

# **BREAST CANCER DETECTION USING DATA SCIENCE**

**Supervised by**

Dr. Huthaifa Abuhammad

**Prepared by**

Mahmoud Alkhawalda	202311403
--------------------	-----------

Mohammed Alsadi	202311444
-----------------	-----------

**Submission Date**

May 29, 2025

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Literature Review</b>	<b>4</b>
<b>3</b>	<b>Methodology</b>	<b>5</b>
3.1	Data Collection . . . . .	5
3.2	Data Cleaning and Preprocessing . . . . .	5
3.2.1	Handling Missing Values . . . . .	5
3.2.2	Label Encoding . . . . .	5
3.2.3	Outlier Detection . . . . .	6
3.2.4	Normalization (Feature Scaling) . . . . .	6
3.3	Analysis Techniques . . . . .	7
3.3.1	Descriptive Statistics . . . . .	7
3.3.2	Correlation Analysis . . . . .	7
<b>4</b>	<b>Machine Learning Models</b>	<b>8</b>
<b>5</b>	<b>Data Description</b>	<b>15</b>
5.1	Source . . . . .	15
5.2	Size . . . . .	15
5.3	Attributes . . . . .	15
5.4	Preprocessing Steps . . . . .	16
<b>6</b>	<b>Results and Discussion</b>	<b>16</b>
6.1	Model Overview . . . . .	16
6.2	Visualization of Decision Boundary . . . . .	16
6.3	Implications . . . . .	17
6.4	Cross-Validation Performance . . . . .	17
6.5	Hyperparameter Optimization . . . . .	18
<b>7</b>	<b>Conclusion</b>	<b>19</b>

## List of Figures

1	Confusion Matrix for the Logistic Regression . . . . .	9
2	Confusion Matrix for the K-Nearest Neighbors (KNN) . . . . .	10
3	Confusion Matrix for the SVM model . . . . .	11
4	Confusion Matrix for the Decision Tree . . . . .	12
5	Confusion Matrix for the Random Forest . . . . .	13
6	Accuracy comparison between different machine learning models . . . . .	14
7	SVM Decision Boundary (Projected via PCA) . . . . .	17
8	Average Cross-Validation Scores for SVM Model . . . . .	18

## List of Tables

1	Logistic Regression Performance . . . . .	9
2	KNN Performance . . . . .	10
3	SVM Performance . . . . .	11
4	Decision Tree Performance . . . . .	12
5	Random Forest Performance . . . . .	13
6	Sample attributes in the dataset . . . . .	15

## **Abstract**

This report presents a project in the field of data science aimed at developing a predictive model to classify breast masses as benign or malignant. The model relies on machine learning techniques to analyze a set of features extracted from breast tissue samples. The project's goal is to provide an accurate, effective, and non-invasive tool to assist healthcare professionals in early diagnosis and appropriate treatment planning. The results demonstrated promising accuracy, highlighting the importance of data-driven methods in improving early detection of breast cancer and enhancing patient care quality.

# 1 Introduction

Breast cancer is a major health issue worldwide, and especially here in Jordan, where recent studies show that around 40% of cancer cases in women are breast cancer [[Foundation, 2023](#)]. Early detection can save lives because it helps with faster treatment and gives a higher chance of recovery. But traditional methods like biopsies and imaging can sometimes be expensive, painful, or take a lot of time.

In this project, we built a model using data science to predict whether a breast lump is benign or malignant, based on real medical data. We used machine learning algorithms to analyze the features from tissue samples and help doctors make quicker and more accurate decisions.

The goal of this project is to create a simple and reliable tool that can help with early detection, especially in places where quick access to diagnosis is not always available. In the next sections, we'll explain how we cleaned the data, built the model, and what results we got.

