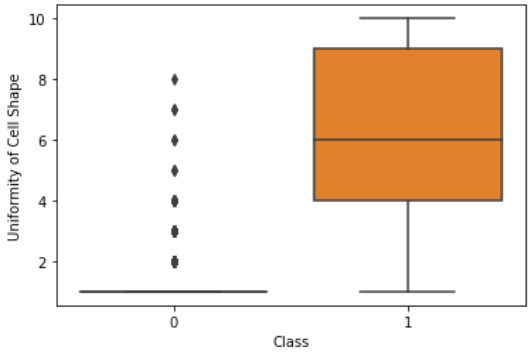
1) Figure 10: Boxplot of Clump Thickness



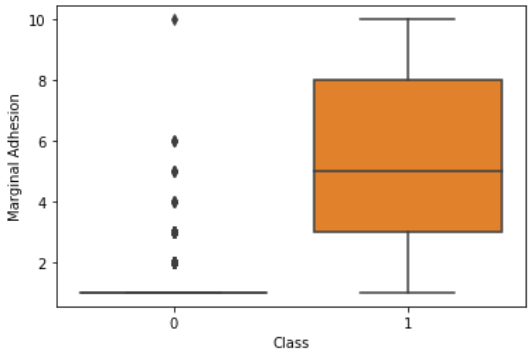
2) Figure 11: Boxplot of Uniformity of Cell Size



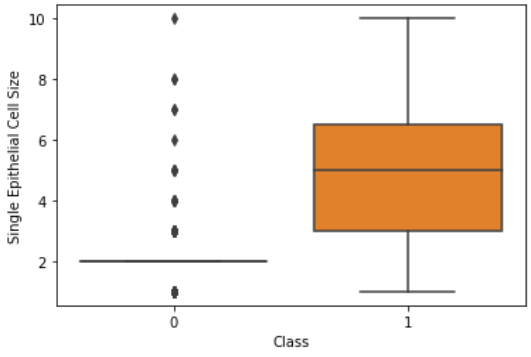
3) Figure 12: Boxplot of Uniformity of Cell Shape



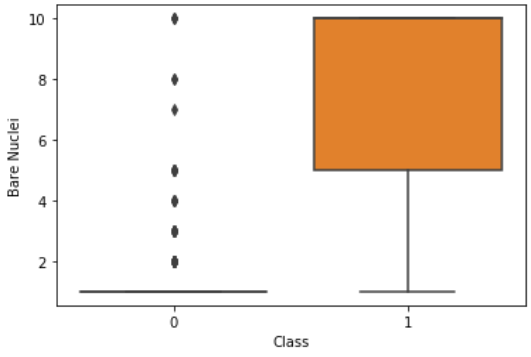
4) Figure 13: Boxplot of Marginal Adhesion



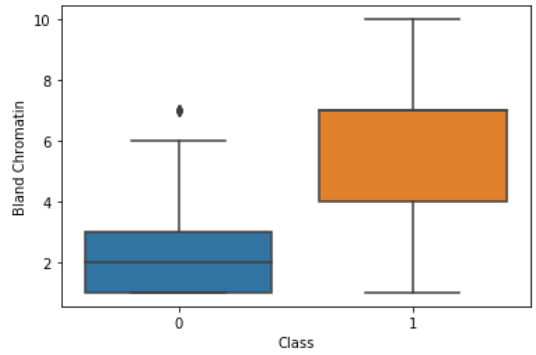
5) Figure 14: Boxplot of Single Epithelial Cell Size



