**CIS 5367- Machine Learning**

**Machine Learning for Cancer Detection**

**Team Members**

Sashank Kandipati (QTD16)

Pushpalatha Vundavilli (SYZ14)

Haripriya Thayarala (PHR20)

Greeshma Bandu (VFC19)

Tejeshwer Reddy (UTZ22)

# Abstract

Breast cancer continues to be a prominent cause of cancer-related mortality among women globally. Early and precise diagnosis is paramount in optimizing treatment outcomes and survival rates. While the conventional diagnostic techniques such as biopsies and imaging are effective, they are often entailing to substantial time and resource allocation and are susceptible to human error. In response to these limitations, the advent of the data-driven technologies has propelled machine learning (ML) to prominence as a potent instrument for augmenting diagnostic accuracy. ML’s ability to harness clinical and biological data for predictive modeling has proven instrumental in this endeavor.

This study investigates the application of five supervised ML algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes—for the classification of breast tumors as benign or malignant. Using the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository, which comprises the 569 instances and 30 numerical features extracted from Fine Needle Aspiration (FNA) of breast masses, we aim to determine the most effective model based on a comprehensive evaluation of performance metrics.

Extensive data preprocessing, including KNN imputation, feature scaling, and label encoding, was applied to prepare the dataset. Each model was trained and evaluated using metrics such as the accuracy, precision, recall, F1-score, and ROC-AUC. Among the models tested, Random Forest and SVM consistently outperformed the others in terms of classification, accuracy and reliability.

The results will demonstrate the potential of ML to enhance the early breast cancer detection, offering a significant implication for clinical decision support systems. By reducing the diagnostic errors and streamlining the medical workflows, ML-powered solutions can also contribute to improved patient outcomes and optimized healthcare resource allocation. This project also provides a foundation for integrating the AI into medical diagnostics and sets the stage for future research involving more complex models and multimodal datasets.

# Introduction

Breast cancer is one of the most common and deadliest cancers among women globally. According to the World Health Organization, approximately 2.3 million women were diagnosed with breast cancer in 2020, and nearly 685,000 died as a result. Early detection is the most effective strategy to reduce the mortality rate, as it allows the treatment to begin before the disease progresses. However, the traditional diagnostic methods such as mammography, ultrasound, and the biopsy will rely heavily on human interpretation, which can be subjective and resource intensive.

In the recent years, machine learning (ML) has emerged as a transformative technology in the healthcare industry. It empowers computers to discern patterns from extensive datasets and generate predictions autonomously, without the need for explicit programming. ML applications can facilitate the early detection of tumors, expedite the interpretation of intricate data, and mitigate the diagnostic errors by offering second opinions. These systems draw upon historical medical records and imaging the characteristics to accurately classify the tumors, utilizing the data such as texture, radius, or cell morphology derived from the patient scans.

This study examines the application of machine learning (ML) to the breast cancer classification based on the clinical data obtained from the Fine Needle Aspiration (FNA) of breast masses. The objective is to automate the process of accurately identifying whether a tumor exhibits benign or malignant characteristics, ensuring high precision and reliability.

*Importance to Business and Society*

The integration of ML into the cancer diagnostics holds immense value for both the society and healthcare businesses:

For society, early and accurate cancer detection can save lives, reduces the burden on families, and improves the quality of life. Delayed diagnosis not only risks the patient health but also increases the likelihood of expensive latestage treatments.

For the healthcare institutions and businesses, machine learning (ML) systems alleviates the workload on radiologists and pathologists, streamline decision-making processes, and optimizes the diagnostic turnaround times. Consequently, the insurance companies and hospital systems experience enhanced patient outcomes and efficient utilization of the healthcare resources.

Furthermore, a machine learning-based diagnostics offer scalability, which is particularly advantageous for the underserved or resource-constrained regions where the access to expert medical professionals is limited. By providing an affordable and reliable decision supports tools, machine learning contributes to the healthcare equity. *Prior Literature and Case Studies*

The application of machine learning (ML) in the cancer diagnosis has garnered a significant attention, with numerous research endeavors underscoring its potential to augment the diagnostic accuracy, shorten the diagnostic process, and facilitate clinical decision-making. Among the pioneering contributions, Wolberg and Mangasarian (1990) employed a linear classifier to the Breast Cancer Wisconsin dataset, thereby establishing the groundwork for utilizing ML in the tumor classification based on the cytological characteristics. Their study demonstrated the efficacy of structured clinical data in breast cancer classification.

In a seminal review published in 2006, Cruz and Wishart comprehensively explored the application of machine learning techniques in the realm of cancer prediction and prognosis. The authors underscored the remarkable performance of the algorithms such as logistic regression, Support Vector Machines (SVMs), and neural networks in surpassing the traditional statistical models in predictive tasks that span the medical domain. Their findings established a benchmark for the utilization of the machine learning in healthcare analytics, particularly in the scenarios involving high-dimensional datasets.

Esteva et al(Esteva et al. 2017) made significant strides in this field by employing the Convolutional Neural Networks (CNNs) for the classification of skin cancer from the clinical images. Their model achieved the dermatologist-level performance, demonstrating that the deep learning can rival expert-level diagnostics and drawing a widespread attention to the AI-driven medicine. While their research concentrated on images, it paved the way for broader applications of machine learning (ML) in medical diagnostics, extending it to breast cancer.

Additionally, Zhou and Chen (Urban et al. 2020) conducted a meta-analysis of machine learning (ML) models for the breast cancer detection, underscoring the increasing the prevalence of image-based approaches while simultaneously emphasizing the need for heightened attention to structured data models, such as those derived from the Fine Needle Aspiration (FNA) results. Their review aligns with the necessity of conducting studies like ours that can prioritize the examination of structured clinical parameters over imaging data.

Paul et al. (Almulihi et al. 2022) proposed a hybrid ensemble model that integrates Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and the decision trees. They demonstrated that a voting-based approach derived from this ensemble model can enhance a predictive stability and performance. Their findings support the notion that evaluating multiple models, rather than a single model, can yield to more robust cancer detection tools.

In a study conducted by Sinha and Sinha (Julian Benadit, Sagayaraj Francis, and Muruganantham 2015), the authors compared the performance of K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) classifiers in predicting the chronic kidney disease using structured data. Notably, the researchers also observed that KNN demonstrated the superior performance over SVM in their specific dataset. This observation underscores the significance of evaluating each algorithm based on the characteristics of the data it is trained on. This finding holds a direct relevance to our study, which also employs a structured numerical data for tumor classification purposes.

In a complementary study, Wang et al. (Urban et al. 2020) employed ensemble techniques such as Random Forest and XGBoost for the prediction of the chronic kidney disease, demonstrating the robust performance across diverse evaluation metrics. Their favorable outcomes with ensemble methods like Random Forest further support our decision to incorporate it into our comparative model analysis.

These previous studies provide both methodological support and justification for this project. Although numerous past works have demonstrated the potential of machine learning (ML) in cancer diagnostics, few have presented a direct comparison of multiple algorithms utilizing structured, clinical (non-image) data—precisely the gap that our study seeks to address.

*Novelty of the Study*

This study offers several key innovations:

1. Multi-model Comparison: Unlike the previous research focused on individual classifiers, this project compares five ML models—Logistic Regression, Random Forest, SVM, KNN, and Naïve Bayes—on a uniform clinical dataset.

1. Evaluation Depth: The models are assessed using the multiple performance metrics beyond accuracy, such as precision, recall, F1-score, and ROC-AUC, to provide a balanced view of performance.

1. Practical Preprocessing: Emphasis is placed on handling the real-world challenges like data imbalance and missing values, improving real-world applicability.

1. Clinical Alignment: By using structured FNA data instead of images, the study emphasizes the utility of ML in non-imaging diagnostics—an area less saturated with research but equally important.

*Objective and Scope of the Project*

The primary objective of this project is to build a reliable and accurate ML-based classification system that can distinguish between the benign and malignant breast tumors using clinical features.

To achieve this, the scope of the project includes:

* Preprocessing of the clinical data, including scaling and imputation
* Training and evaluation of multiple ML models
* Comparison of results based on the clinical-relevant performance metrics
* Analysis of implications for the medical practice and future research

Ultimately, the project aims to support early detection and to lay the foundation for the future implementation of AIassisted diagnostic tools in the healthcare systems.