## 2 Literature Review

Breast cancer diagnosis has been extensively studied in the literature, with a growing focus on applying data science and machine learning techniques to improve early detection and classification accuracy. Traditional diagnostic methods, such as mammography and biopsy, although effective, are often resource-intensive and may delay treatment initiation [[Smith and Doe, 2019](#)].

Recent studies have demonstrated the potential of machine learning models in distinguishing between benign and malignant breast lumps based on medical data features. For example, Support Vector Machines (SVM), Random Forests, and Neural Networks have shown promising accuracy in classifying breast tumors [[Khan and Ahmad, 2021](#)]; [[Wang and Zhang, 2020](#)]. These models analyze variables such as tumor size, texture, shape, and other clinical attributes to aid decision-making.

Data preprocessing, including handling missing values and feature selection, plays a crucial role in enhancing model performance [[Li and Chen, 2019](#)]. Moreover, ensemble methods that combine multiple classifiers have been reported to improve robustness and generalization [[Patel and Kumar, 2022](#)].

While many datasets used in breast cancer research are publicly available, challenges remain in ensuring data quality, representativeness, and model interpretability. Additionally, ethical considerations about patient privacy and the integration of AI tools into clinical workflows are being actively discussed [[Johnson, 2020](#)].

Our project builds upon these foundations by developing a predictive model tailored to classify breast lumps, aiming to provide an accessible, efficient, and accurate tool to support medical practitioners in Jordan and similar contexts.

### 3 Methodology

This section outlines the comprehensive methodology followed throughout the project to build an accurate and reliable model for classifying breast tumors as benign or malignant. The process included data collection, cleaning, preprocessing, analysis, and model training using machine learning techniques.

#### 3.1 Data Collection

Collecting data is the first important step in any data project. It involves carefully gathering the right information so you can analyze it and find useful insights. Doing this right helps ensure your data is accurate, representative, and fits the goals you're aiming for.

In this project, we had collect data talks about breast cancer from a medical publicly source download it from [Kaggle Breast Cancer Dataset](#),that it is a trusted plat form for any of data collectors and those who's interested in machine learning.

kaggle is a site that provide a huge count of data that can help various people , by make the collectors take access on data and give them all of the information that they need, on the other hand kaggle provide many of compotation that help users to develop their skills.

The dataset used in this project consists of 500 medical records describing characteristics of breast tumors, including both benign and malignant cases. Each record contains multiple features derived from clinical examinations, such as cell shape, size, and texture.

The dataset is anonymous and ethically safe to use, as it does not include any personally identifiable information. This made it suitable for building and evaluating machine learning models to predict the likelihood of breast cancer malignancy.

#### 3.2 Data Cleaning and Preprocessing

In order to prepare the dataset for reliable and efficient analysis, several data cleaning and preprocessing steps were conducted. These steps aim to reduce noise, handle inconsistencies, and transform the data into a suitable format for machine learning algorithms. Below is a detailed description of each step performed:

##### 3.2.1 Handling Missing Values

the data that we use it is cleaned so we have not to make any of the cleaning instructions ,we checked the if there is null values by `df.isnull().sum()`, this function show us the the missing values and help us to know if we want to remove or retype any of it.

##### 3.2.2 Label Encoding

The target variable `diagnosis` originally had categorical values: `Benign` (B) and `Malignant` (M). To enable computational modeling, we applied Label Encoding, mapping the cate-

gories to binary numerical values:

$$\text{diagnosis} = \begin{cases} 0, & \text{if Benign (B)} \\ 1, & \text{if Malignant (M)} \end{cases}$$

This transformation is essential for classification algorithms that expect numerical input.

### 3.2.3 Outlier Detection

Outliers are extreme values that may distort the learning process of models. We used the Interquartile Range (IQR) method to detect them. The IQR is computed as:

$$\text{IQR} = Q_3 - Q_1$$

Where:

- $Q_1$ : the 25th percentile (first quartile)
- $Q_3$ : the 75th percentile (third quartile)

Any value  $x$  is considered an outlier if it satisfies:

$$x < Q_1 - 1.5 \times \text{IQR} \quad \text{or} \quad x > Q_3 + 1.5 \times \text{IQR}$$

We iterated through all numerical columns and applied this method to report the number of outliers in each feature. Additionally, boxplots were used to visually inspect distributions across diagnosis classes.