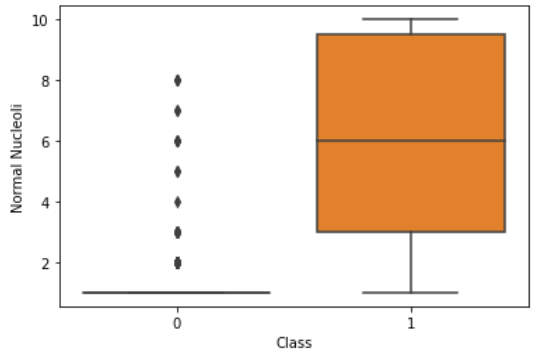
6) Figure 15: Boxplot of Bare Nuclei



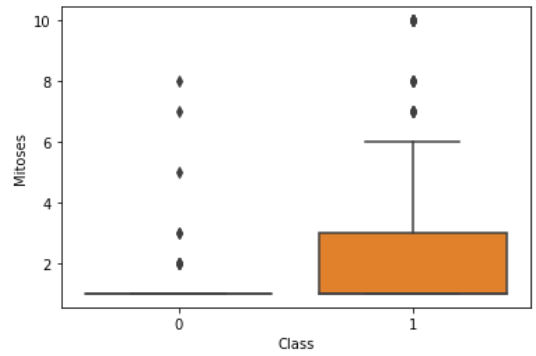
7) Figure 16: Boxplot of Bland Chromatin



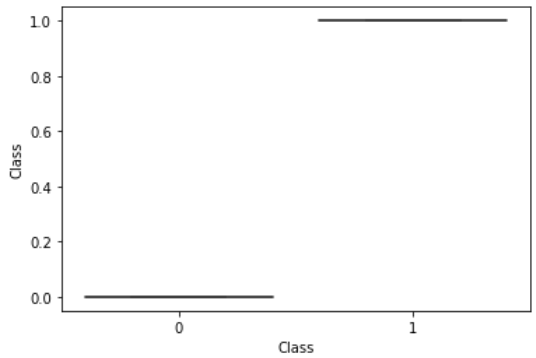
8) Figure 17: Boxplot of Normal Nucleoli



9) Figure 18: Boxplot of Mitoses



10) Figure 19:Boxplot of Class



Looking through the boxplots, we can see that benign cases have more outliers than the malignant cases especially in the mitoses, normal nucleoli, uniformity of cell size, uniformity of cell shape, single epithelial cell size, and marginal adhesion.

1) Attribute Relationship

To further visualize the dataset, a pairplot and correlation table is done by separating the distinct 'Class' variable.

