# Literature Review, Data Description, and Project Approach

**By:**

Rama Barham

500903499

Toronto Metropolitan University

10 June 2024

**Submitted to:**

Dr. Ceni Babaoglu

In partial fulfillment for the requirements of:

CIND820: Big Data Analytics Project

## I. PROBLEM:

According to the American Cancer Society, breast cancer is the most commonly diagnosed cancer among women, with millions of new cases reported each year. Breast cancer affects about one in eight women, making it one of the most prevalent and life-threatening cancers impacting females worldwide [1]. Despite the advanced progress in early diagnosis, screening, and patient management, it is still the leading cause of cancer-related deaths among women. Recently, however, new studies have shown that Machine Learning (ML) approaches can aid in detecting breast cancer at an early stage and with high accuracy, allowing for increased survival rate because better treatment can be provided.

Therefore, the objective of this study is to analyze the Breast Cancer Wisconsin (Diagnostic) dataset (WDBC) by applying various machine learning (ML) techniques to predict the presence of breast cancer, classifying tumor types, and identifying key features that influence diagnosis. The chosen theme is Predictive Analytics, with a focus on Classification.

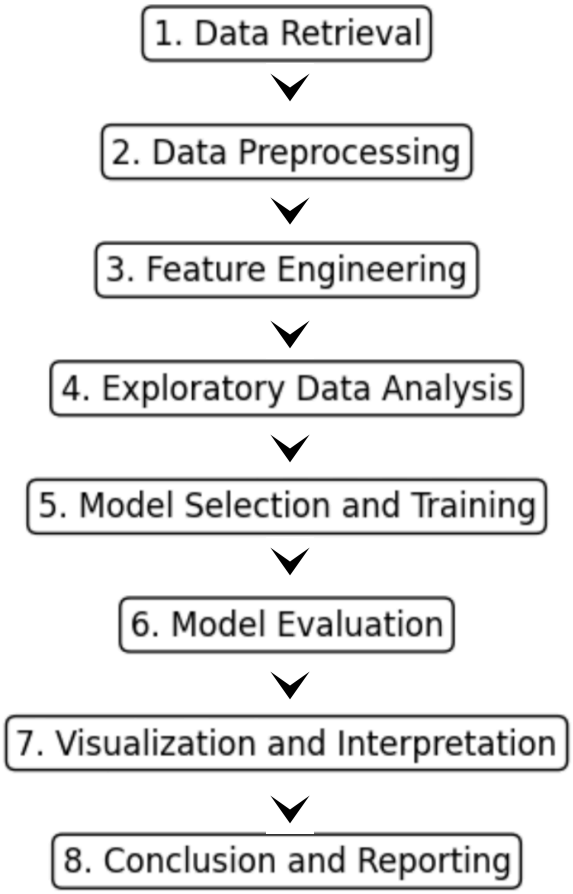
## II. RESEARCH QUESTIONS & IMPLICATIONS:

The primary problem addressed in this project is how to accurately classify breast tumors as benign or malignant using machine learning (ML) techniques. The project seeks to answer the following research questions:

- What features are most indicative of whether a tumor is benign or malignant?
- How accurately can tumors be classified based on their computed features?
- To what extent can Fine Needle Aspiration (FNA) results, when used as a stand-alone diagnostic test, accurately determine the classification of breast tumors as benign or malignant?
- What are the key differences in value features between benign and malignant tumors?
- Are there value features that provide more insights to breast cancer detection over others?
- How can visualization techniques help in understanding the distribution and significance of features?

## III. METHODOLOGY:

To achieve the objective of accurately classifying breast tumors as benign or malignant using machine learning techniques, the following methodology will be employed:




**1. Data Retrieval**

- Retrieve or fetch the data from the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository.

**2. Data Preprocessing**

- Handle missing values: Identify and fill or remove missing data to ensure a complete dataset.
- Normalize/scale data: Normalize the data to ensure all features contribute equally to the analysis.