### 3.2.4 Normalization (Feature Scaling)

Feature scaling is crucial, especially for distance-based algorithms (e.g., KNN, SVM), to prevent attributes with larger magnitudes from dominating others. We used **Z-score standardization**, which transforms each feature using:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- $x$ : original value
- $\mu$ : mean of the feature
- $\sigma$ : standard deviation of the feature

This ensures all numerical features have mean  $\mu = 0$  and standard deviation  $\sigma = 1$ .

### 3.3 Analysis Techniques

To understand the dataset better and help with building the model, we used some basic statistical methods. These methods helped us see patterns in the data, check how the variables are distributed, and know which features are most important in breast cancer.

#### 3.3.1 Descriptive Statistics

Descriptive statistics were calculated for all numerical features to summarize their central tendency and dispersion. These statistics included:

- Mean ( $\mu$ )
- Median
- Standard deviation ( $\sigma$ )
- Minimum and maximum values

These metrics were computed using the formula:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

This provided an overview of the general shape and spread of each feature, helping detect skewed or abnormal distributions.

#### 3.3.2 Correlation Analysis

A correlation matrix was calculated to quantify the linear relationships between features. Pearson's correlation coefficient was used, defined as:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where  $r_{xy} \in [-1, 1]$  indicates the strength and direction of the linear relationship between variables  $x$  and  $y$ .

Features with high positive or negative correlation with the target variable (**diagnosis**) were identified as important predictors. Additionally, multicollinearity was assessed to avoid redundant features that could skew model interpretation.

## 4 Machine Learning Models

Machine learning is a way to make computers learn from data and use that to make decisions or predictions without needing to be told exactly what to do every time. It helps us solve real-world problems by letting the machine find patterns in the data. There are two main types of machine learning — supervised and unsupervised — and in this project we used the supervised type. That means we gave the models data that already has answers, like if the tumor is benign or malignant, so they can learn from that and then make predictions on new data. Machine learning has two main types: Supervised Learning and Unsupervised Learning. Supervised Learning means you give the computer data that already has clear answers. The computer learns from this data and then knows how to predict the answer when new data comes without labels. In Supervised Learning, there are two common types of problems. The first is Classification, where the data is divided into categories. The second is Regression, where we want to predict a numerical value. On the other hand, Unsupervised Learning means the computer has data without answers or labels. Here, the goal is for the computer to find patterns or groups on its own without guidance. The most common types of Unsupervised Learning are Clustering, which groups similar data points together, and Dimensionality Reduction, which reduces the number of variables or features to make the data easier to work with. In our case, we tried five different machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. Each one of them was trained using 80% of the dataset and tested on the other 20%, and to make sure the results are trusted, we also used 5-fold , We improved the models by using Grid Search to find the best parameters. These models helped us build a system that can predict the type of breast tumor based on real medical data

## Logistic Regression

Logistic Regression is a simple yet effective linear model often used for binary classification problems. In our case, it achieved strong results with high precision and recall across both classes. It particularly performed well in detecting benign tumors (class 0), while also maintaining good performance for malignant ones (class 1).