Figure 20: Pair Plot By Class Type - Benign and Malignant

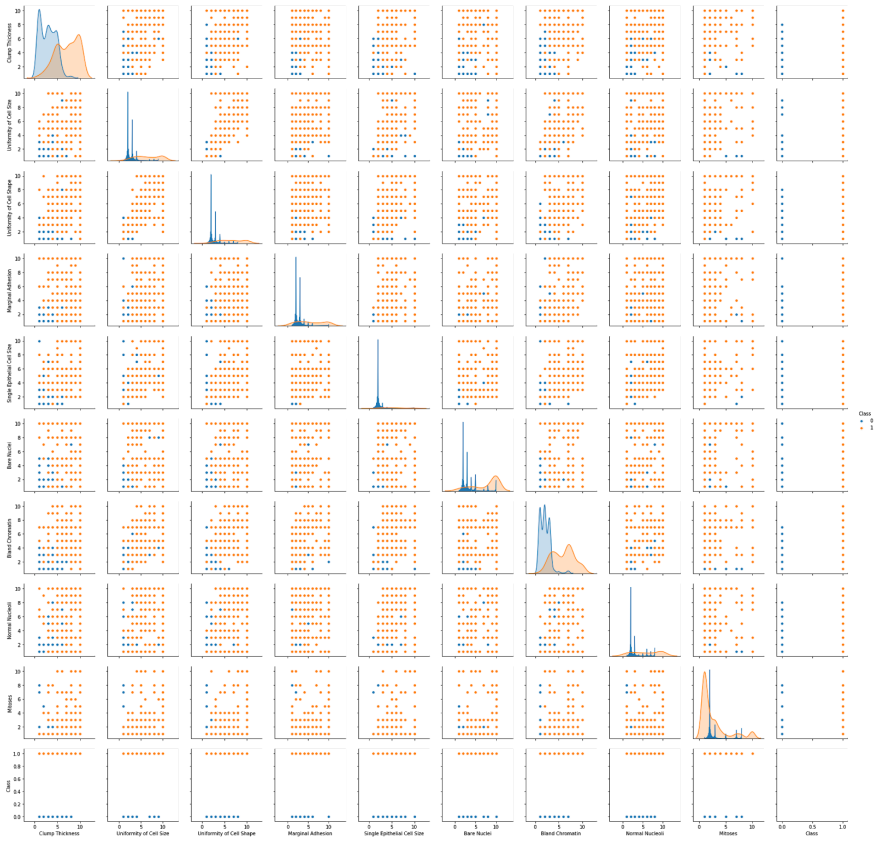


Figure 20 shows the pairplot obtained by class type of benign and malignant.

Figure 21: Correlation Plot of the Dataset

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
Clump Thickness	1	0.642481	0.65347	0.487829	0.523596	0.593091	0.553742	0.534066	0.350957	0.71479
Uniformity of Cell Size	0.642481	1	0.907228	0.706977	0.753544	0.691709	0.755559	0.719346	0.460755	0.820801
Uniformity of Cell Shape	0.65347	0.907228	1	0.685948	0.722462	0.713878	0.735344	0.717963	0.441258	0.821891
Marginal Adhesion	0.487829	0.706977	0.685948	1	0.594548	0.670648	0.668567	0.603121	0.418898	0.706294
Single Epithelial Cell Size	0.523596	0.753544	0.722462	0.594548	1	0.585716	0.618128	0.628926	0.480583	0.690958
Bare Nuclei	0.593091	0.691709	0.713878	0.670648	0.585716	1	0.680615	0.58428	0.33921	0.822696
Bland Chromatin	0.553742	0.755559	0.735344	0.668567	0.618128	0.680615	1	0.665602	0.346011	0.758228
Normal Nucleoli	0.534066	0.719346	0.717963	0.603121	0.628926	0.58428	0.665602	1	0.433757	0.718677
Mitoses	0.350957	0.460755	0.441258	0.418898	0.480583	0.33921	0.346011	0.433757	1	0.423448
Class	0.71479	0.820801	0.821891	0.706294	0.690958	0.822696	0.758228	0.718677	0.423448	1

Figure 21 shows the detailed correlation between the variables.

Looking into the heatmap, we can get a better visualization of the correlated variables as well.