**3. Feature Engineering**

- Feature Selection:
    - Use univariate feature selection to remove irrelevant features.
    - Apply Recursive Feature Elimination (RFE) to identify significant features for classification.
    - Use L1 regularization (Lasso) to refine feature selection.
    - Evaluate the selected features using cross-validation to ensure they generalize well.

**4. Exploratory Data Analysis (EDA)**

- Perform exploratory analysis to understand the data distribution and feature relationships.
- Use visualization techniques such as scatter plots, heatmaps, and box plots to identify patterns and anomalies.

**5. Model Selection and Training**

- Split the data into training and testing sets.
- Train multiple machine learning models such as Logistic Regression and Decision Tree.
- Use grid search or random search for hyperparameter tuning to optimize model performance.

**6. Model Evaluation**

- Evaluate models using metrics such as confusion matrix, accuracy, precision, recall, and F1-score.
- Compare model performance to identify the best model for classification.

**7. Visualization and Interpretation**

- Use visualization tools to present the distribution and significance of features.
- Create visual aids such as ROC curves and feature importance plots to interpret the model results.

**8. Conclusion and Reporting**

- Summarize findings and implications of the study.
- Prepare a detailed report documenting the methodology, results, and conclusions.

## IV. DATA DESCRIPTION:

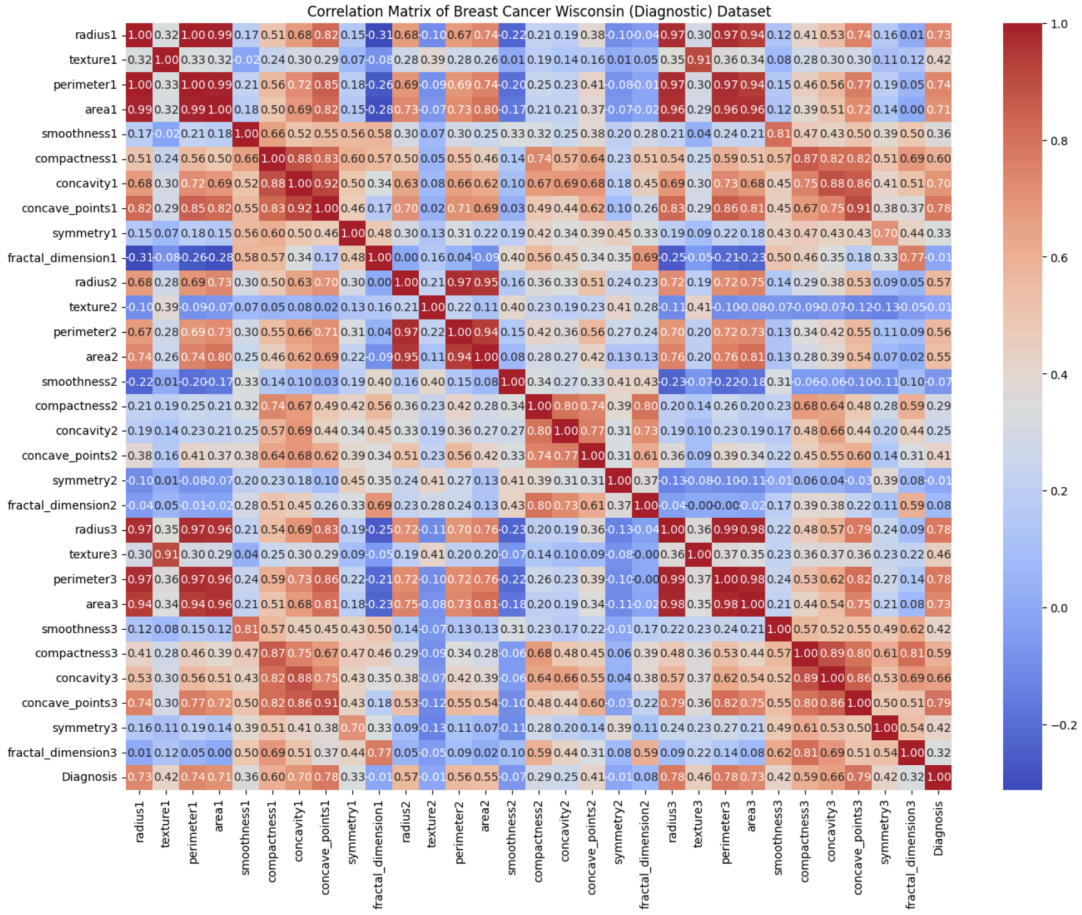
In order to study the problem, this project will utilize the Breast Cancer Wisconsin (Diagnostic) dataset. This dataset includes 569 instances and 32 real-valued attributes. The real-valued attributes were computed from Fine Needle Aspiration (FNA) of breast masses, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Among the 569 patients tested, 37% were diagnosed with malignant breast cancer and 63% were diagnosed with benign breast cancer. There are no missing values in the dataset, ensuring completeness. The attributes are shown in the following table:

| ATTRIBUTES | | | |
|---|---|---|---|
| radius1 | concavity2 | symmetry1 | smoothness3 |
| texture1 | concave_points2 | fractal_dimension1 | compactness3 |

| | | | |
|---|---|---|---|
| perimeter1 | symmetry2 | radius2 | concavity3 |
| area1 | fractal_dimension2 | texture2 | concave_points3 |
| smoothness1 | radius3 | perimeter2 | symmetry3 |
| compactness1 | texture3 | area2 | fractal_dimension3 |
| concavity1 | perimeter3 | smoothness2 | Diagnosis |
| concave_points1 | area3 | | |

**Highly Correlated Groups:**

Features such as Radius, Perimeter, and Area across different measurements (1, 2, 3) are highly correlated with each other, suggesting redundancy as these features provide similar information. Additionally, features related to the shape of the cell nuclei, such as Concavity, Concave Points, and Symmetry, show moderate to high correlations among themselves. Attributes like Radius1, Perimeter1, and Area1, which have strong correlations with the target variable (Diagnosis), are particularly important. In the further stages of this study, I will use the LASSO method to filter out unnecessary features and retain the most useful ones.



**Outliers:** Abnormal data points (outliers) have been identified using the IQR method. Whether the outliers will be treated or not, depends on their impact on model performance.

**The github repository can be found **here**, and the working file can be found **here**.*

## V. LITERATURE REVIEW:

Numerous studies have used the Wisconsin Breast Cancer Database (WDBC) dataset, and similar datasets to study pattern classification of malignant or benign tumors in patients using ML techniques. These studies have taken different approaches to the given problem and achieved high accuracy rates. Summaries of previous research works are detailed as follows:

In Yedjou et al., [6] the study aimed to investigate the use of ML techniques to classify breast cancer by analyzing feature values derived from digitized images of fine-needle aspiration (FNA) samples of breast masses taken from the WDBC dataset.This study focused on analyzing the different real-value features including, the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, and their effectiveness in diagnosing breast cancer. The study found that the features such as radius, texture, perimeter, area, compactness, concavity, and concave points of the cell were strong in distinguishing between malignant and benign tumors. However, other features like smoothness, symmetry, and fractal dimension of the cell image were not strong features in differentiating the type of tumor.

Wanebo et al., [5] also focused on FNA cytology, with the aim of evaluating its efficacy and accuracy as a diagnostic tool for breast cancer. The study compared FNA to the traditional open biopsy method. The primary goal was to determine if FNA could reliably replace open biopsy in the diagnosis and management of primary breast cancer. The sample size included 398 patients who underwent FNA of the breast and the study found FNA correctly diagnosed 100 out of 136 cancers, with no false positives. FNA was found to be safe, rapid, and cost-effective, reducing the need for more invasive procedures like open biopsies.