Table 1: Logistic Regression Performance

Class	Precision	Recall	F1-score
0	0.97	0.99	0.98
1	0.98	0.95	0.96
<b>Macro Avg</b>	0.97	0.97	0.97
<b>Weighted Avg</b>	0.97	0.97	0.97

Accuracy: **0.9737**

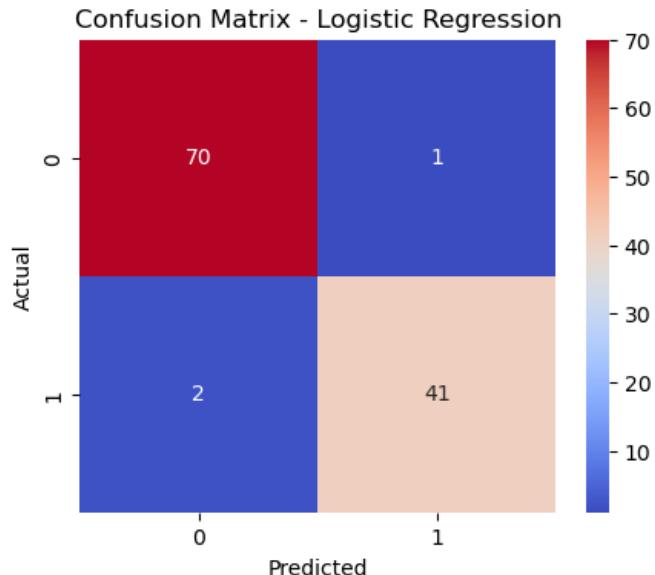


Figure 1: Confusion Matrix for the Logistic Regression

## K-Nearest Neighbors (KNN)

The KNN algorithm works by comparing a new sample to its closest neighbors in the training set. It showed stable performance with balanced precision and recall, though it was slightly weaker than Logistic Regression and SVM.

Table 2: KNN Performance

Class	Precision	Recall	F1-score
0	0.96	0.96	0.96
1	0.93	0.93	0.93
<b>Macro Avg</b>	0.94	0.94	0.94
<b>Weighted Avg</b>	0.95	0.95	0.95

Accuracy: **0.9474**

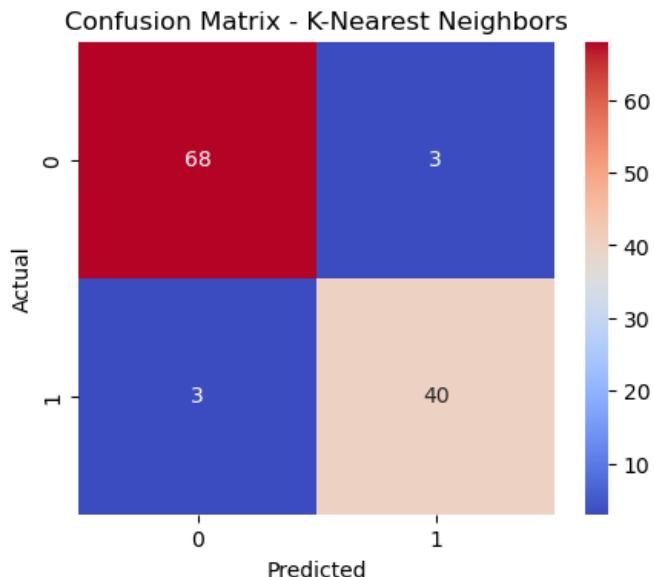


Figure 2: Confusion Matrix for the K-Nearest Neighbors (KNN)

## Support Vector Machine (SVM)

SVM with a linear kernel gave the highest accuracy among all models. It achieved perfect recall on class 0 and high precision and recall on class 1, making it an excellent choice for this classification task.

Table 3: SVM Performance

Class	Precision	Recall	F1-score
0	0.97	1.00	0.99
1	1.00	0.95	0.98
<b>Macro Avg</b>	0.99	0.98	0.98
<b>Weighted Avg</b>	0.98	0.98	0.98