Figure 22: Heatmap Variable Correlation

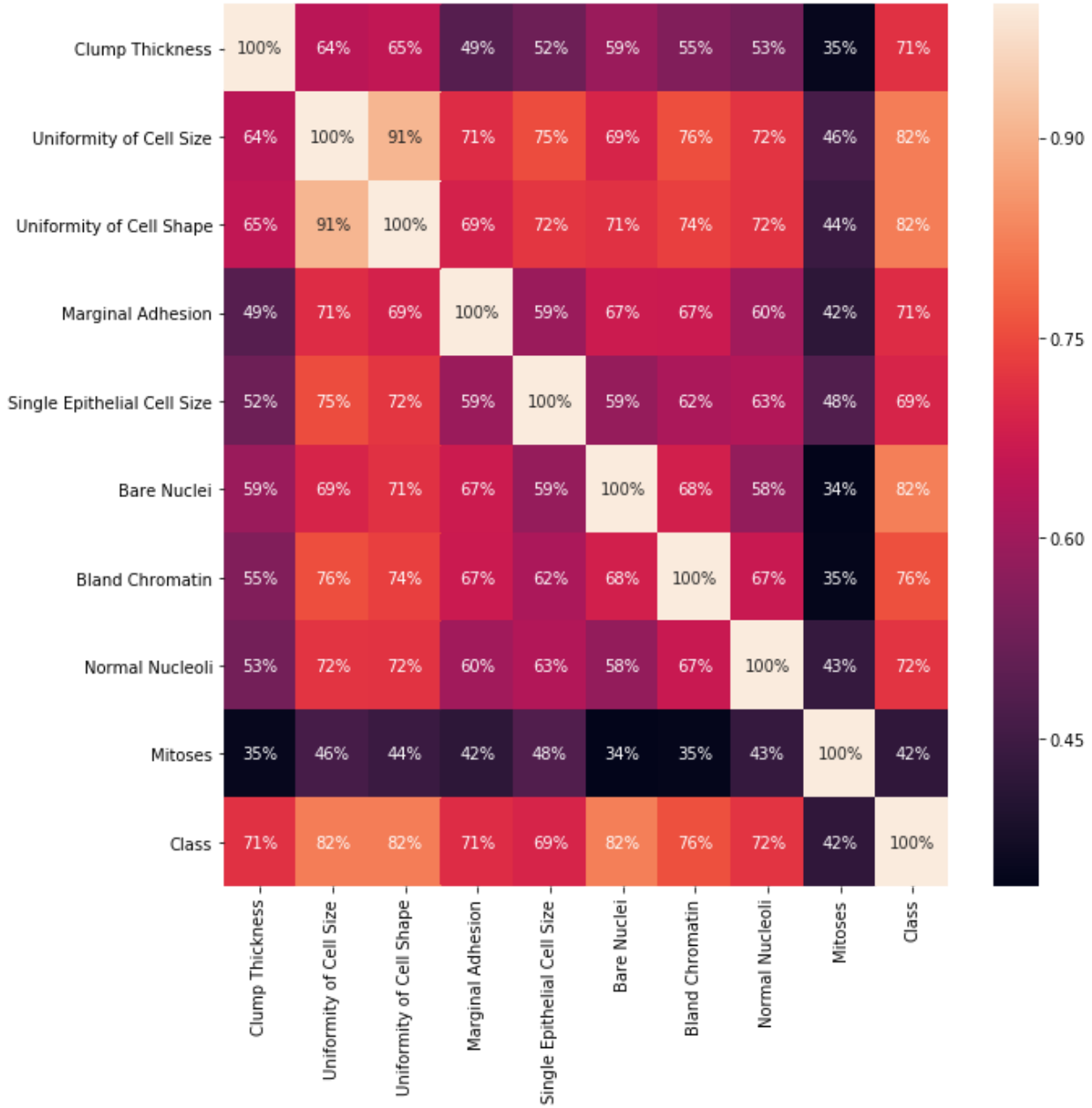


Figure 22 shows the heatmap of the variable correlation in %. The warmer (red) the indicator, the higher the correlation.

Looking at the heatmap and correlation table , we can see the class correlation with other variables. The following is the arrangement of variables affecting the class.:-

1) Bare Nuclei (82.23%)
2) Uniformity of cell shape (82.19%)
3) Uniformity of cell size (82.02%)
4) Bland Chromatin (75.82%)
5) Normal Nucleoli (71.87%)
6) Clump Thickness (71.48%)
7) Marginal Adhesion (70.63%)
8) Single Epithelial Cell Size (69.10%)
9) Mitoses (42.34%)

We can see there are strong correlations between 2 variables with the following order :-

1) Uniformity of Cell Size and Uniformity of Cell Shape (91%)
2) Bland Chromatin and Uniformity of Cell Size (76%)
3) Uniformity of Cell Size and Single Epithelial Cell Size (75%)

Step 3: Train and Test Data

Before attempting the models the datasets were split into train and test data. Using a 80/20 split (80% training and 20% testing), the is now separated to 137 test data and 546 train data. As the value in the datasets are discrete, standard scaling need not be performed as the data are within a good range.

Step 4: Defining Model Evaluation Metrics

With the train and rest data defined, a metric to determine the performance of the model is considered. The cost of error is important especially when the result will lead to fatalities. For an example, there are huge repercussions for the misdiagnosis of a patient with malignant breast cancer as benign cancer versus the diagnosis of a benign patient as malignant in this project scenario.