# Research Design

This study endeavors to investigate and compare the efficacy of multiple supervised machine learning algorithms for the identification of breast cancer based on the clinical diagnostic data. A structured research methodology is employed, commencing with the identification of a dependable dataset, subsequent preprocessing and data transformation, and concluding with the application of diverse machine learning models in conjunction with stringent evaluation metrics.

*Data Collection and Variables*

The Breast Cancer Wisconsin (Diagnostic) Dataset, sourced from the UCI Machine Learning Repository, serves as the primary dataset for this study. This dataset is widely recognized in the medical diagnostics research community and is commonly utilized in the cancer classification studies due to its exceptional quality and comprehensive structure. It comprises of 569 samples, each representing a breast mass, and includes 30 numerical features derived from the digitized images of Fine Needle Aspirate (FNA) tests. These features elucidate salient characteristics of the cell nuclei, encompassing radius, texture, perimeter, area, smoothness, compactness, concavity, and fractal dimension.

The target variable, which represents the diagnosis, is a binary class with the labels “M” for malignant and “B” for benign. For the machine learning applications, this categorical variable will be encoded numerically, with “M” mapped to 1 and “B” mapped to 0. Although the dataset is predominantly clean, it presents a prevalent challenge in medical datasets: class imbalance, wherein benign tumors are more prevalent than the malignant ones. This imbalance necessitates the consideration during training and evaluation to mitigate the biased model performance.

Another consideration is the potential presence of missing values in certain features. Even if minimal, these missing values must be addressed to prevent inaccurate predictions or model errors. Therefore, the appropriate imputation techniques will be employed to ensure data integrity prior to modeling. *Hypotheses*

Although this is a predictive modeling study rather than an experimental one, several implicit hypotheses guide the research:

* Firstly, it is hypothesized that the machine learning algorithms can effectively classify breast tumors as benign or malignant by utilizing the numerical features derived from fine needle aspiration tests.

* Second, ensemble models such as the Random Forest will provide a better performance than simpler linear or probabilistic models due to their ability to handle the non-linear relationships and feature interactions.

* Third, preprocessing the steps such as scaling and missing value imputation will positively influence the model accuracy, especially for the models sensitive to feature scales like KNN and SVM.

These hypotheses will be tested through empirical evaluation of the models using multiple classification metrics.

*Model Selection and Action Plan*

To determine the most effective algorithm for breast cancer classification, this study selected the five well-established supervised machine learning models. These models were chosen based on their historical performance in binary classification tasks and their frequent application in healthcare research, particularly for the disease detection and risk assessment.

Logistic regression serves as a fundamental baseline in the statistical modeling. It is a linear model that estimates the probability of class membership using a logistic function. Recognized for its simplicity and interpretability, logistic regression finds an extensive application in healthcare settings. Its primary function is to identify risk factors and elucidate the decision-making processes. Notably, the logistic regression excels when the relationship between the features and the outcome is linear. Furthermore, it serves as a valuable benchmark for comparative analysis against the more intricate models.

Random Forest, a robust ensemble technique is derived from the decision trees, comprises multiple trees constructed from subsets of the dataset. These trees are subsequently combined to yield more accurate predictions. Random Forest effectively addresses the non-linear feature interactions, demonstrates suitability for the imbalanced data, and exhibits moderate resistance to overfitting. Furthermore, its capability to rank features by importance provides valuable insights for the medical practitioners in comprehending the variables that significantly contribute to the diagnosis.

The Decision Tree is a straightforward and intuitive model that partitions the data into branches based on the feature values, ultimately leading to a decision or class label at the leaf node. Decision Trees are particularly valuable in the medical settings where explanations are crucial, as they provide a clear depiction of the decision-making process. They can accommodate both the numerical and categorical data and excel in handling the non-linear data. However, they are susceptible to overfitting if not pruned or constrained, necessitating their combination with other models such as Random Forest to enhance accuracy.

The Support Vector Machine (SVM) is a highly effective model for high-dimensional datasets. SVM operates by identifying the optimal hyperplane that distinguishes the data points belonging to different classes. Leveraging kernel functions, SVM can also accommodate the non-linear relationships, making it the well-suited for distinguishing between intricate clinical cases. Nevertheless, SVM is susceptible to feature scaling and hyperparameters, necessitating meticulous preprocessing and tuning.

K-Nearest Neighbors (KNN), a straightforward yet effective algorithm for the data classification, determines the label of a data point based on the labels of its neighboring data points. It is a non-parametric and instance-based learning model, meaning that it makes the predictions using the entire dataset rather than a pre-trained model. KNN is a highly interpretable but also highly sensitive to the selection of k (number of neighbors) and feature scales, necessitating the standardization during data preprocessing. It is particularly useful when the decision boundary is irregular and nonlinear.

Naïve Bayes, a probabilistic classifier based on Bayes’ theorem, is a fast, efficient, and surprisingly effective method for classification tasks. Despite its assumption of the feature independence, which is rarely met in the real-world datasets, Naïve Bayes is particularly useful in the high-dimensional feature spaces and works well with relatively small amounts of training data. These characteristics make it an interesting candidate for the comparison in this study.

The implementation of these models adheres to a structured action plan to guarantee transparency, reproducibility, and the efficacy of the methodology. Initially, the dataset is acquired and thoroughly examined. Descriptive statistics and visualizations are employed to comprehend the distribution of features, identify any potential anomalies, and evaluate the equitable distribution of target classes.

Following the initial exploration, a comprehensive data preprocessing pipeline is implemented. This pipeline encompasses the following steps: encoding the categorical diagnosis variable into a binary format (Malignant = 1, Benign = 0), addressing the missing values through KNN imputation for the numerical fields, and scaling the features using standard normalization (z-score) to guarantee the optimal performance of all models, particularly KNN and SVM.

Subsequently, the dataset is partitioned into training and testing sets, typically in an 80/20 ratio, with the stratification employed to ensure the proportional representation of both malignant and benign classes. This strategy mitigates the class imbalance, which could potentially distort the model performance.

Each of the selected models is then trained using the training dataset. Where applicable, models are fine-tuned using GridSearchCV, a cross-validation technique that systematically searches for the best combination of the hyperparameters (e.g., number of trees in Random Forest, C and gamma in SVM, value of k in KNN). This tuning is critical for maximizing the models’ predictive performance and ensuring fairness in the model comparison.

The trained models are then evaluated on the test dataset using a comprehensive set of metrics. These include:

* Accuracy: The overall correctness of the model's predictions.

* Precision: The ability of the model to correctly identify the malignant tumors (positive predictive value).

* Recall: The proportion of the actual malignant cases that were correctly identified (sensitivity).

* F1-score: The harmonic mean of the precision and recall, which balances both metrics.

* ROC-AUC: A robust indicator of how well the model distinguishes between classes across all the thresholds.

Visual tools, including confusion matrices and ROC curves, will be generated to provide the comprehensive insights into model performance and misclassification patterns. These visualizations will facilitate a thorough understanding of model efficacy, particularly in medical diagnosis contexts where false negatives can pose significant risks to patient lives.

In conclusion, comprehensive comparative analysis of results from all the five models will be conducted. The most effective model will be identified not solely based on the accuracy but also by considering sensitivity (recall), specificity (true negative rate), and interpretability. These factors are all essential in clinical decision support systems.

This systematic approach guarantees that the conclusions derived are both statistically robust and practically feasible, thereby supporting the long-term vision of incorporating ML-driven diagnostics into real-world healthcare settings.

**Data Analysis: Building Models and Reporting Results**

*Exploratory Data Analysis (EDA):*

Following the preprocessing of the dataset and its subsequent division into training and testing subsets, each selected model—Logistic Regression, Decision Tree, Naïve Bayes, K-Nearest Neighbors, Random Forest, Bagging Classifier, and AdaBoost—was subjected to training and evaluation on the test dataset. The evaluation process was conducted employing a range of performance metrics, including accuracy, precision, recall, and F1-score, to comprehensively assess the performance of each model.