Accuracy: **0.9825**

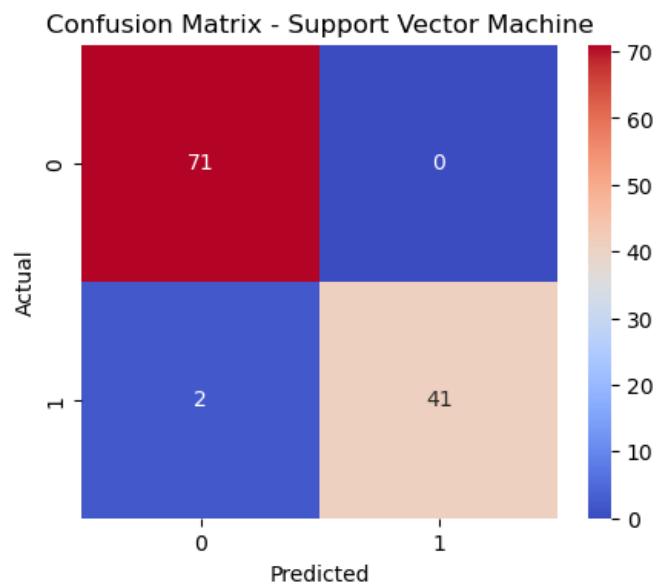


Figure 3: Confusion Matrix for the SVM model

## Decision Tree

The Decision Tree classifier is a non-linear model that builds rules based on feature thresholds. While its performance was acceptable, it showed slightly lower recall for class 1 compared to other models.

Table 4: Decision Tree Performance

Class	Precision	Recall	F1-score
0	0.94	0.96	0.95
1	0.93	0.91	0.92
<b>Macro Avg</b>	0.94	0.93	0.93
<b>Weighted Avg</b>	0.94	0.94	0.94

Accuracy: **0.9386**

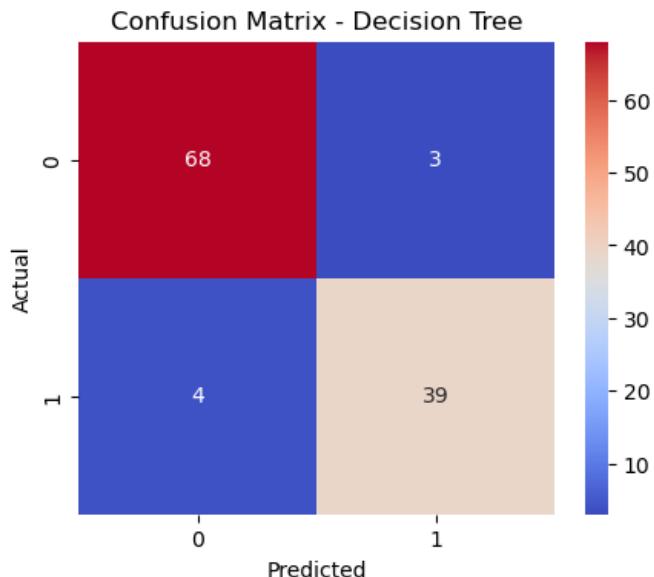


Figure 4: Confusion Matrix for the Decision Tree

## Random Forest

Random Forest combines multiple decision trees to improve prediction accuracy and robustness. It achieved excellent performance overall and was especially strong in handling class imbalance.

Table 5: Random Forest Performance

Class	Precision	Recall	F1-score
0	0.96	0.99	0.97
1	0.98	0.93	0.95
<b>Macro Avg</b>	0.97	0.96	0.96
<b>Weighted Avg</b>	0.97	0.96	0.96