1) Confusion Matrix.

The confusion matrix takes into consideration the instance of correct and incorrect classified diagnosis for the models. These instances are divided into true positives and true negatives (correctly classified) and false positives and false negatives (incorrectly classified).

Table 3 : WCBO Confusion Matrix

Diagnosis	Predicted Benign - 0	Predicted Malignant - 1
Actual Benign - 0	TN	FP
Actual Malignant - 1	FN	TP

Table 3 shows the confusion matrix for the models tested for diagnosis classification.

For this project, the confusion matrix would provide the following breakdowns:-

True Positive (TP) : The predicted class and the actual class is Malignant
False Positive (FP) : The predicted class is Malignant and the actual class is Benign
True Negatives (TN) : The predicted class and the actual class is Benign
False Negatives (FN) : The predicted class is Benign and the actual class is Malignant

Therefore, in this scenario, the lower the false negative rate, the better.

2) Accuracy.

By far this is the most popularly used metrics to test model performance. From the confusion matrix, the following formula can then be applied :-

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

However, as 444 of the samples are benign (65%) and 239 are malignant (35%), there is a slight imbalance in the dataset which may affect the accuracy of the models.

3) Precision

This metric works best for false positive instances. Of the samples classified as malignant, how many were actually malignant.

$$Precision = (TP) / (TP + FP)$$

4) Recall

This method works best when the focus is on the false negatives. In this scenario, out of the samples classified as malignant, how many did the model classified correctly. The higher the recall indicates a lower false negative.

$$Recall = (TP) / (TP + FN)$$

5) F1 Score

A combination of both the recall and precision of the model to measure the performance of the model.

$$F1\ Score = 2/((1/Recall)+(1/Precision))$$

Recall is therefore the better metric for this project as the incidence of sample misclassification as benign instead of malignant needs to be monitored the most.

Step 5: Feature Selection

Using the Select K Best feature, each variable will be scored to determine its importances in the model's decisions.

	Features	Scores
5	Bare Nuclei	1729.066174
1	Uniformity of Cell Size	1370.064587
2	Uniformity of Cell Shape	1279.767704
7	Normal Nucleoli	1143.866712
3	Marginal Adhesion	986.417879
6	Bland Chromatin	682.978239
0	Clump Thickness	624.135704
4	Single Epithelial Cell Size	497.536763
8	Mitoses	228.994346

Figure 23:Snippet of feature selection ranking from Jupyter notebook

Now that we have the feature importance, we can perform backward elimination, which is the removal of variables from the least important to the most important. Then the analysis of the accuracy of the model based on reduction of variables can be tabulated. Also note that there are minor differences

between the rankings attained through this step in comparison to the rankings obtained with the correlation table.

Step 6: Model Selection.

The following models were used to test the dataset in hand. In this project, supervised classification techniques are applied as the outcomes of cell types are known to be either benign or malignant. Therefore, the following models are taken into consideration:-

Table 4 : Supervised Classification Models Used

Supervised Classification Models	Reasonings
Logistic Regression	Generally used for binary classifications.
Random Forest	Obtain robust model which is much less likely to overfit
Support Vector Machine	This model works well for binary labeled points.
Naive bayes	Popularly used model which is good for multivariate datasets
Decision Tree	Easy to interpret especially with 9 variables
K Nearest Neighbors	Commonly used supervised classification technique

Table 4 shows the models used for this project and the reasons they were selected.

The following comparison table was obtained with all the models. Variables with the lowest importance (from Mitoses to Uniformity of Cell Size) are then removed and the recall and accuracy was tested again.

RESULTS

Best Performing Model Selection

The following comparison tables were obtained for both recall and accuracy by applying the backward elimination to remove the least important variable as determined earlier using the KBestFeatures. Highlighted in pink are the highest scores results obtained on recall and accuracy for each model.