Other studies have compared several ML algorithms to determine which provide high classification accuracies. For example, in Chaurasia et al., [1] the study looked at how to accurately predict breast cancer survivability using various data mining techniques. To accomplish this, they developed three prediction models: Naive Bayes, Radial Basis Function (RBF) Network, and J48 Decision Tree to determine breast cancer survivability on two parameters: benign and malignant cancer patients, using the WDBC dataset. The models relied on features such as clump thickness, uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal

nucleoli, and mitosis. These features were critical in distinguishing between benign and malignant tumors. The methodologies and results from Chaurasia et al.'s study provides valuable insights into effective algorithms and feature selection that can enhance the accuracy of the predictive models used within my project.

Similarly, this study [3] also explores the application of ML techniques to predict breast cancer. This study uses a dataset of 500 patients from Dhaka Medical College Hospital and evaluates five supervised ML methods: decision tree, random forest, logistic regression, naive bayes, and XGBoost. The goal is to compare the performance criterion of the outlined supervised learning classifiers and determine which method classifies breast tumors as benign or malignant and understand the impact of various features on the model's predictions using SHAP analysis. The study found that XGBoost achieved the highest model accuracy (97%) and performed well across other metrics such as precision, recall, and F1 score. SHAP analysis was used to interpret the model's predictions, revealing that features like mean_perimeter and mean_radius significantly impact the model's output, providing valuable insights to understanding likelihood of predicting early-stage breast cancer.

Kadhim & Kamil's (2022) [4] study aimed to evaluate the performance of various ML classifiers in predicting and diagnosing breast cancer using the WDBC dataset. The study sought to identify the most effective ML algorithms for breast cancer classification based on specificity, sensitivity, precision, accuracy, and F1 score. Eleven ML classifiers were evaluated, including Decision Tree (DT), Extra Randomized Trees (ERT), Naive Bayes (NB), and k-Nearest Neighbors (KNN). The study found that the ERT classifier outperformed other classifiers in terms of accuracy and F1 score, making it the most effective model for breast cancer classification on the WDBC dataset. The results indicate that ensemble methods like ERT, which combine multiple models to improve prediction accuracy, are particularly effective for this type of classification task.

Hasan and Tahir [2] developed a feature extraction algorithm using Principal Component Analysis (PCA) and Artificial Neural Network (ANNs) as classifiers to improve the differentiation between benign and malignant tumors, using the WDBC Database. They used cumulative variance, scree test, and Kaiser Guttman rule as feature selection. The results showed that the method can distinguish between benign and malignant cases strongly, showing that PCA-ANN is an effective way in pattern classification. This study emphasizes the importance of dimensionality reduction techniques like PCA in enhancing the performance of ML models by removing noise and redundant features which will be considered in my feature selection process.

## VI. REFERENCES:


**[1]** Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119-126.

**[2]** Hasan, H., & Tahir, N. M. (2010, May). Feature selection of breast cancer based on principal component analysis. In *2010 6th International Colloquium on Signal Processing & its Applications* (pp. 1-4). IEEE.

**[3]** Islam, T., Sheakh, M. A., Tahosin, M. S., Hena, M. H., Akash, S., Bin Jardan, Y. A., ... & Bourhia, M. (2024). Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Scientific Reports*, *14*(1), 8487.

**[4]** Kadhim, R. R., & Kamil, M. Y. (2022). Comparison of breast cancer classification models on Wisconsin dataset. *Int J Reconfigurable & Embedded Syst ISSN*, *2089*(4864), 4864.

**[5]** Wanebo, H. J., Feldman, P. S., Wilhelm, M. C., Covell, J. L., & Binns, R. L. (1984). Fine needle aspiration cytology in lieu of open biopsy in management of primary breast cancer. *Annals of surgery*, *199*(5), 569-579.

**[6]** Yedjou, C. G., Tchounwou, S. S., Aló, R. A., Elhag, R., Mochona, B., & Latinwo, L. (2021). Application of machine learning algorithms in breast cancer diagnosis and classification. *International journal of science academic research*, *2*(1), 3081.