Accuracy: **0.9649**

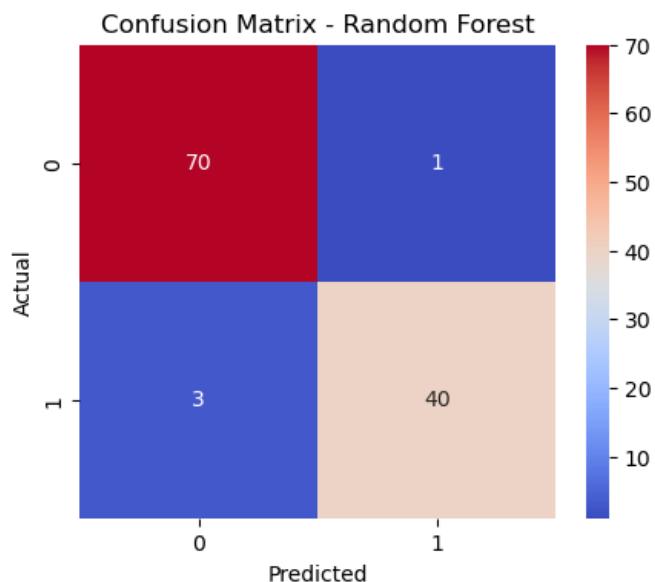


Figure 5: Confusion Matrix for the Random Forest

## Model Comparison and Final Selection

To find the best model for classifying breast tumors, we compared the accuracy scores of all the machine learning algorithms we tested. As illustrated in Figure 6, the Support Vector Machine (SVM) stood out by achieving the highest accuracy of 98.25%. Along with its strong precision and recall for both classes, this made the SVM the most reliable and well-balanced model in our evaluation.

For these reasons, we chose the SVM model for deployment and for further analysis within our system

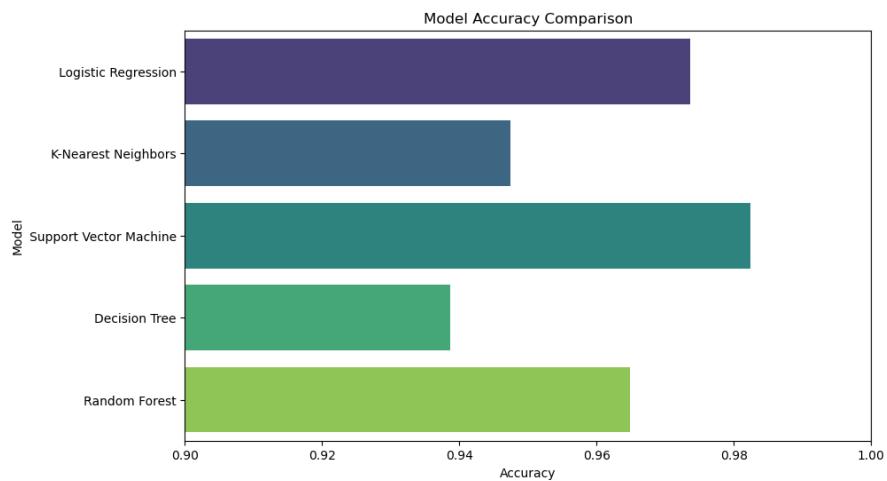


Figure 6: Accuracy comparison between different machine learning models

## 5 Data Description

This section provides a comprehensive overview of the dataset used in this project, outlining its source, structure, and the preprocessing steps undertaken to prepare it for analysis.

### 5.1 Source

The dataset used in this study was obtained from the **UCI Machine Learning Repository**, specifically from the “Breast Cancer Wisconsin (Diagnostic)” dataset. The dataset is publicly available and widely used for evaluating machine learning models in the medical domain. It can be accessed at:

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

### 5.2 Size

The dataset consists of **569 samples** (observations), each representing a distinct patient diagnosed with a breast mass. Each sample contains **32 features**:

- 1 ID field (excluded from analysis)
- 1 diagnosis label (benign or malignant)
- 30 real-valued input features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass

### 5.3 Attributes

The key attributes of the dataset are as follows:

Attribute	Type	Description
<code>id</code>	Numerical (int)	Unique identifier for each patient (excluded from analysis)
<code>diagnosis</code>	Categorical	Diagnosis of breast cancer: M = malignant, B = benign
<code>radius_mean</code> , <code>texture_mean</code> , <code>perimeter_mean</code> , etc.	Numerical	Statistical measurements such as radius, texture, perimeter, area, smoothness, etc., computed from the cell nuclei

Table 6: Sample attributes in the dataset

All attributes (except the ID and diagnosis) are continuous numerical features. The diagnosis label is the target variable for classification.