RECALL

Table 5 : Model Recall Scores

	Variables	Logistic Regression	Random Forest	SVM	Naive Bayes	Decision Tree	KNN
0	All Features	0.896552	0.913793	0.913793	0.948276	0.87931	0.896552
1	Features -1	0.87931	0.931034	0.931034	0.965517	0.862069	0.913793
2	Features -2	0.87931	0.931034	0.931034	0.965517	0.896552	0.896552
3	Features -3	0.844828	0.896552	0.87931	0.931034	0.87931	0.896552
4	Features -4	0.844828	0.896552	0.87931	0.913793	0.87931	0.896552
5	Features -5	0.862069	0.896552	0.87931	0.913793	0.87931	0.913793
6	Features -6	0.844828	0.913793	0.87931	0.896552	0.810345	0.896552
7	Features -7	0.862069	0.913793	0.87931	0.87931	0.896552	0.913793
8	Features -8	0.724138	0.810345	0.810345	0.810345	0.810345	0.706897

Table 5 shows the comparison table of each model and their recall scores by carrying out backward elimination.

From the table 5, Naive Bayes has the best recall scores when mitoses and/or single epithelial cell size was removed with 96.6%. This is followed by both random forest and support vector machine with the same variable constraint at 93.1% recall score. The weakest recall score was seen in both logistic regression (89.7%) and decision tree with two and seven variables removed (89.7%).

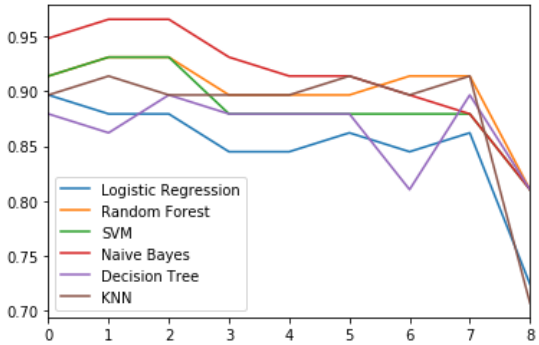


Figure 24: The comparison plot of each model and their recall scores by carrying out backward elimination.

ACCURACY

As the Naive Bayes had two same valued high recall scores with the removal of 2 and three variables, the accuracy score was generated for the models as well to see the best model. This is then compared combined with the recall score to determine the best model.

Table 6 : Model Accuracy Scores

	Variables	Logistic Regression	Random Forest	SVM	Naive Bayes	Decision Tree	KNN
0	All Features	0.948905	0.956204	0.948905	0.956204	0.934307	0.948905
1	Features -1	0.941606	0.963504	0.956204	0.970803	0.927007	0.956204
2	Features -2	0.941606	0.963504	0.956204	0.963504	0.941606	0.948905
3	Features -3	0.927007	0.934307	0.934307	0.948905	0.934307	0.941606
4	Features -4	0.927007	0.934307	0.934307	0.948905	0.941606	0.941606
5	Features -5	0.927007	0.934307	0.934307	0.941606	0.934307	0.948905
6	Features -6	0.919708	0.941606	0.927007	0.934307	0.905109	0.934307
7	Features -7	0.927007	0.948905	0.927007	0.927007	0.948905	0.941606
8	Features -8	0.868613	0.89781	0.89781	0.89781	0.89781	0.824818

Table 6 shows the comparison table of each model and their accuracy performance by carrying out backward elimination.

Naive Bayes had the highest accuracy overall at 97% with the removal of mitoses followed by random forest with 96.4% (with the removal of one and/or two variables). The least accurate model is the decision tree at (94.9%) which produced better accuracy with only two variables.