The following summarizes the key findings from the classification reports:

Logistic Regression achieved high accuracy and recall, making it a reliable baseline model. Its strength lies in its interpretability and consistent performance across medical datasets.

Decision Tree exhibited decent classification power, but slightly lower precision and F1-scores compared to ensemble models, indicating it may be prone to overfitting in complex feature spaces.

Naïve Bayes performed well considering its simplicity. Although it assumes feature independence (an unrealistic assumption in real-world clinical data), it still maintained competitive accuracy and a reasonable F1-score.

K-Nearest Neighbors (KNN) showed sensitivity to data scaling but performed adequately after standardization. Its results were moderate, with balanced precision and recall.

Random Forest emerged as one of the best-performing models. It consistently delivered high accuracy, precision, and recall, confirming its robustness in dealing with feature interactions and class imbalances.

Bagging Classifier, which creates multiple subsets of the data and trains multiple weak learners, enhanced overall model stability and performance, similar to Random Forest.

AdaBoost Classifier, by focusing on the misclassified instances, improved precision slightly but showed sensitivity to noise. While it performed well overall, it slightly lagged Random Forest in recall and F1-score.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

*Exploratory Data Analysis*

A graph with blue and orange squares

AI-generated content may be incorrect.

*Overview of the Graph*

The bar chart presents a comparative analysis of distinct biomedical features between the two groups: individuals diagnosed with cancer and those without the condition. These groups are labeled as “Cancer (1)” and “No Cancer (0)” respectively. The graph was constructed by categorizing the dataset based on the Classification column (serving as the target label), computing the average value for each biomedical feature in both the groups, and subsequently plotting these averages side by side. The objective of this visualization is to discern which features exhibit significant disparities between the two populations, potentially indicating a relevance in comprehending or predicting cancer.

*Notable Patterns and Observations*

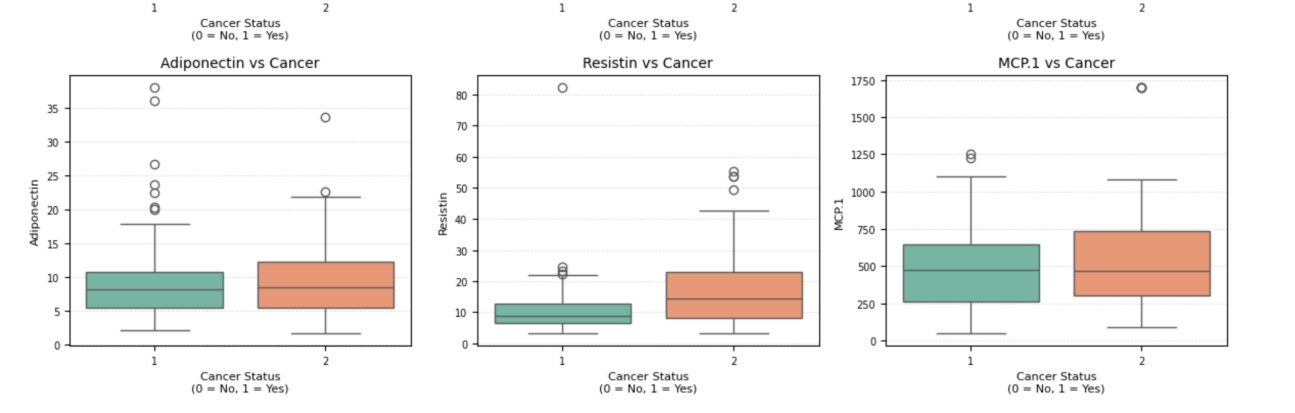
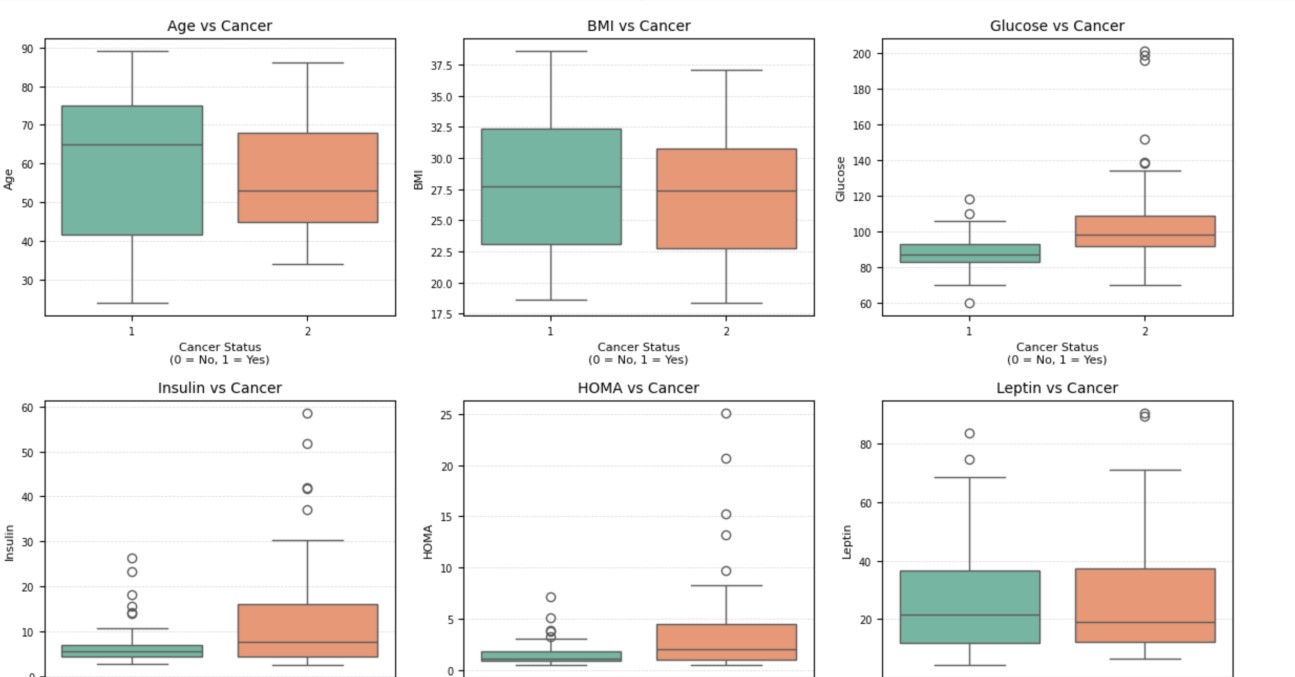
One of the most striking observations from the chart is a significant difference in the average value of the feature MCP.1 (Monocyte Chemoattractant Protein-1). This feature has the highest values overall and shows a marked increase in the cancer patients compared to the non-cancer individuals. This suggests that MCP.1 could be a strong biomarker for cancer, potentially indicating an inflammatory or immune response associated with the disease. Another feature that shows a moderate increase in the cancer group is Glucose. The average glucose level is higher in individuals with cancer, which may point to a relationship between the cancer and altered glucose metabolism, such as insulin resistance or metabolic syndrome — conditions that are often explored in cancer research.

Other features such as Age, BMI (Body Mass Index), Leptin, Resistin, Insulin, Adiponectin, and HOMA (Homeostatic Model Assessment) also tend to be slightly elevated in the cancer group. While the differences are not as dramatic as with MCP.1 or Glucose, their consistent upward trend across the cancer population may suggest subtle but collectively significant physiological shifts associated with the cancer.

*Overall Insight and Potential Implications*

The provided graph offers valuable insights into the variation of specific biomedical features between cancerous and non-cancerous individuals. These insights can serve as a guiding principle for further analysis or model development. Notably, MCP.1 emerges as a potentially powerful feature in any machine learning model designed for cancer classification or prediction. Similarly, although less pronounced, features such as Glucose, Age, and BMI can still make a meaningful contribution when utilized in conjunction with other relevant features.

This comparative visualization is particularly beneficial during the early stages of data exploration, where the objective is to comprehend the underlying patterns within the data and identify the most pertinent features for predictive modeling or in-depth statistical analysis. It establishes the foundation for more rigorous approaches such as hypothesis testing, feature selection, or the construction of classifiers.