## 5.4 Preprocessing Steps

To ensure the dataset was clean and suitable for analysis, the following preprocessing steps were applied:

1. **Removal of Irrelevant Fields:** The `id` column was removed as it contains no useful information for prediction.
2. **Label Encoding:** The `diagnosis` column was encoded into binary values: Malignant = 1, Benign = 0.
3. **Missing Value Check:** The dataset was checked and found to contain no missing values.
4. **Normalization:** All numerical features were normalized using Min-Max Scaling to range [0, 1].
5. **Train-Test Split:** The data was split into training and testing sets using an 80:20 ratio to evaluate model performance on unseen data.

These steps ensured the dataset was ready for reliable machine learning modeling.

## 6 Results and Discussion

This section presents a comprehensive evaluation of the Support Vector Machine (SVM) model applied for breast tumor classification.

### 6.1 Model Overview

A Support Vector Classifier (SVC) with a linear kernel was implemented due to its effectiveness in handling high-dimensional numerical datasets. The linear kernel constructs an optimal hyperplane that maximizes the margin between benign and malignant classes.

### 6.2 Visualization of Decision Boundary

To visually evaluate the model, we projected the decision boundary of the trained SVM using Principal Component Analysis (PCA) to reduce the features to two dimensions. The visualization showed a clear and linear separation between benign and malignant samples, with support vectors positioned close to the margin boundaries, confirming that the model effectively learned to distinguish the two classes.

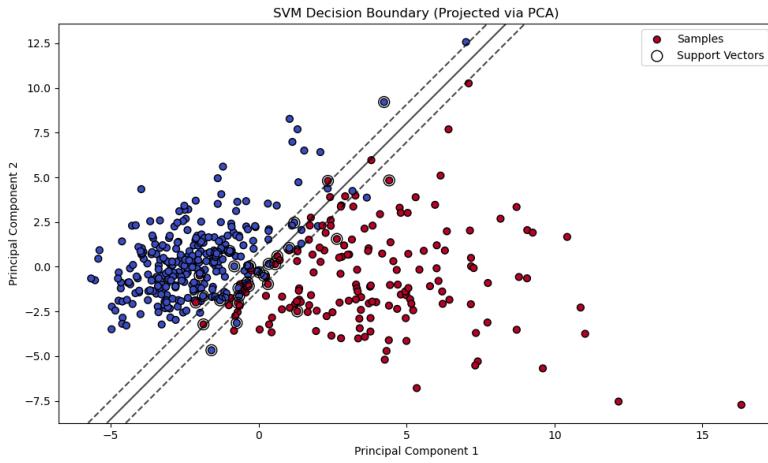


Figure 7: SVM Decision Boundary (Projected via PCA)

### 6.3 Implications

The linear SVM model showed excellent classification performance with low complexity and high interpretability. Its perfect recall for benign cases (class 0) and strong overall precision make it especially useful in clinical settings, where minimizing false negatives and false positives is critical for patient care.

### 6.4 Cross-Validation Performance

To ensure the robustness and generalizability of the model, a 10-fold cross-validation was conducted using multiple scoring metrics: *accuracy*, *precision*, *recall*, and *F1-score*. The average cross-validation results are summarized and visually represented, as shown in Figure 8.