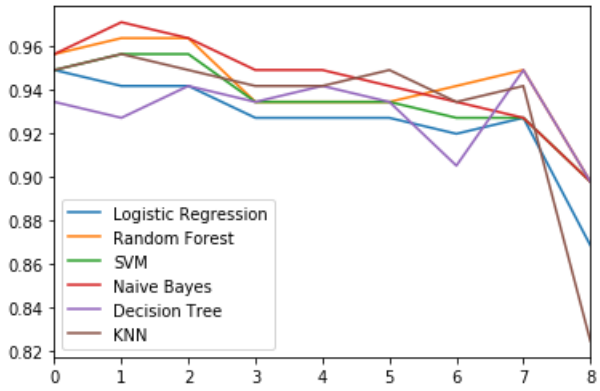


Figure 25: Shows the comparison of accuracy in each model.

Looking at both the recall and accuracy scores, it can be noted that in both cases, naive bayes performed the best followed by random forest and support vector machine. The least performing was the decision tree.

CONCLUSION

The goal of the project is to determine the best performing model with backward elimination to determine relevant variables in the diagnosis of benign and malignant cancer. Based on the results attained, Naive Bayes produced the least false negatives as determined by the recall score (96.6%) and the highest accurate score diagnosis at 97.1% with the removal of a single variable.

By applying machine learning models, we are able to alleviate pressures of breast cancer diagnosis by using the economical, time efficient and non-invasive FNA biopsy. With more data in hand and further training and testing, a better model can be determined to further strengthen the diagnosis prediction.

REFERENCES

"Breast Cancer Statistics." *The Canadian Cancer Society.*, 2019, www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on.

"Cancer and COVID19 (Novel Coronavirus)." *The Canadian Cancer Society*, 11 Mar. 2020, www.cancer.ca/en/support-and-services/support-services/cancer-and-covid19/?region=on.

"Cancer Facts & Figures 2020." *The American Cancer Society*, 2020, www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.

Casaubon, J. T., and J. P. Regan. "Fine Needle Aspiration Of Breast Masses." *StatPearls [Internet]*, 2020, pp. 1–7. *StatPearls Publishing LLC.*, https://www.ncbi.nlm.nih.gov/books/NBK470268/#_NBK470268_pubdet_.

"Core Needle Biopsy of the Breast." *The American Cancer Society*, 2017, www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/core-needle-biopsy-of-the-breast.html/.

"Fine Needle Aspiration (FNA) Biopsy of the Breast." *The American Cancer Society*, 2017, www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html.

Garbar, C, and H Curé. "Fine-Needle Aspiration Cytology Can Play a Role in Neoadjuvant Chemotherapy in Operable Breast Cancer." *ISRN Oncology*, 2013, 10 July 2013, pp. 1–5. *Hindawi Publishing Corporation*, doi:10.1155/2013/935796.

Grant, K. "Ontario Hospitals Warn COVID-19 Trumps Cancer Care in Event of Outbreak." *The Globe and Mail*, 17 Mar. 2020, www.theglobeandmail.com/canada/article-ontario-hospitals-warn-covid-19-trumps-cancer-care-in-event-of/.

Karâa, W., and N. Dey. *Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes*. IGI Global, 2015, pp. 214.

Kourou, K., et al. "Machine Learning Applications in Cancer Prognosis and Prediction." *Computational and Structural Biotechnology*, vol. 13, 2015, pp. 8–17. *Science Direct*, https://doi.org/10.1016/j.csbj.2014.11.005.

Ming, C., et al. “Machine Learning Techniques for Personalized Breast Cancer Risk Prediction: Comparison with the BCRAT and BOADICEA Models.” *Breast Cancer Research*, vol. 21, no. 75, 2019, pp. 1–11. *BMC - Springer Nature*, https://doi.org/10.1186/s13058-019-1158-4.

Sharma, S., et al. “Breast Cancer Detection Using Machine Learning Algorithms.” *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 114–118. *IEEE Xplore*, doi:10.1109/CTEMS.2018.8769187.

University of Wisconsin Hospitals. “Breast Cancer Wisconsin (Original) Data Set.” *UC Irvine Machine Learning Repository*, 1992, archive.ics.uci.edu/ml/datasets/Breast Cancer Wisconsin (Original).