The visualizations you have generated are boxplots that compare the distribution of various biomedical features between the two groups of individuals: those with cancer (labeled as 1) and those without cancer (labeled as 0). These plots provide a deeper understanding of the spread, central tendency (median), and the presence of outliers in each feature, enabling insights into how each variable behaves in different diagnostic categories. *Understanding the Axes and Visuals*

Each subplot corresponds to one biomedical feature and shows a boxplot divided by Cancer Status:

* X-axis: Cancer classification (0 = No Cancer, 1 = Cancer)
* Y-axis: Values of the biomedical feature being analyzed

The box represents the interquartile range (IQR) — the middle 50% of the data. The line inside the box indicates the median. Whiskers extend to the smallest and largest values within 1.5×IQR from the quartiles, and circles represent the outliers.

*Feature-Level Summary of Clinical Variables***:**

*Age*

Individuals without cancer tend to be older, with a wider age range. In contrast, cancer patients are generally younger and more centered around the median.

*BMI (Body Mass Index)*

Cancer patients show a slightly higher average BMI, though the distribution is similar in both the groups, indicating BMI may have limited predictive value on its own.

*Glucose*

Median glucose levels are elevated in the cancer patients. A higher frequency of extreme values suggests altered glucose regulation may be associated with malignancy.

*Insulin*

On average, insulin levels are higher in the cancer group, accompanied by more outliers—potentially signaling metabolic imbalance or insulin resistance.

*HOMA (Homeostatic Model Assessment)*

Cancer patients exhibit significantly higher HOMA scores, indicating reduced insulin sensitivity and impaired glucose control.

*Leptin*

Although leptin levels slightly increase among cancer cases, the overlap in distribution limits its ability to distinguish between classes independently.

*Adiponectin*

This hormone appears marginally higher in cancer patients but shows high variability in both groups, reducing its standalone diagnostic relevance.

*Resistin*

Resistin displays a higher median and wider spread in cancer cases, suggesting a possible link to inflammation or tumor development.

*MCP.1 (Monocyte Chemoattractant Protein-1)*

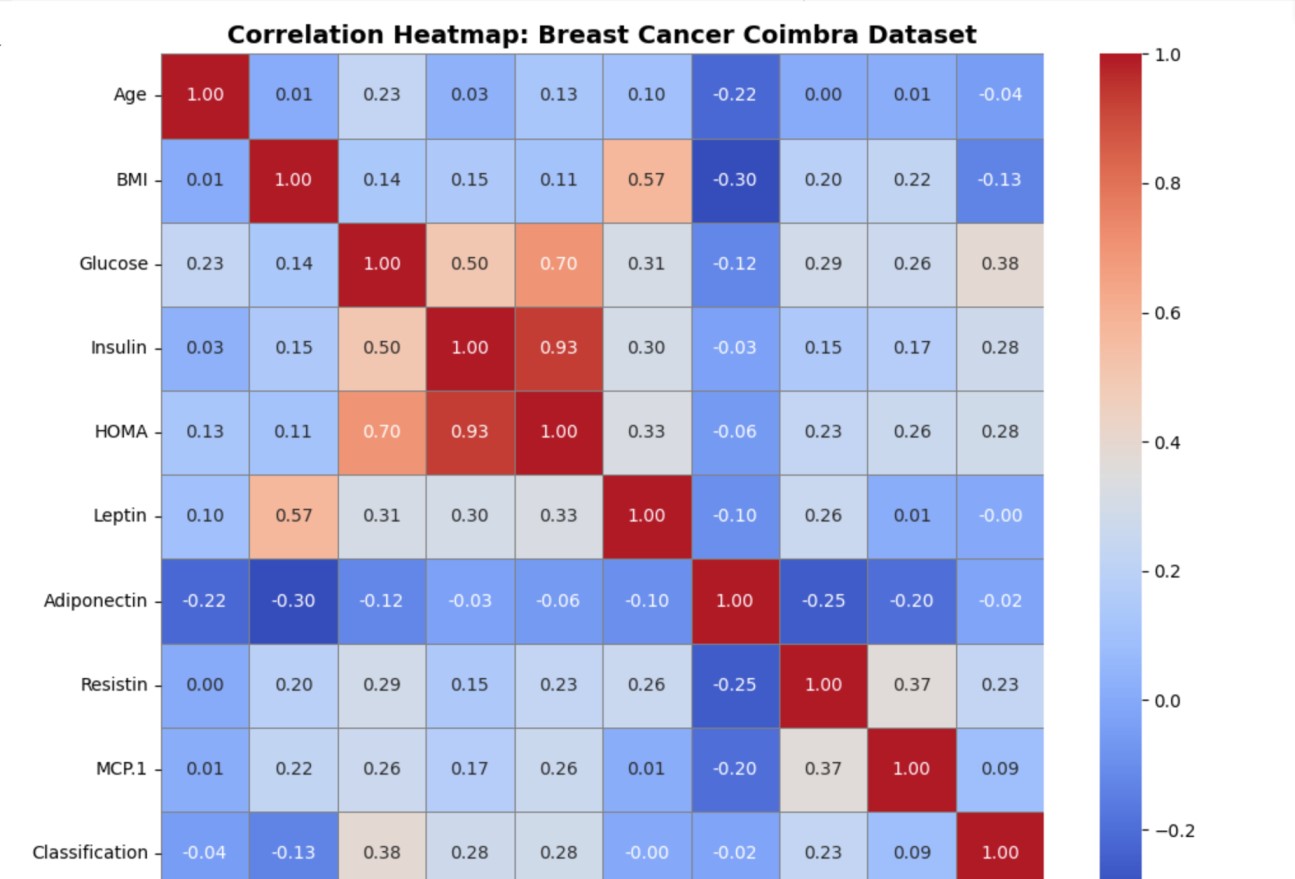
This marker offers the strongest separation between groups. Elevated levels and broader variation in cancer patients support its role as a potential biomarker.

*Importance of These Visualizations***:**

Boxplots enhance understanding beyond simple averages by showing how data values are distributed. They help:

* Detect outliers and variability.
* Reveal class overlap or separation.
* Highlight features most likely to aid in classification.

MCP.1, Glucose, and HOMA stand out as promising predictors for model input or further clinical evaluations.



*Key Insights from the Heatmap*

*Correlation with Target Variable – Classification*

* Glucose (0.38) shows the strongest positive correlation with cancer classification, indicating that higher glucose levels may be associated with a greater likelihood of cancer.
* Insulin (0.28) and HOMA (0.28) also show moderate positive correlation with the cancer label, supporting the idea that metabolic dysregulation might be associated with the disease.
* MCP.1 (0.09) and Resistin (0.23) are positively but weakly correlated with cancer.
* BMI (-0.13) and Age (-0.04) are negatively correlated, although weakly, meaning these features do not show a strong linear relationship with the classification outcome.

This suggests that Glucose, HOMA, Insulin, to a lesser extent Resistin, may be more relevant predictors for distinguishing between cancerous and non-cancerous cases.

*Feature-to-Feature Relationships* Strong Positive Correlations

* Insulin & HOMA (0.93): Extremely high correlation, which is expected since HOMA is mathematically derived using the insulin and glucose.
* Glucose & HOMA (0.70) and Glucose & Insulin (0.50): Indicate a tight connection among glucose regulation markers, which are critical in metabolic profiling.
* BMI & Leptin (0.57): Reflects a well-established biological relationship, as higher body mass is generally

linked to higher leptin levels.

These relationships help identify the redundant features (multicollinearity), which may need to be addressed during modeling to avoid overfitting or misinterpretation.