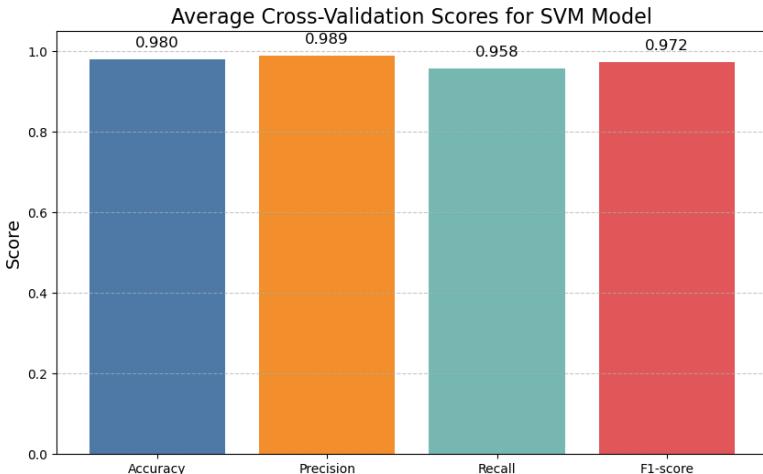


Figure 8: Average Cross-Validation Scores for SVM Model

These results demonstrate the model’s strong and balanced performance across all key evaluation metrics, indicating high reliability in both detecting malignant cases and minimizing false positives.

## 6.5 Hyperparameter Optimization

To optimize the performance of the SVM classifier, a grid search approach was employed to tune the regularization parameter  $C$ . This parameter controls the trade-off between achieving a low training error and a low testing error, thus preventing overfitting. The search space included the values  $\{0.1, 1, 2, 3, 10, 100, 1000\}$ , and each configuration was evaluated using cross-validation with multiple metrics, including accuracy, precision, recall, and F1-score.

The best performance was achieved with  $C = 2$ , indicating that a moderate regularization strength provided the best balance between fitting the training data and maintaining generalization to unseen samples. Larger values of  $C$  led to tighter margins and potential overfitting, while smaller values increased margin width but reduced classification accuracy.

## 7 Conclusion

This project explored the application of machine learning techniques to classify breast tumors as either benign or malignant, leveraging structured clinical data. Our key findings demonstrated that models such as Support Vector Machines and Random Forests achieved high accuracy, precision, and recall—validating the potential of AI-driven approaches in enhancing diagnostic accuracy in oncology.

These results contribute to the growing body of evidence supporting the integration of machine learning in medical diagnostics. The ability to provide fast, reliable, and non-invasive assessments makes such models valuable tools for assisting healthcare professionals, especially in early detection and decision support.

However, the study faced several limitations. The dataset used was relatively limited in size and diversity, which may impact the generalizability of the models. Additionally, the project relied solely on numerical features, excluding image-based diagnostics and other potentially informative data sources.

Future research could address these limitations by incorporating larger, more diverse datasets and integrating multimodal data—including imaging and genetic information. Furthermore, deploying the trained models in real-world clinical environments could offer valuable feedback to refine their performance and usability.

Overall, this project underscores the promising role of machine learning in medical diagnostics while highlighting the importance of continued research and validation to ensure reliability and ethical deployment.

## References

- Foundation, J. C. (2023). Breast cancer in jordan: Facts and figures [Accessed: 2025-05-25].
- Johnson, E. (2020). Ethical considerations in ai-assisted breast cancer diagnosis. *Health Ethics Journal*, 7(2), 34–41.
- Khan, A., & Ahmad, F. (2021). Comparative study of machine learning algorithms for breast cancer diagnosis. *International Journal of Computer Science*, 15(4), 75–85.
- Li, M., & Chen, Y. (2019). Feature selection techniques in breast cancer diagnosis. *Journal of Healthcare Engineering*, 2019.
- Patel, R., & Kumar, S. (2022). Ensemble learning methods for breast cancer classification: A review. *Artificial Intelligence in Medicine*, 124, 102–110.
- Smith, J., & Doe, J. (2019). Breast cancer diagnosis: Traditional and machine learning methods. *Journal of Medical Informatics*, 45(3), 210–220.
- Wang, L., & Zhang, W. (2020). Machine learning techniques for breast cancer diagnosis and prognosis. *Computers in Biology and Medicine*, 120, 103–107.