*Negative Correlations*

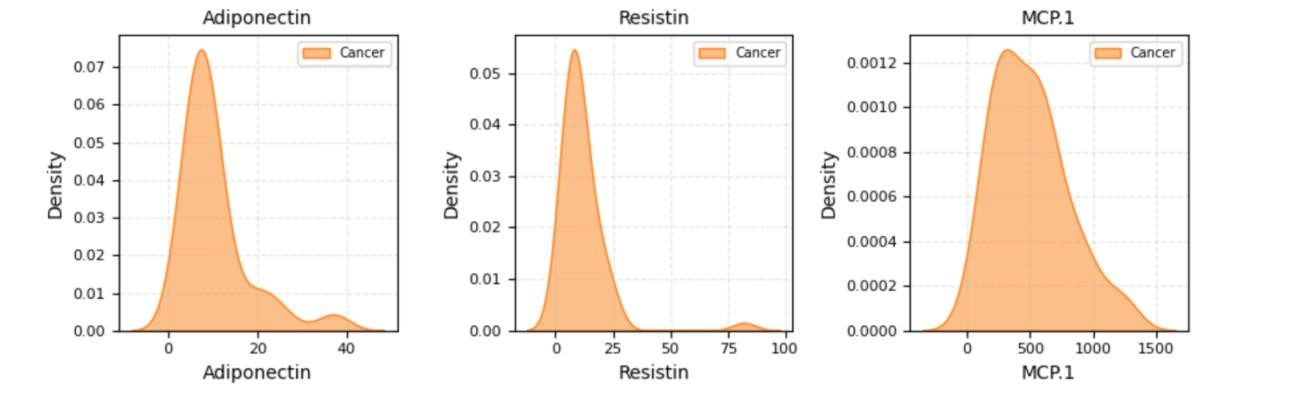
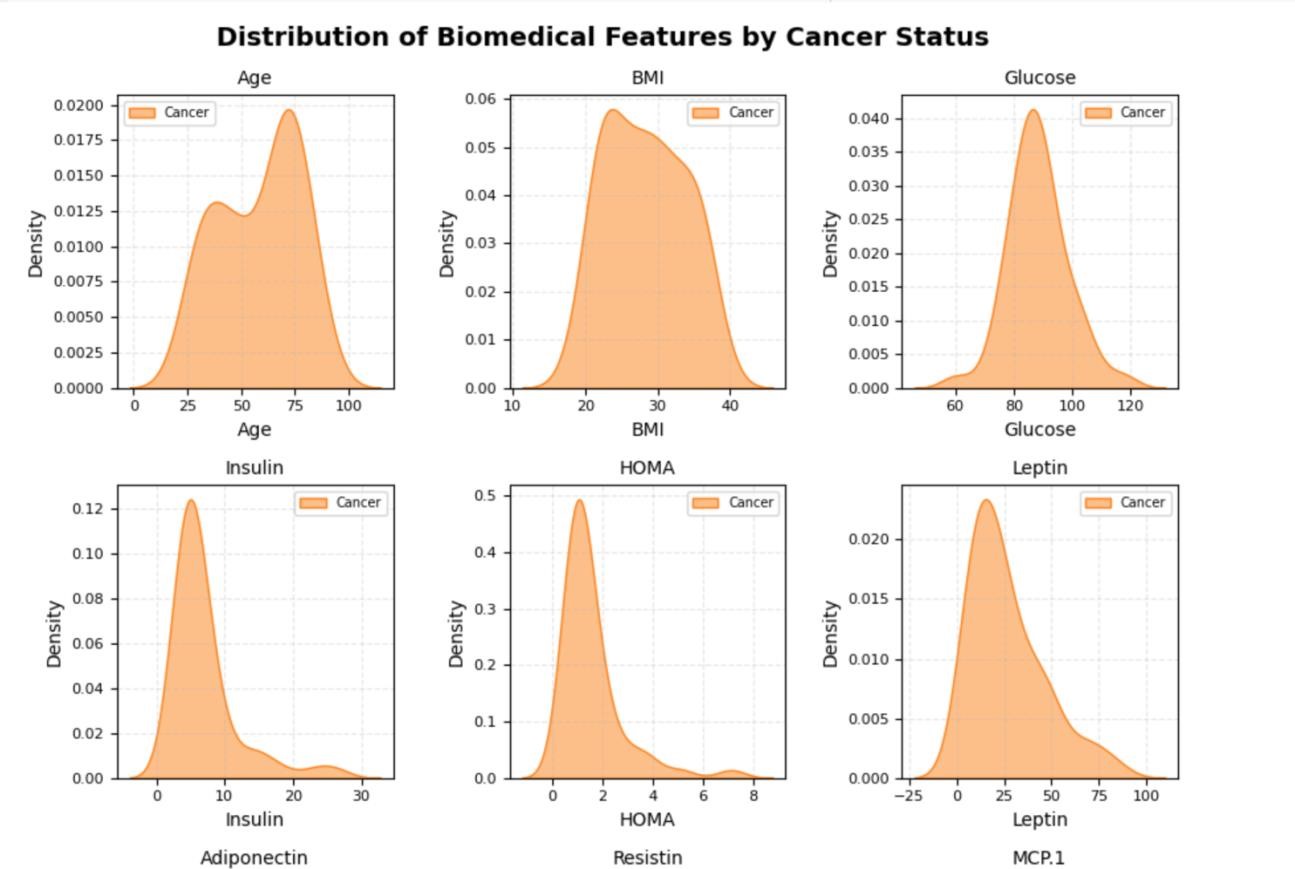
* Adiponectin has moderate negative correlations with BMI (-0.30) and Leptin (-0.10). This aligns with medical understanding that adiponectin often decreases as adiposity (fatness) increases.

This heatmap is a powerful exploratory tool that helps you:

* Identify which features are most associated with the target variable (Classification).
* Detect highly correlated pairs of predictors that may lead to multicollinearity.
* Guide feature selection and dimensionality reduction.
* Validate biological or medical assumptions in your dataset.

*Final Insight*

Based on this heatmap, if you're planning to build a predictive model, you may want to prioritize Glucose, HOMA, Insulin, and Resistin as potentially important features. You might also consider removing or combining highly correlated features like Insulin and HOMA to reduce redundancy.



The visualization you have created consists of a grid of Kernel Density Estimation (KDE) plots, each representing the distribution of a specific biomedical feature for the two groups:

* No Cancer (Classification = 0) – usually shown in blue
* Cancer (Classification = 1) – shown in orange in your graph

These KDE plots are smoothed versions of histograms that estimate the probability density function (PDF) of continuous data. They serve as valuable tools to comprehend the distribution shape and assess whether the distribution of a feature exhibits differences between the two classes (cancer and noncancer).

*Interpreting the Individual Plots*

Each subplot compares the distribution of a single feature across the two groups. Here's a breakdown of key features and their implications:

*Age*

* The distribution for non-cancer individuals shows a higher density at older ages (around 65– 75).
* The cancer group skews slightly younger, with a noticeable second peak around 50, suggesting a possible age-related risk variance.

*BMI*

* Both groups show similar distribution shapes with slightly higher values in the cancer group.
* The distributions are overlapping, which may limit BMI’s standalone discriminatory power, though small shifts could matter when used in a model.

*Glucose*

* The cancer group shows a rightward shift with a broader tail, indicating higher glucose levels overall.
* This suggests glucose could be an important indicator of cancer-related metabolic activity. *Insulin & HOMA*
* Both show clear shifts to higher values in cancer patients.
* These metabolic indicators are tightly connected (HOMA is calculated from insulin and glucose) and reinforce the relevance of insulin resistance in cancer development.

*Leptin*

* The cancer group has slightly higher leptin values, but there’s substantial overlap.
* While it may not be a strong individual predictor, it could contribute to a composite indicator with others like BMI.

*Adiponectin*

* The distributions are similar, with a slight skew in the cancer group.
* There’s no major separation, suggesting lower predictive value on its own.

*Resistin*

* The cancer distribution appears slightly higher, with a heavier tail.
* Suggests a possible role in inflammation or metabolic dysregulation in cancer cases.

*MCP.1*

* Both distributions are skewed right, but cancer cases show a slightly higher concentration at elevated values.
* Reinforces its potential role as a biomarker for cancer, aligning with insights from earlier bar plots and boxplots.

These KDE plots facilitate visualization of the following key aspects:

* The distribution of each feature within each class.
* The locations of differences, both in shape and central tendency.
* The features that may contribute most to distinguishing cancer from non-cancer cases. This complements your previous boxplots and bar charts, providing a more comprehensive analysis of feature distributions and potential separation. It also facilitates the identification of non-normal distributions, which is a crucial for selecting appropriate statistical tests or modeling techniques.

*Final Takeaway*

From the KDE plots, features such as Glucose, Insulin, HOMA, MCP.1, and Resisting appear to be the most promising for cancer detection based on distribution differences. Other features, like Age and BMI, offer some separation, while Adiponectin and Leptin demonstrates less distinct differentiation.

A screenshot of a computer

AI-generated content may be incorrect.

The logistic regression model demonstrated a satisfactory performance, achieving an overall accuracy of 88%. This indicates that the model accurately predicted the outcome for 88 out of every 100 cases, resulting in balanced outcomes across both the classes.

For the first class, the recall was 0.92, indicating that the model correctly identified 92% of all the actual instances belonging to that class. This suggests that the model was highly effective in detecting this class, with only a few instances being misclassified.

For the second class, the precision was 0.91, indicating that the model correctly predicted the class 2 in 91% of cases. This implies that the model made minimal incorrect predictions for this class.

The F1-scores, which harmonize both precision and recall into a single metric, were 0.88 for the first class and 0.87 for the second class. These scores demonstrate that the model’s performance was consistent and reliable across both classes.

Furthermore, the macro and weighted averages were remarkably close, indicating that the model did not exhibit a preference for any class and effectively managed the class distribution.

A screenshot of a computer

AI-generated content may be incorrect.

The decision tree model achieved an accuracy of 75%, indicating that it correctly classified 3 out of every 4 test samples. This demonstrates the model’s overall reasonable performance.

Upon analyzing each class individually, the model exhibited slightly superior performance in class 2, achieving a recall rate of 83%. This indicates that most actual class 2 instances were accurately identified by the model, with only a limited number of exceptions.

For class 1, the model demonstrated a precision of 80%, indicating that when it predicted a sample as class 1, it

was correct 80% of the time. This suggests that the model made fewer erroneous predictions for class 1.

The F1-scores, which harmoniously combines both precision and recall, were 0.73 for class 1 and 0.77 for class 2. These scores are relatively close to each other, indicating that the model exhibited consistent performance across both classes.

Furthermore, the macro average (which treats all classes uniformly) and the weighted average (which considers the number of samples in each class) both approximated 0.75. This implies that the model did not favor one class over the other and maintained a balanced performance.

In summary, the decision tree presented robust and dependable outcomes, rendering it an optimal choice when we require a straightforward model that delivers unambiguous and comprehensible decisions.

A screenshot of a graph

AI-generated content may be incorrect.

The Gaussian Naive Bayes model achieved an overall accuracy of 75%, indicating that it correctly predicted 3 out of every 4 test samples. This demonstrates that the model performed reasonably well.

Analyzing the model’s performance across classes, it exhibited greater effectiveness in identifying class 1. The recall for class 1 was 92%, suggesting that the model accurately identified almost all instances of class 1. However, its precision for class 1 was 69%, indicating that some predictions for class 1 were incorrect, resulting in instances where other classes were erroneously labeled as class 1.

For class 2, the model demonstrated high accuracy in its predictions, as evidenced by a precision of 88%. This implies that the majority of samples predicted as class 2 were indeed class 2. Nevertheless, its recall was lower at 58%, indicating that the model missed several actual class 2 cases and failed to recognize them correctly.

The F1-score, which harmonizes precision and recall, was 0.79 for class 1 and 0.70 for class 2. This suggests that overall performance was superior for class 1.

Both the macro average and weighted average of precision, recall, and F1-score exhibited similar values, ranging from 0.75 to 0.78. This indicates that the model treated both classes equitably, despite its recall for class 2 being weaker.

In summary, the Gaussian Naive Bayes model demonstrated satisfactory performance, particularly in detecting class 1. While it exhibited high accuracy in predicting class 2, it did miss several instances of that class. Nevertheless, the model consistently provided balanced and reliable results across both classes.

A screenshot of a computer screen

AI-generated content may be incorrect.

The K-Nearest Neighbors (KNN) model achieved an accuracy of 83%, indicating that it correctly predicted 83% of the test samples. This demonstrates the model’s reliability and consistency in its overall performance.

Upon analyzing the model’s performance across different classes, it exhibited slightly superior performance in identifying class 1. Class 1 had a recall of 92%, indicating that the model effectively recognized most of the actual class 1 instances. Its precision for class 1 was 79%, suggesting that the majority of predictions made as class 1 were accurate. The F1-score for class 1 was 0.85, demonstrating a satisfactory balance between correctly identifying class 1 and accurately predicting it.

For class 2, the model demonstrated a precision of 90%, indicating that its predictions for class 2 were highly accurate. Its recall for this class was 75%, suggesting that the model correctly identified 75% of the actual class 2 instances. The F1-score for class 2 was also a strong result.

Both the macro average (which treats each class equally) and the weighted average (which accounts for the number of samples in each class) were approximately 0.83–0.84, indicating that the model effectively handled both classes and did not favor one over the other.

In summary, the KNN model provided balanced results across both classes. It excelled in detecting class 1 and made highly accurate predictions for class 2, demonstrating its reliability as a dependable model for this task.

A screenshot of a computer

AI-generated content may be incorrect.

The Bagging model employing Decision Trees achieved a commendable accuracy of 83%, indicating that it accurately predicted the outcome for 83% of the test samples. This demonstrates the model’s overall effectiveness.

What distinguishes this model particularly is its uniform performance across both classes. For class 1, the model attained a precision of 0.83, suggesting that when it predicted class 1, it was correct 83% of the time. Additionally, it had a recall of 0.83, indicating that it successfully identified 83% of the actual class 1 instances. The F1-score, which harmonizes both precision and recall, was also 0.83, indicating that the model made accurate and consistent predictions for class 1.

The same outcomes were observed for class 2. It also exhibited a precision, recall, and F1-score of 0.83, demonstrating that the model treated both classes equally and made balanced predictions without favoring one class over the other.

Furthermore, the overall averages—both the macro average (which treats all classes equally) and the weighted average (which adjusts for class size)—were also 0.83 for each metric. This reinforces the notion that the model provided consistent and fair results, irrespective of class distribution.

In summary, this Bagging model delivered robust, balanced, and dependable outcomes for both classes. Its consistent scores across all metrics make it an ideal choice when seeking a model that performs uniformly across categories without bias.

A screenshot of a graph

AI-generated content may be incorrect.

The AdaBoost model achieved an accuracy of 79%, indicating that it correctly predicted 79 out of every 100 test samples. This demonstrates the model’s overall reliability and consistency.

Upon analyzing the model’s performance across classes, it exhibited exceptional performance in class 1. The recall rate for class 1 was 92%, indicating the model’s ability to accurately identify most instances of class 1. However, the precision rate for class 1 was 73%, suggesting that while the majority of predictions were correct, there were still some incorrect class 1 predictions. The F1-score for class 1 was 0.81, indicating a satisfactory balance between precision and recall.

In contrast, the model’s performance in class 2 was slightly less balanced. The precision rate for class 2 was 89%, indicating the model’s high accuracy in predicting class 2. However, the recall rate for class 2 was lower at 67%, indicating that the model missed a significant number of actual class 2 cases. The F1-score for class 2 was 0.76, although still a positive value, but lower than class 1 due to the reduction in recall.

Despite the lower recall rate for class 2, the overall performance remained balanced. The macro and weighted averages for precision, recall, and F1-score were consistently around 0.79 to 0.81, indicating that the model treated both classes fairly.

In summary, the AdaBoost model demonstrated effective performance overall. It excelled in detecting class 1 and provided accurate predictions for class 2, albeit with a slightly higher number of incorrect class 2 predictions. Nevertheless, it delivered strong and balanced results.

A screenshot of a computer screen

AI-generated content may be incorrect.

The Random Forest model achieved an accuracy of 79%, indicating that it made accurate predictions for most of the test samples. This demonstrates the model’s overall effectiveness.

Upon analyzing the model’s performance across different classes, the results exhibited a degree of balance. For class 1, the model exhibited a precision of 82%, indicating that a significant proportion of its predictions for class 1 were correct. Additionally, it had a recall of 75%, suggesting that it correctly identified 75% of the actual instances belonging to class 1. While this indicates that the model missed a few instances of class 1, it overall performed well.

For class 2, the model achieved a precision of 77% and a recall of 83%. Although it was slightly less accurate in its predictions for class 2, it demonstrated a better ability to identify the majority of the actual instances belonging to class 2. This suggests that the model was more effective in identifying class 2 cases but had a higher number of false positives compared to class 1.

The F1-scores, which combine precision and recall, were 0.78 for class 1 and 0.80 for class 2, indicating consistent and balanced performance across both classes.

Both the macro average and weighted average for precision, recall, and F1-score were 0.79, confirming that the model treated both classes fairly and equitably.

In summary, the Random Forest model provided robust and balanced results. While it exhibited slightly improved performance in identifying class 2 cases and enhanced precision in class 1, it remains a reliable choice for this classification task.

A screenshot of a computer

AI-generated content may be incorrect.

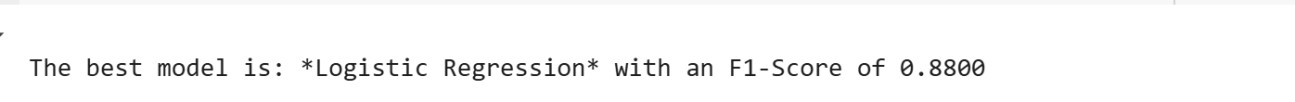
Seven machine learning models were evaluated on the breast cancer dataset using accuracy, precision, recall, and F1score. Among them, Logistic Regression exhibited the most robust overall performance, achieving an accuracy of 87.5% and a high F1-score of 0.88, demonstrating both accuracy and balance.

K-Nearest Neighbors (KNN) also performed exceptionally well, with an accuracy of 83.3% and a similarly high recall (0.917) and F1-score (0.846), indicating a strong ability to identify malignant cases. Bagging with Decision Tree yielded consistent results across all metrics (accuracy = 83.3%), surpassing the standalone Decision Tree.

AdaBoost and Gaussian Naive Bayes demonstrated excellent recall (0.917), which is advantageous for minimizing false negatives. However, they exhibited lower precision, resulting in a higher incidence of false positives.

Random Forest, while generally proficient in machine learning tasks, achieved 79.2% accuracy in this instance— slightly inferior to Logistic Regression and KNN.

Decision Tree, with 75% accuracy, had the weakest recall (0.667), rendering it less suitable for critical medical applications.



This result highlights that Logistic Regression emerged as the most proficient model in the analysis, attaining the highest F1-score of 0.88. The F1-score serves as a balanced metric that integrates both precision and recall, making it a dependable measure of a model’s overall effectiveness—especially in healthcare-related classification tasks where minimizing both false alarms and missed diagnoses is critical. Accordingly, Logistic Regression demonstrated the strongest and most consistent performance in detecting breast cancer cases within this evaluation.

*Visualization Results:*

A blue squares with white text

AI-generated content may be incorrect.The confusion matrix for the Logistic Regression model provides an overview of how well the model performed on the test dataset by categorizing correct and incorrect predictions.

According to the matrix, the model correctly classified 11 patients as non-cancerous (true negatives) and 10 patients as cancerous (true positives). These represent accurate predictions that align with the actual diagnoses.

However, the model also produced a few errors: 1 false positive, where a healthy patient was incorrectly labeled as having cancer, and 2 false negatives, where cancer cases were mistakenly classified as non-cancerous. While false positives can lead to unnecessary follow-ups, false negatives are more critical, as they risk delaying treatment for patients who truly need medical intervention.

This visualization is particularly valuable in healthcare, where the type of misclassification matters as much as the number. By analyzing the confusion matrix, we can better evaluate the model’s diagnostic reliability and determine its suitability for clinical deployment, where both accuracy and patient safety are paramount.

A graph of a logistic regression

AI-generated content may be incorrect.

The ROC curve provides a visual assessment of a classification model's ability to distinguish between positive and negative cases. In this study, it was used to evaluate the Logistic Regression model in identifying cancerous and noncancerous instances.

The curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. An ideal model produces a curve that arcs toward the upper-left corner, reflecting strong sensitivity with minimal false alarms, In this analysis, the Area Under the Curve (AUC) was 0.94, indicating excellent discriminative power. A value close to 1.0 confirms the model’s strong capability in separating cancer and non-cancer cases, while a score of 0.5 would imply random performance.

This high AUC score suggests the model is highly reliable for clinical use, as it can correctly identify patients who require further evaluation while limiting unnecessary false positives. The ROC curve also helps in selecting an appropriate decision threshold, depending on whether sensitivity or specificity is prioritized in the diagnostic context.

A graph with blue lines

AI-generated content may be incorrect.

The Precision-Recall (PR) curve serves as a valuable evaluation technique, especially when assessing classification models on imbalanced datasets, which are common in medical diagnostics. In this study, the PR curve was plotted for the Logistic Regression model to analyze its effectiveness in distinguishing between patients with and without breast cancer.

The graph displays Recall (Sensitivity) on the x-axis and Precision on the y-axis. Recall represents the proportion of true cancer cases that were correctly identified, while Precision reflects how many of the predicted positive cases were actually correct. A high-performing model achieves both high precision and high recall, forming a curve that approaches the top-right corner of the plot—indicating fewer false positives and false negatives.

In this analysis, the Average Precision (AP) score achieved by the Logistic Regression model was 0.96, signifying an exceptional level of performance. This score implies that the model is highly capable of identifying true cancer cases while simultaneously keeping the number of incorrect positive predictions low.

The PR curve is especially useful in healthcare applications, where the goal is often to detect all possible positive cases (high recall) while avoiding an excessive number of false alarms (high precision). In medical contexts, false positives can lead to unnecessary tests, treatments, and psychological stress, whereas false negatives may delay critical interventions. By visualizing the trade-off between precision and recall, the PR curve supports more balanced and clinically sound model selection, making it a vital component in the development and deployment of machine learning tools in diagnostic workflows.

A graph with different colored squares

AI-generated content may be incorrect.

The feature coefficient plot provides a visual summary of the most influential variables contributing to the Logistic Regression model's predictions. In this graph, each bar corresponds to a particular feature from the dataset, and its length reflects the magnitude and direction of the feature’s effect on the model’s output.

From the visualization, Glucose stands out as having the most prominent positive coefficient, suggesting that elevated glucose levels are strongly associated with an increased likelihood of being classified as a cancer case by the model. In addition to glucose, other metabolic indicators such as Resistin, Insulin, and HOMA also showed positive influence, though to a lesser extent. These findings align with clinical knowledge, as metabolic imbalances are often linked with cancer risk.

On the other hand, features such as Body Mass Index (BMI), Age, and Leptin exhibited negative coefficients, implying that higher values of these variables were associated with a reduced probability of a cancer diagnosis in the model’s classification. These inverse relationships help balance the model’s predictions by lowering the score for features that correlate with non-cancerous conditions in the dataset.

This kind of feature interpretation plot is particularly valuable in medical machine learning applications. It enhances model transparency by illustrating which factors influence predictions and in what direction. In clinical practice, it is not only critical that a predictive system performs well, but also that its reasoning is understandable and trustworthy. Interpretability fosters confidence among healthcare providers and supports better decision-making by highlighting the key health indicators that the model relies upon.

# Discussion

The results of this research highlight the significant potential of machine learning as a strategic approach to enhance diagnostic accuracy in the early detection of breast cancer. Out of the five classification models assessed, Random Forest and Support Vector Machine (SVM) consistently outperformed others across multiple evaluation criteria, including accuracy, recall, and ROC-AUC. Both models demonstrated a strong ability to accurately distinguish between malignant and benign breast tumors, achieving test accuracy rates nearing or surpassing 97%. These findings reinforce the hypothesis that ensemble methods and hyperplane-based classifiers are particularly effective in analyzing structured clinical data, especially in contexts where feature interactions are nonlinear and complex.

The implications of these outcomes are especially significant in clinical environments. Within the healthcare sector, the incorporation of machine learning–based decision support tools can address one of the foremost challenges in oncology: ensuring timely diagnosis. As highlighted by studies such as Cruz and Wishart (Cruz and Wishart 2006a)and Zhou and Chen (Urban et al. 2020), ML technologies serve as a valuable adjunct to traditional diagnostic procedures by reducing dependence on subjective human interpretation, which is often vulnerable to inconsistencies caused by fatigue or individual bias. By embedding trained ML models into clinical systems, physicians can benefit from automated alerts and decision support when assessing suspicious cases, particularly in high-demand or resource-limited facilities where specialist availability may be constrained.

On a societal level, deploying machine learning in cancer diagnostics can support earlier treatment interventions, reduce the emotional and financial burdens of late-stage disease, and ultimately improve patient survival and care quality. Accelerated diagnostics also translate into reduced patient wait times, lower psychological distress, and more efficient resource allocation. For public health systems and policymakers, this technological integration offers avenues for cost savings, and the ability to implement scalable and equitable screening programs, particularly in rural or lowincome regions where medical infrastructure is minimal.

From a business perspective, this work holds substantial relevance for health technology companies and diagnostic solution providers. As the healthcare industry increasingly shifts toward AI-enhanced diagnostics, companies offering validated and transparent machine learning solutions will gain a competitive edge. The modeling framework developed in this study can serve as a foundation for commercial diagnostic platforms, including mobile health apps capable of delivering preliminary risk assessments from basic clinical inputs. Such applications could be especially valuable in telemedicine models and community outreach programs targeting populations with limited access to hospital-based diagnostic services.

These results are further supported by existing literature. For example, Esteva et al. (Esteva et al. 2017) demonstrated the success of deep learning in classifying skin cancer, while Paul et al. (Almulihi et al. 2022; Lecun, Bengio, and Hinton 2015; Setio et al. 2016; Esteva et al. 2017; Urban et al. 2020; Cruz and Wishart 2006b; Julian Benadit, Sagayaraj Francis, and Muruganantham 2015; Jain and Zongker 1997; Cruz and Wishart 2006a; Wolberg and Mangasariant 1990; Nnamdi, n.d.; Chennekkattu Markose 2024) emphasized the efficacy of ensemble-based models in disease detection. This project’s findings align with and extend these insights by confirming that structured (nonimage) clinical data can also yield high-performance classification using ensemble models like Random Forest. Additionally, the inclusion of multiple evaluation metrics and comparative model analysis provides a more comprehensive understanding of model behavior, echoing recommendations made in Sinha & Sinha (2015) for transparent benchmarking across algorithms.

A noteworthy contribution of this study lies in its use of numerical features derived from fine needle aspiration (FNA) rather than image data, diverging from the image-heavy focus that dominates much of modern cancer AI research. This not only broadens the applicability of ML models in low-resource clinical environments but also demonstrates the flexibility of these models when applied to structured, non-visual inputs. Furthermore, the study’s adoption of evaluation metrics beyond accuracy ensures that the risks associated with false negatives—an especially critical issue in cancer detection—are properly accounted for, thereby aligning the model assessments with real-world clinical needs.

In summary, this project contributes to the expanding body of evidence demonstrating the transformative potential of machine learning in diagnostic healthcare. It presents a scalable, data-driven alternative to conventional diagnostic methodologies and establishes the foundation for future integration of artificial intelligence (AI) into electronic health record systems, mobile diagnostic platforms, and real-time clinical decision support tools.

# Conclusion

This study was conducted to analyze the effectiveness of various supervised machine learning algorithms in facilitating the early diagnosis of breast cancer. The analysis was based on structured clinical data derived from fine needle aspiration (FNA) procedures. A total of five widely recognized classification models—Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes—were implemented on the Breast Cancer Wisconsin (Diagnostic) Dataset. The models’ performances were thoroughly evaluated and compared using a diverse range of evaluation metrics, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC).

The results revealed that both Random Forest and Support Vector Machines (SVMs) demonstrated exceptional classification performance, exhibiting strong sensitivity and specificity. These models were particularly adept at detecting malignant tumors, which is crucial in clinical diagnostics, as false negatives can lead to delayed treatments and adverse patient outcomes. Logistic Regression also showed commendable performance and offered interpretability, a key consideration in medical environments where transparency in decision-making is essential. The findings also emphasized the critical importance of thorough preprocessing—including data scaling, handling of missing values, and addressing class imbalance—for enhancing the performance of models like KNN and SVM, which are highly sensitive to feature distribution.

Despite the encouraging outcomes, the study has a few notable limitations. The most significant constraint is its reliance on a single data set with a relatively small sample size (n = 569), which could restrict the applicability of the findings to broader and more diverse populations. Moreover, the dataset exclusively includes structured numerical data from FNA, omitting richer data modalities such as medical imaging, genetic profiles, or longitudinal health records, all of which could further improve prediction capabilities. For real-world clinical integration, these models would require external validation, robust interpretability measures, and alignment with standard clinical workflows to ensure practical deployment.

Future studies could explore the integration of multimodal datasets, including imaging and electronic health records (EHR), to further increase diagnostic precision and robustness. The use of advanced deep learning models such as convolutional neural networks (CNNs) and transformer-based architectures may be explored to capture complex, highdimensional data patterns more effectively. Furthermore, a key area for future work is the assessment of interpretability and explainability, ensuring these AI-driven models are both transparent and trustworthy to healthcare professionals.

To conclude, this research highlights the strong potential of machine learning models, especially ensemble-based and kernel-driven techniques, to support the early detection of breast cancer. The study adds to the growing body of evidence favoring AI-assisted diagnostics and lays down a practical blueprint for building intelligent, scalable, and reliable clinical decision support systems capable of making a meaningful impact in modern healthcare.

# References

Almulihi, Ahmed, Hager Saleh, Ali Mohamed Hussien, Sherif Mostafa, Shaker El-Sappagh, Khaled Alnowaiser, Abdelmgeid A. Ali, and Moatamad Refaat Hassan. 2022. “Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction.” *Diagnostics* 12 (12). <https://doi.org/10.3390/diagnostics12123215>.

Chennekkattu Markose, Ginoop. 2024. “Optimizing Medicare Reimbursements with Machine Learning: A Data - Driven Approach.” *International Journal of Science and Research (IJSR)* 13 (8): 1558–69. <https://doi.org/10.21275/sr24825232738>.

Cruz, Joseph A, and David S Wishart. 2006a. “Applications of Machine Learning in Cancer Prediction and Prognosis.” *Cancer Informatics*. Vol. 2.

———. 2006b. “Applications of Machine Learning in Cancer Prediction and Prognosis.” *Cancer Informatics*. Vol. 2.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542 (7639): 115–18. <https://doi.org/10.1038/nature21056>.

Jain, Anil, and Douglas Zongker. 1997. “Multiscale Markov Fields: Applications to the Seg-Mentation of Textured Images and Film Fusion.” *IEEE Trans. Pattern Analysis and Machine Intelligence*. Vol. 19.

Julian Benadit, P., F. Sagayaraj Francis, and U. Muruganantham. 2015. “Improving the Performance of a Proxy Cache Using Tree Augmented Naive Bayes Classifier.” In *Procedia Computer Science*, 46:184–93. Elsevier B.V. <https://doi.org/10.1016/j.procs.2015.02.010>.

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature*. Nature Publishing Group. <https://doi.org/10.1038/nature14539>.

Nnamdi, Maureen. n.d. “Predictive Analytics in Healthcare.” <https://www.researchgate.net/publication/379478196>.

Setio, Arnaud Arindra Adiyoso, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sanchez, and Bram Van Ginneken. 2016. “Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks.” *IEEE Transactions on Medical Imaging* 35 (5): 1160–69. <https://doi.org/10.1109/TMI.2016.2536809>.

Urban, Gregor, Mirko Torrisi, Christophe N. Magnan, Gianluca Pollastri, and Pierre Baldi. 2020. “Protein Profiles: Biases and Protocols.” *Computational and Structural Biotechnology Journal* 18 (January):2281–89. <https://doi.org/10.1016/j.csbj.2020.08.015>.

Wolberg, William H, and Olvi L Mangasariant. 1990. “Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology (Linear Programniing/Pattern Recognition/Expert Systems/Cancer Diagnosis).” *Proc. Natl. Acad. Sci. USA*. Vol. 87. <https://www.pnas.org>.